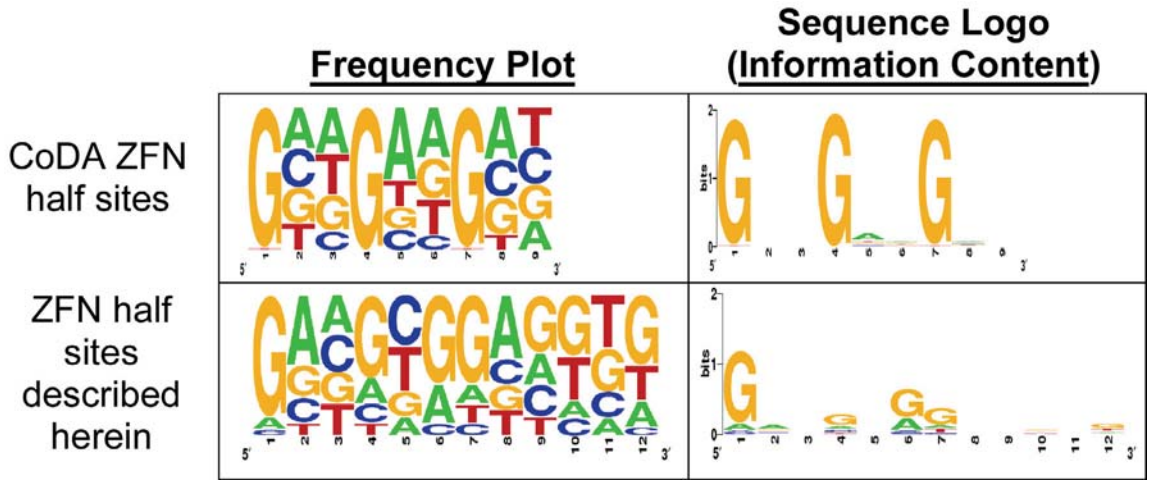
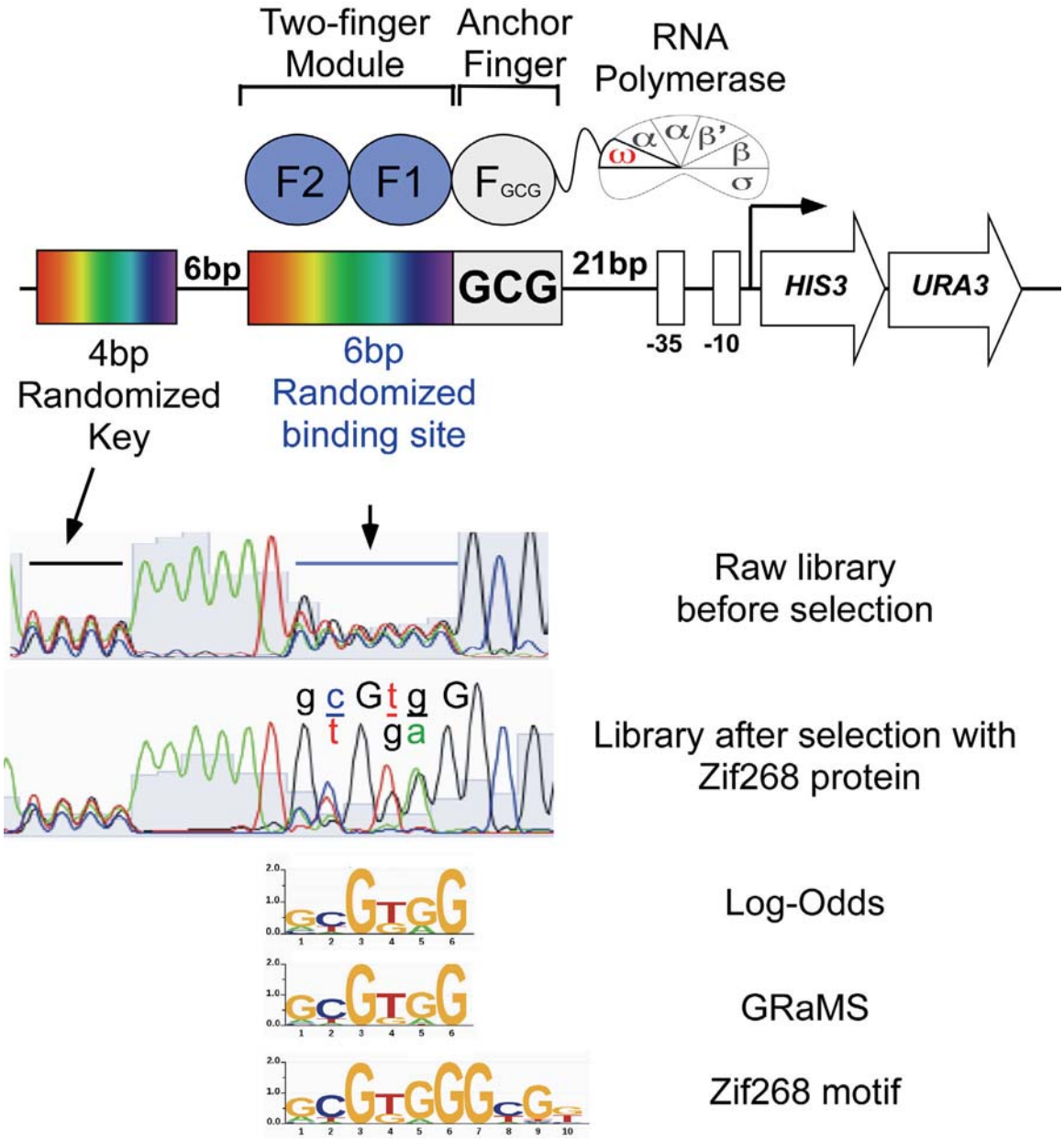
Supplementary Figure 1: Comparison of target site composition for CoDA-ZFNs and our tested ZFNs.



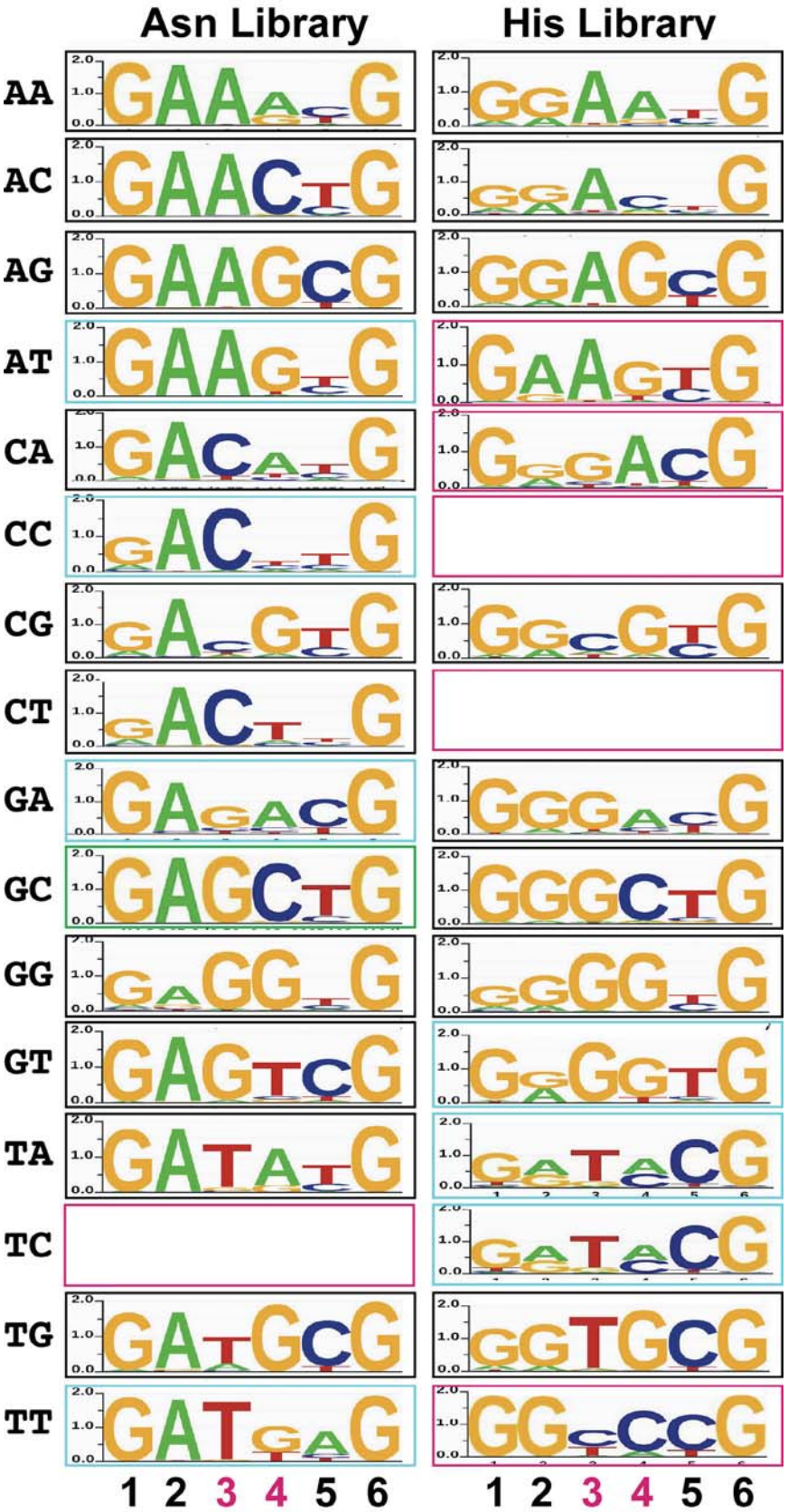
The half sites for ZFNs constructed in Sander et. al. using the CoDA strategy[11] and ZFNs constructed in this paper were compiled by aligning their 5' ends. Frequency plots and Sequence Logos displaying information content on a 2-bit scale were generated for each set of sites using Weblogo[24].

Supplementary Figure 2: Identification of DNA binding specificity for 2F-modules using the CV-B1H method[25].



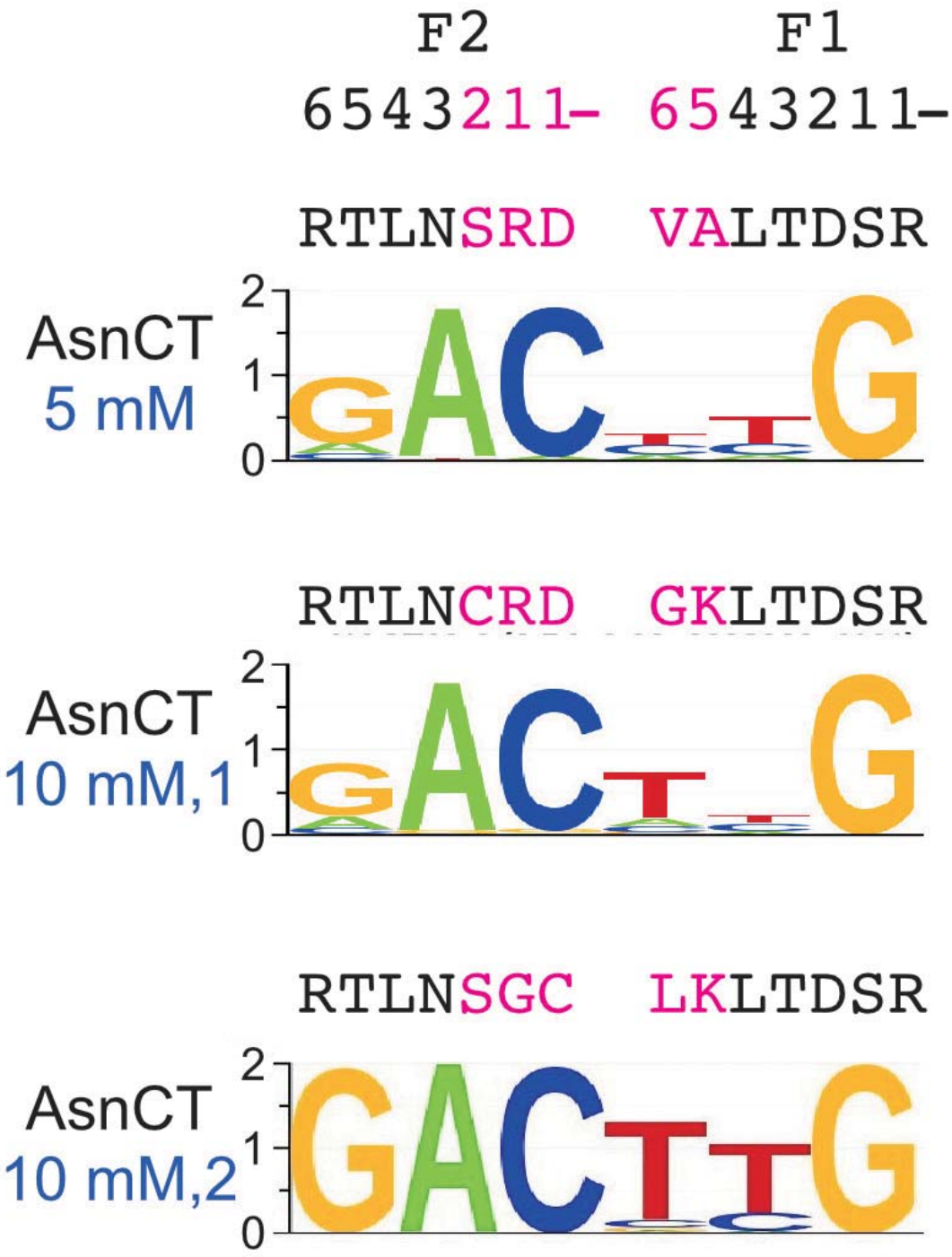
The 2F-module is fused to an N-terminal finger (RSDTLAR) that binds to the 'GCG' triplet adjacent to the 6bp randomized zinc finger binding region on the reporter plasmid. Also included is a 4bp randomized region (key region) that serves as an internal control to identify biases in the recovered DNA sequences due to jackpot effects. Following selection, the surviving colonies are pooled and the distribution of bases recovered at each position within the selected binding sites can be evaluated in a single sequencing reaction as shown here for finger 2 and finger 3 of Zif268. The recovered binding sites are determined by Illumina sequencing and then a binding site motif is calculated from these sequences using either log-odds-like or GRaMS (Growth Rate Modeling of Specificities) method[25]. For Zif268 F2 and F3, the binding site model obtained using the log-odds-like and GRaMS method closely matches the motif obtained by HT-SELEX[26].

Supplementary Figure 3: Montage showing the binding site specificities of the best 2F-modules selected from the Asn+3F2 and the His+3F2 library for each 2bp junction.



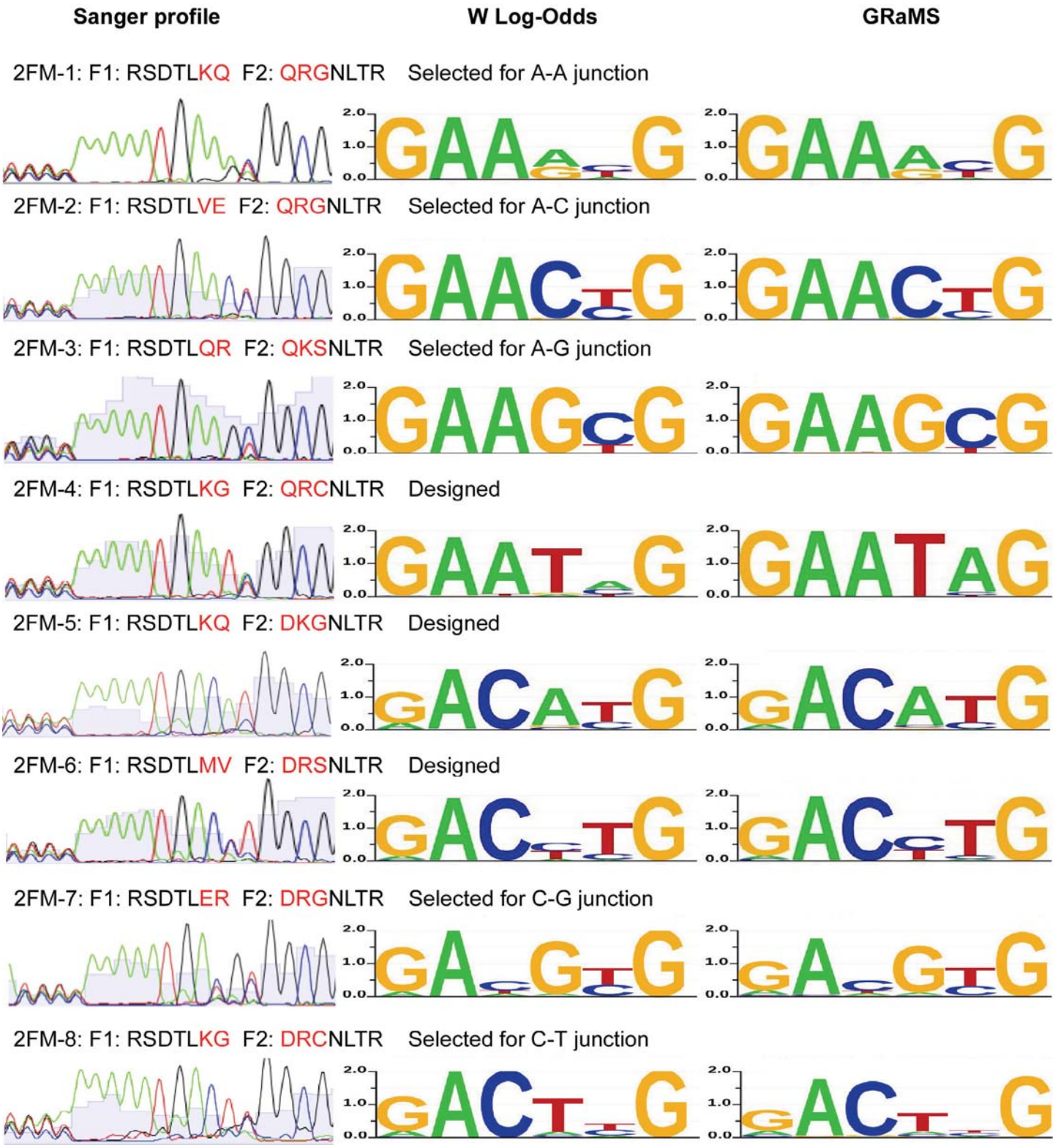
The 2F-modules are designated as having 'preferential specificity' (black box), 'compatible specificity' (cyan box) or 'poor specificity' (magenta box) for the desired target sequence.

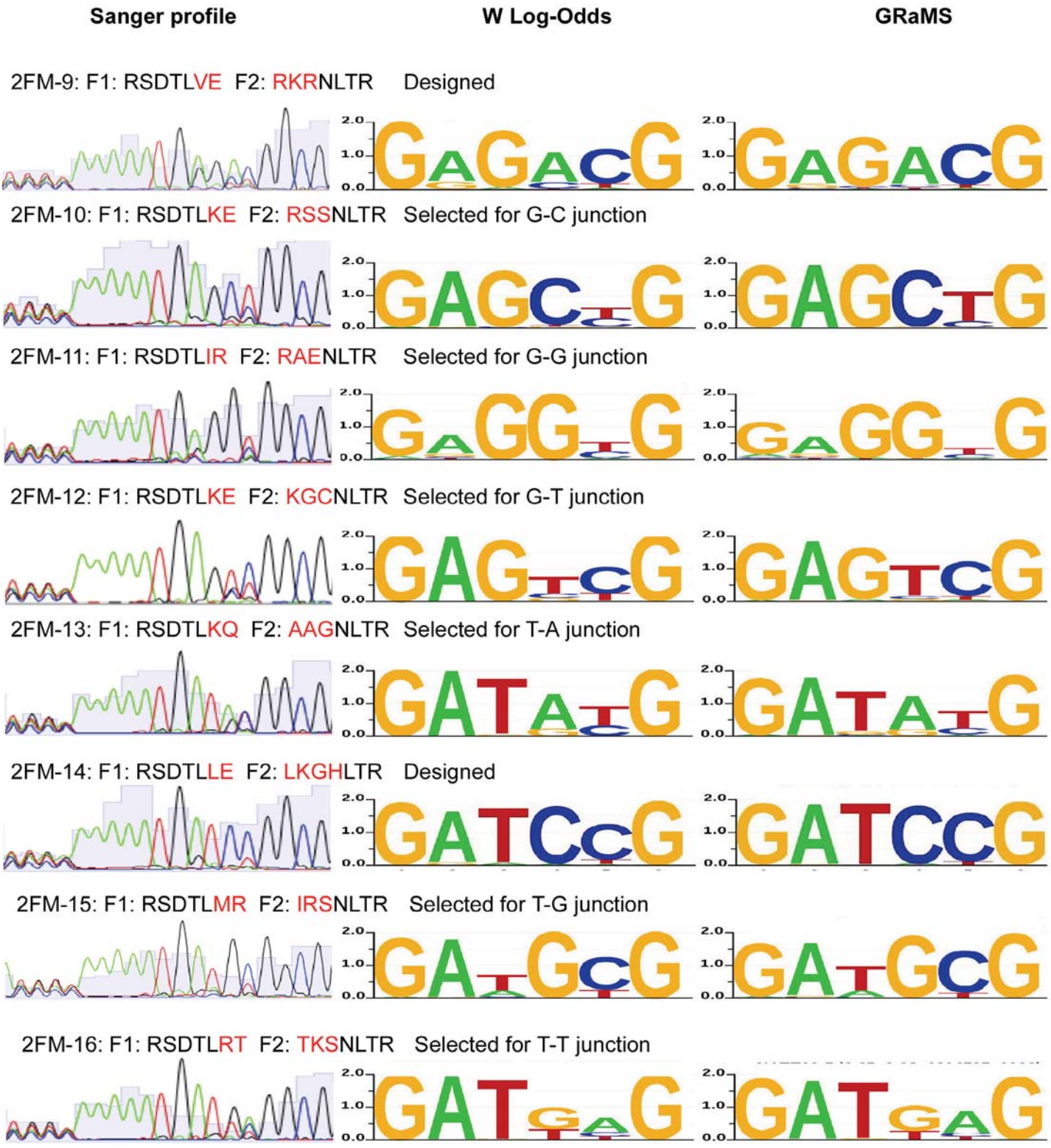
Supplementary Figure 4: Influence of stringency on the specificity of the recovered 2F-module.



Comparison of specificities of one 2F-module selected at low stringency (5mM 3-AT) and two modules (10 mM,1 & 10 mM,2) selected at high stringency (10mM) for a 'CT' junction sequence from the Asn+3F2-library. The modules recovered from the higher stringency selection display increased preference for their target site.

Supplementary Figure 5: DNA-binding site specificities for 2F modules that bind GAN-NYG and GGN-NYG sequences.




Sanger profile
W Log-Odds
GRaMS
2FM-9: F1: RSDTLVE F2: RKRNLTR Designed
2FM-10: F1: RSDTLKE F2: RSSNLTR Selected for G-C junction
2FM-11: F1: RSDTLIR F2: RAENLTR Selected for G-G junction
2FM-12: F1: RSDTLKE F2: KGCNLTR Selected for G-T junction
2FM-13: F1: RSDTLKQ F2: AAGNLTR Selected for T-A junction
2FM-14: F1: RSDTLLE F2: LKGHLTR Designed
2FM-15: F1: RSDTLMR F2: IRSNLTR Selected for T-G junction
2FM-16: F1: RSDTLRT F2: TKSNLTR Selected for T-T junction

Sanger profile **W Log-Odds** **GRaMS**

2FM-17: F1: RSDTLTQ F2: QRGHLTR Selected for A-A junction

2FM-18: F1: RSDTLRE F2: QRGHLTR Selected for A-C junction

2FM-19: F1: RSDTLVR F2: QSGHLTR Selected for A-G junction

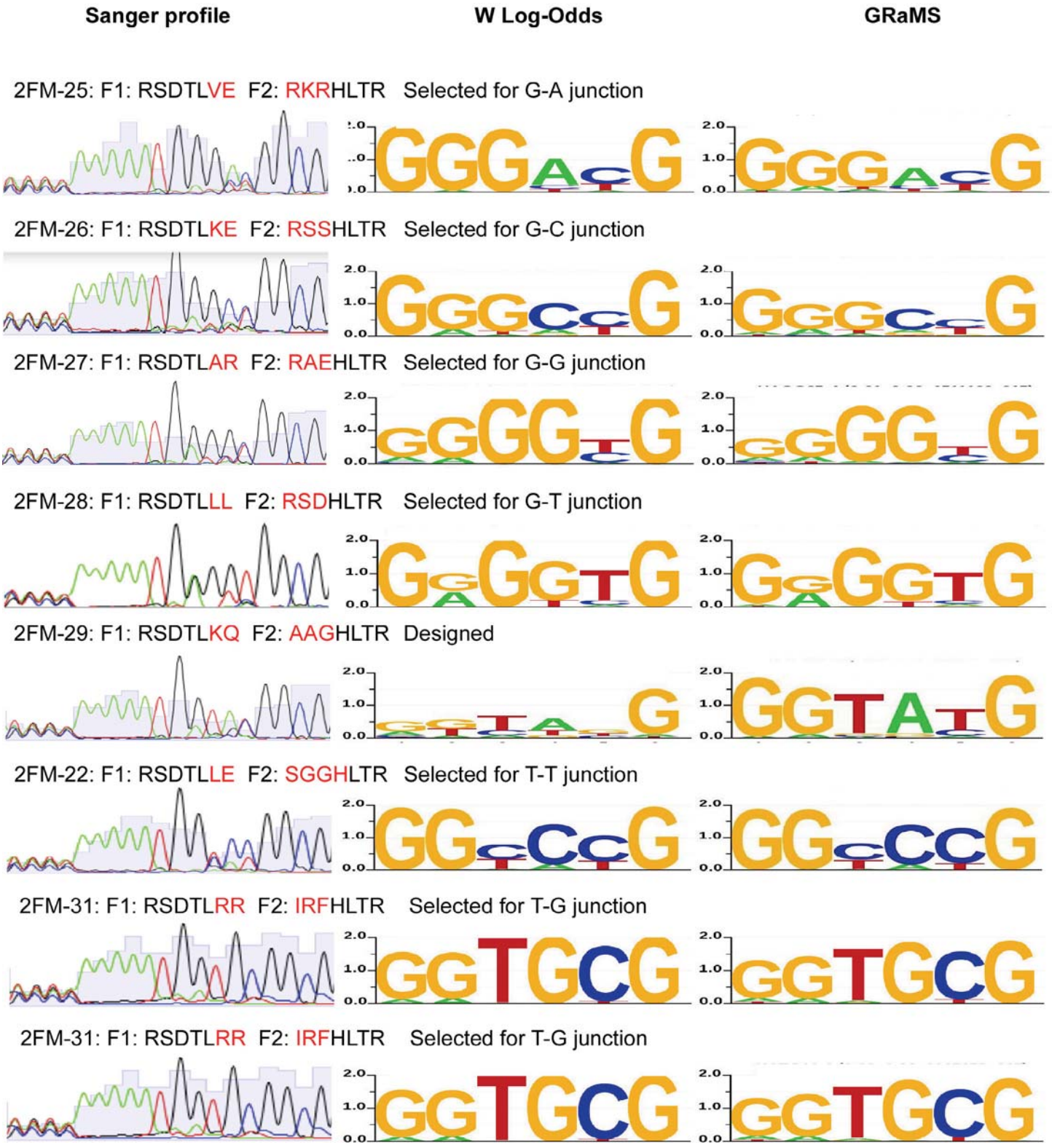
2FM-20: F1: RSDTLKG F2: QRCHLTR Designed

2FM-1014: F1: RSDTLKQ F2: ARRNLTR selected for C-A with Asn lib

2FM-21: F1: RSDTLMV F2: DRSHLTR Designed

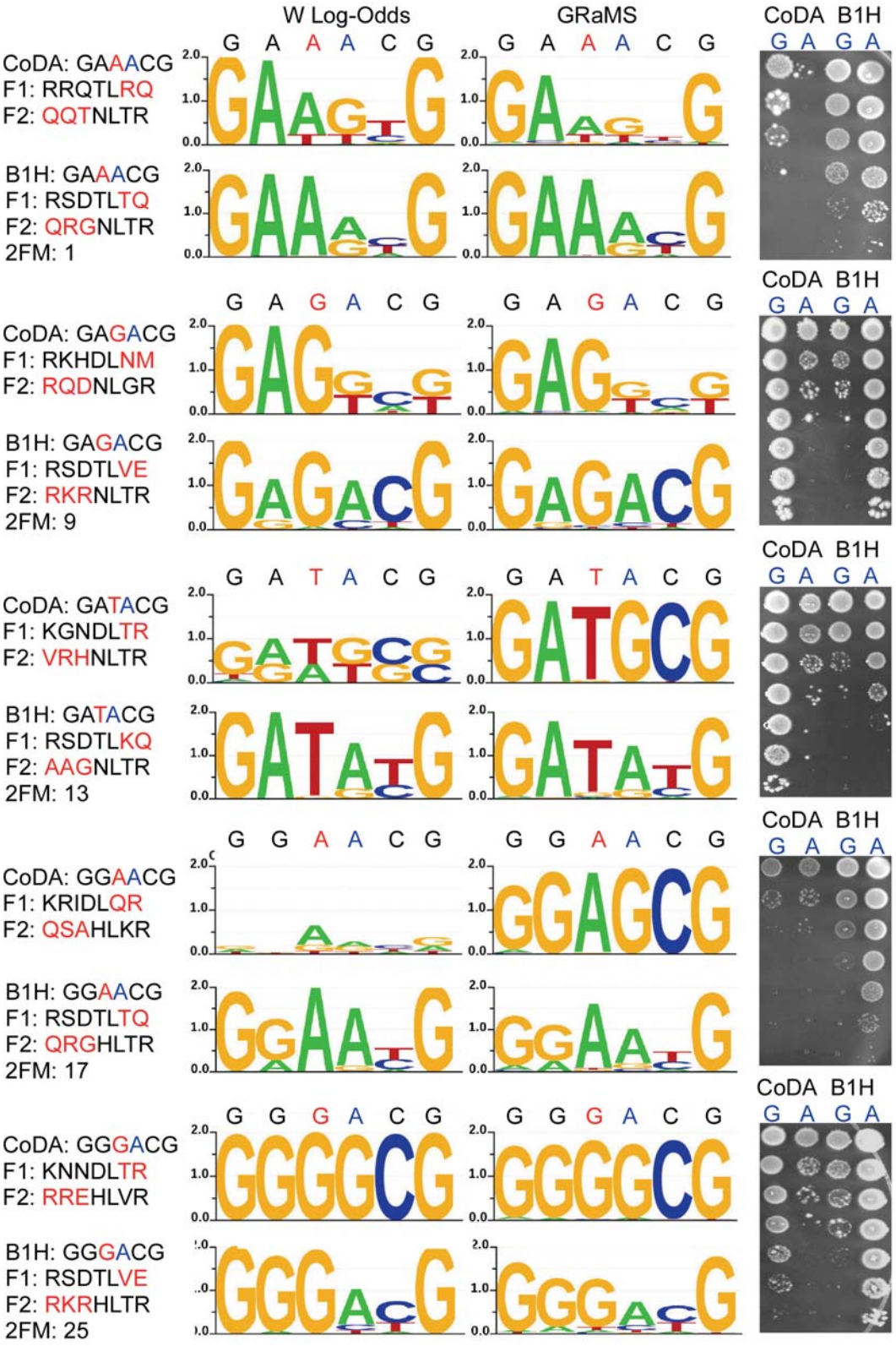
2FM-23: F1: RSDTLLR F2: ESGHLTR Selected for C-G junction

2FM-24:: F1: RSDTLKG F2: DRCHLTR Designed



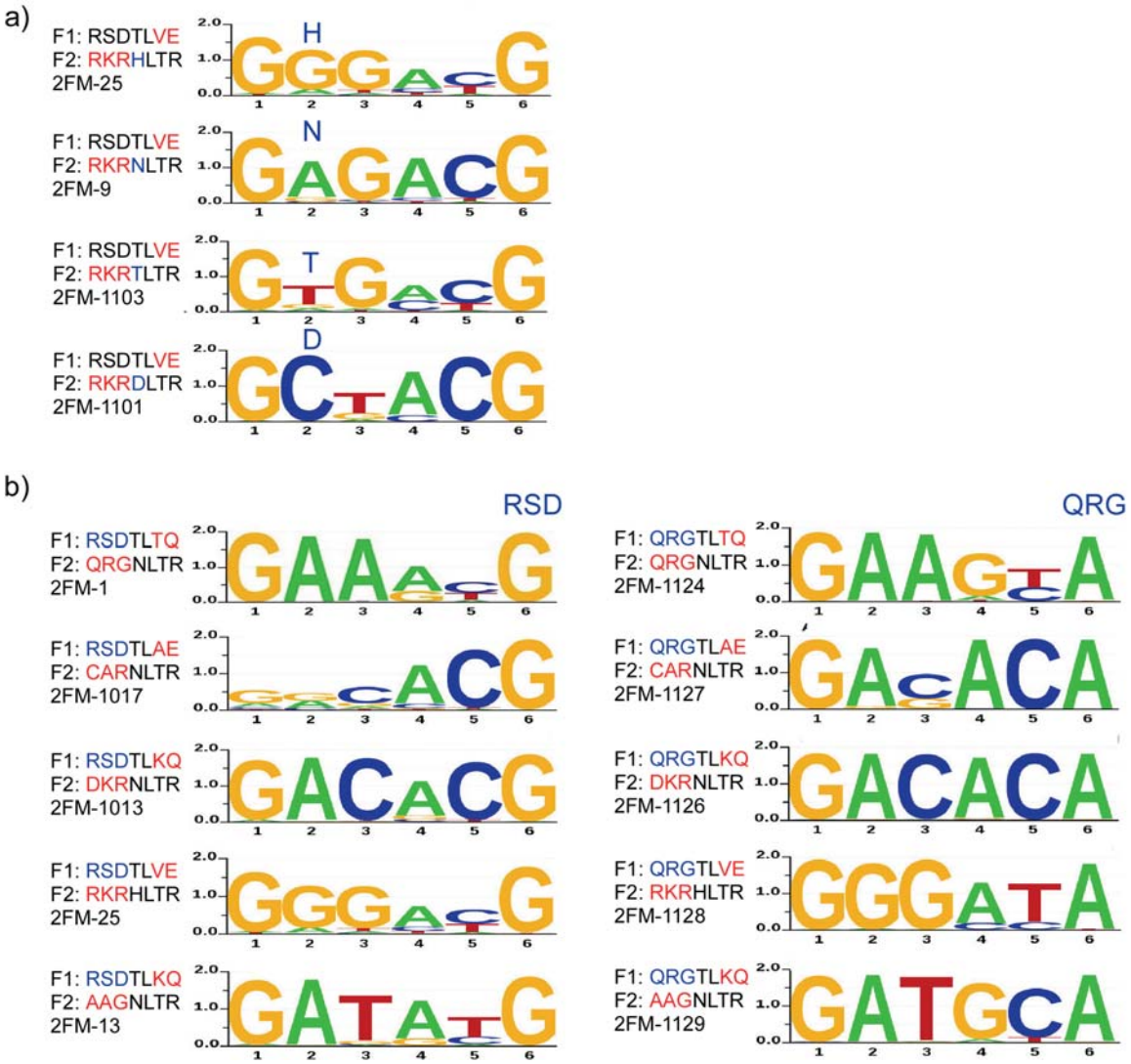
The 2F modules obtained via B1H-selections or rational design that bind each of 16 GAN-NYG and GGN-NYG sequences with highest specificity are displayed. The recognition helix sequences (positions -1, 1, 2, 3, 4, 5 and 6) for the F1 and F2 are shown, and the finger origin is indicated beside each sequence. The randomized interface positions are shown in red. Binding site specificities were determined using the CV-B1H method. The chromatograms are binding site profiles obtained by Sanger sequencing the pools of selected binding sites. Binding site logos were obtained via log-odds-like and GRaMS modeling post Illumina sequencing. For the GGN-NYG 2F-modules, all modules display 'good specificity' except for the CC, GT and TC modules that display 'compatible specificity' and the AT and TT modules that display 'poor specificity' for their target sites.

Supplementary Figure 6: Comparison of specificities of CoDA-2F modules and our 2F modules.



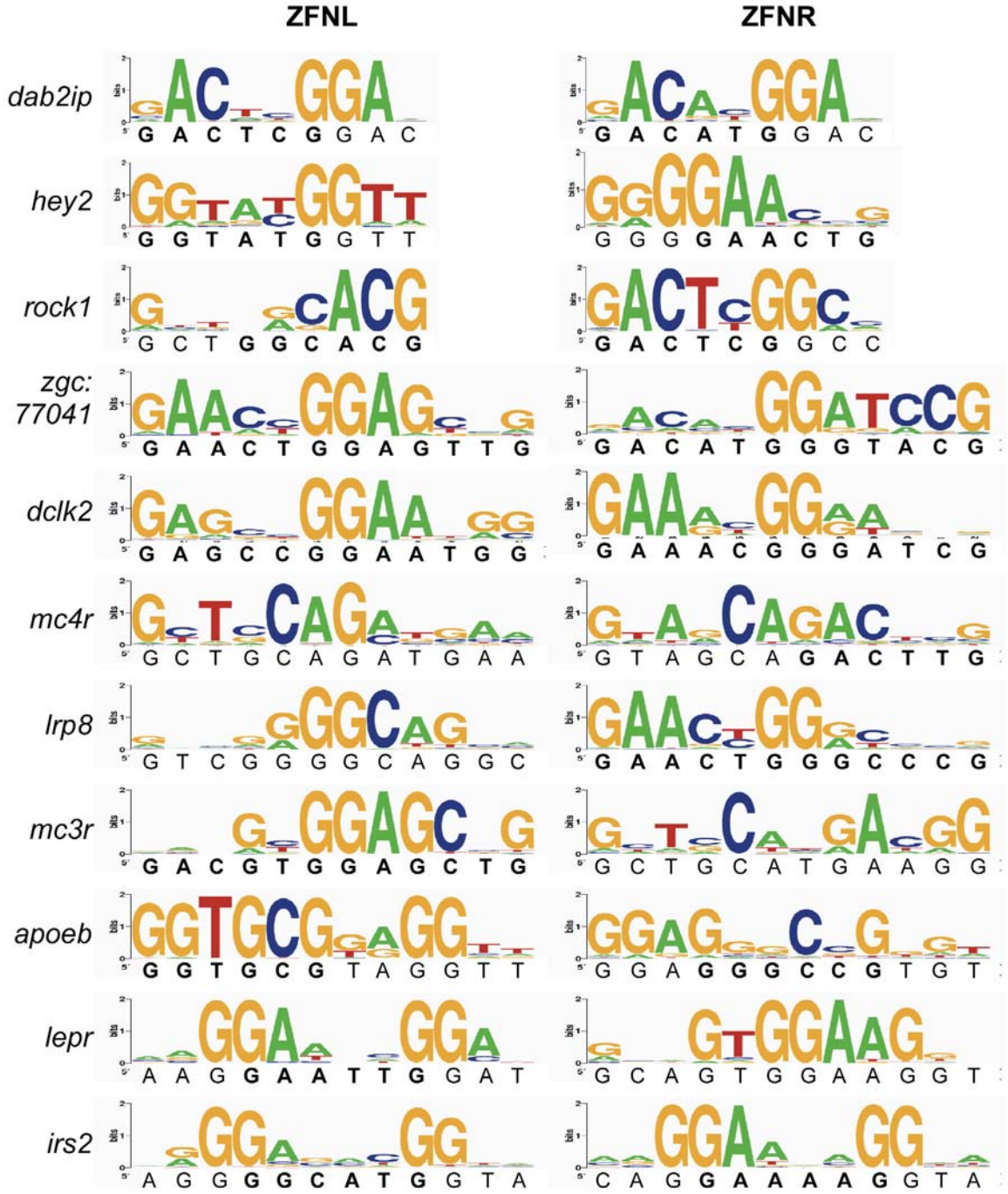
Five CoDA[11] 2F modules were fused to the 'GCG'-binding F1 followed by their binding site analysis via the CV-B1H assay. The binding site logos obtained through GRaMS analysis are displayed for both the CoDA modules and the equivalent B1H-selected modules, where the recognition helices are shown for comparison. A B1H-activity assay was performed for the CoDA and B1H-selected 2F modules (in combination with the 'GCG'-binding F1) against fixed binding sites with either Adenine or Guanine at the 4th position (GAXYCG, where Y is either A or G) to determine the relative activity of the 2F module on each sequence variant. Each row in the assay represents 10-fold dilution of bacterial cells on plates containing the His3 inhibitor 3-AT to provide a stringent challenge to ZFA-driven reporter activity.

Supplementary Figure 7: Examples of module alterations designed to expand the archive of targetable sequences through rational design.



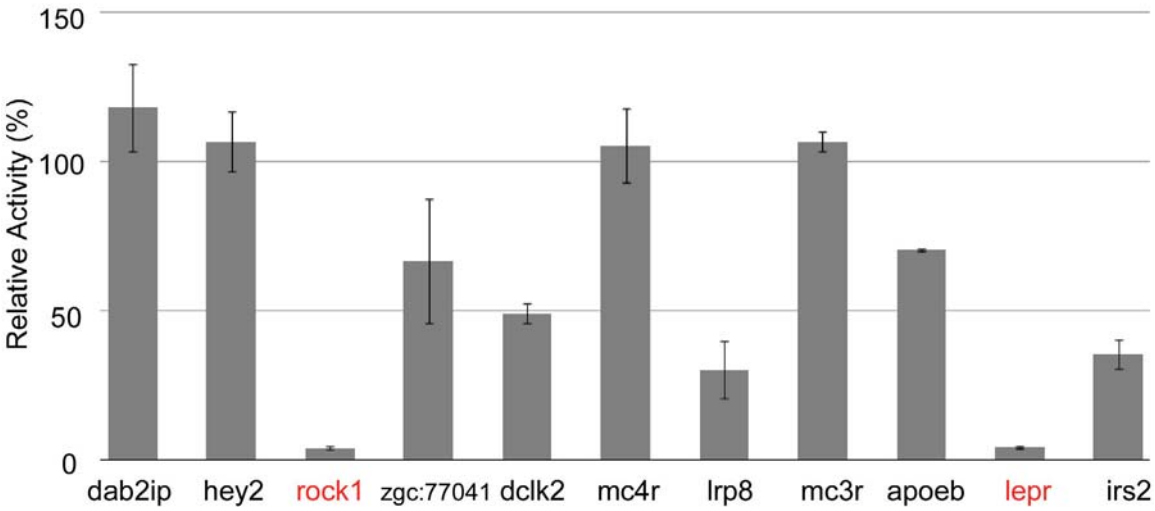
The specificity determinants that were constant in the original libraries were replaced by other residue to expand the repertoire of targetable sequences. DNA binding specificity of new 2F-modules was determined using CV-B1H method and the logos were obtained using GRaMS modeling. (**a**) Examples of the influence of substitution of determinants at position 3 of F2 (shown in blue) on the specificity of the 2FM-25 2F-module. In three instances this results in a desired change in the specificity only at base 2, however in 2FM-1101 the introduction of Asp results in a change in the preference of base position 3, akin to the effects observed for the D20A mutation in Zif268[27]. (**b**) Substituting the N-terminal cap residues in F1 (RSD at positions -1, 1 and 2) with a QRG cap results in a concomitant change in base preference from G to A at the 6th base position without severely compromising the specificity for the junction sequence.

Supplementary Figure 8: Binding site specificities of ZFAs incorporated into each ZFN pair.



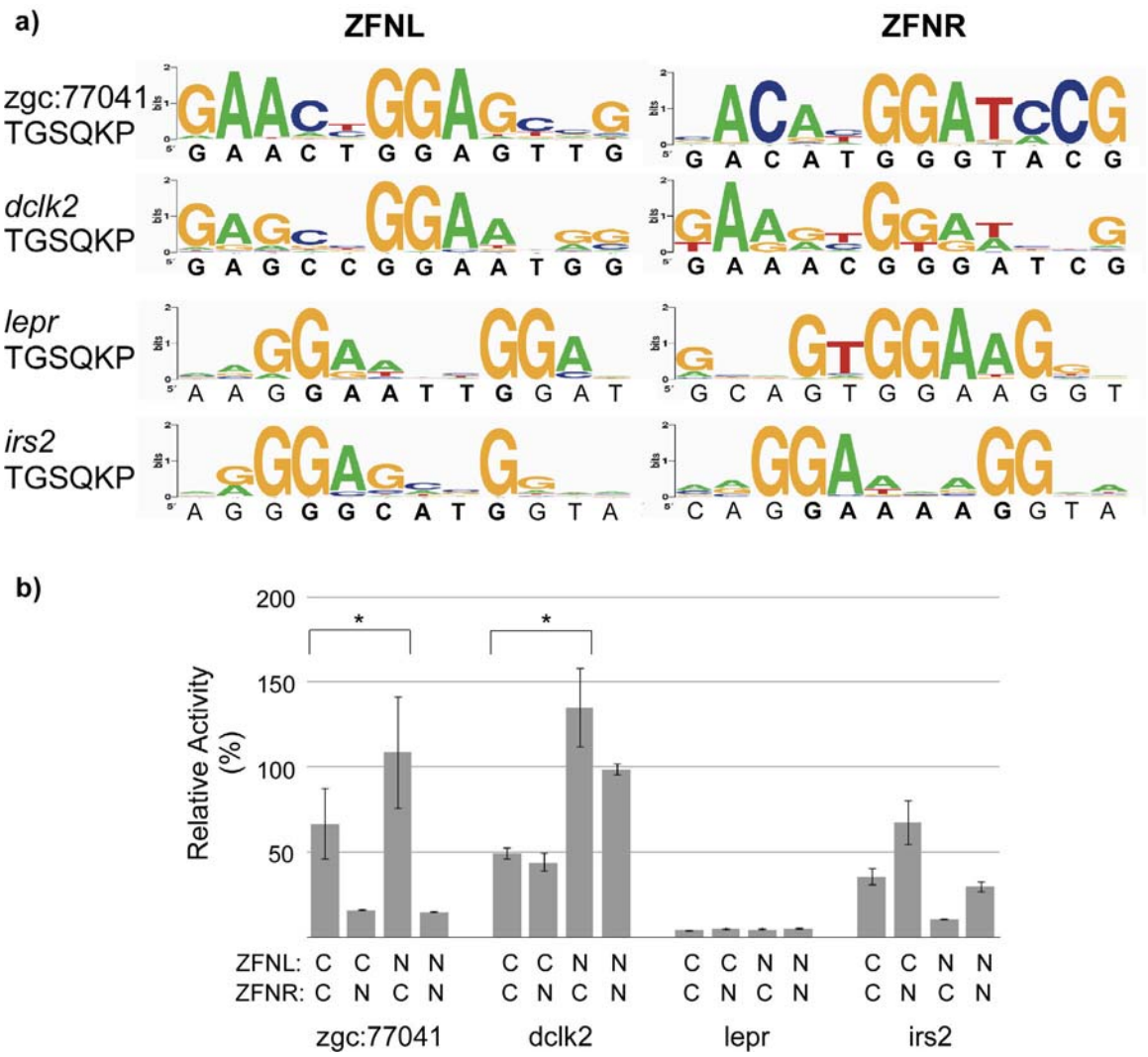
The binding site specificities for the ZFAs incorporated into ZFNs were determined via a B1H assay using the randomized 28bp library followed by Illumina[28]. The desired target sites are provided below each Sequence logo where the portion recognized by a 2F-module is highlighted in bold. The target gene is listed to the left of each ZFN pair.

Supplementary Figure 9: Assessment of ZFN activity using the yeast based chromosomal reporter assay.



The test ZFN target site along with the target site for the positive control ZFN was integrated into the yeast genome where ZFN activity is measured by the reconstitution of α-galactosidase activity[21]. ZFN expression was induced by treating yeast cells with galactose for 30 minutes. The activity relative to the positive control ZFN pair that yields ~10% lesion frequency in zebrafish is displayed as a mean of three experiments. Bars represent standard deviation. The rock1 and lepr ZFNs (shown in red) were inactive based on comparison to a GFP control.

Supplementary Figure 10: Influence of non-canonical linker on ZFN specificity and activity.



(**a**) For 4 4F-ZFNs, the canonical linker ('C') between the 2nd and the 3rd finger was replaced by a non-canonical linker ('N'; TG*S*QKP). DNA binding specificities were determined for the modified ZFNs using the B1H assay as previously described. (**b**) The activity of the modified ZFNs with non-canonical linker was assessed using the yeast based reporter assay. The activities are relative to the positive control ZFN activity and is displayed as a mean of three experiments. Bars represent standard deviation and * indicates $P < 0.05$ as determined by paired student's *t*-test.

Supplementary Table 3: List of all ZFNs and their target sites.

Gene	Target site	ZFNL binding site	ZFNR binding site	Spacer length (bp)	non-GNN fingers	non-N-G junct-ions	ZFNL-F0	ZFNL-F1	ZFNL-F2	ZFNL-F3	ZFNR-F0	ZFNR-F1	ZFNR-F2	ZFNR-F3
dab2ip	GTCCGAGTCcctgtaGACATGGAC	**GACTCG**GAC	**GACATG**GAC	6	2	2		LKGNLTR	**RSDTLKG**	**DRCNLTR**		LKGNLTR	**RSDTLKQ**	**DKGNLTR**
hey2	AACCATACCgaccgtGGGGAACTG	**GGTATG**GTT	GGG**GAACTG**	6	2	2		TSGSLSR	**RSDTLKQ**	**AAGHLTR**		**RSDTLVE**	**QRGNLTR**	RSDHLTR
rock1	CGTGCCAGCtgctccGACTCGGCC	GCT**GGCACG**	**GACTCG**GCC	6	2	2		**RSDTLQE**	**TARNLTR**	HRQSLTR		DRSDLSR	**RSDTLKG**	**DRCNLTR**
zgc77041	CAACTCCAGTTCattttgGACATGGGTACG	**GAACTGGAGTTG**	**GACATGGGTACG**	6	4	4	**RSDTLKE**	**KGCNLTR**	**RSDTLVE**	**QRGNLTR**	**RSDTLKD**	**LKRHLTR**	**RSDTLKQ**	**DKGNLTR**
dclk2	CCATTCCGGCTCtcgggGAAACGGGATCG	**GAGCCGGAATGG**	**GAAACGGGATCG**	5	4	4	**RSDHLTQ**	**QRGNLTR**	**RSDTLKE**	**RSSNLTR**	**RSDTLKG**	**QRCHLTR**	**RSDTLTQ**	**QRGNLTR**
mc4r	TTCATCTGCAGCttggctGTAGCAGACTTG	GCTGCAGATGAA	GTAGCA**GACTTG**	6	1	1	QKCNLVR	HRNNLTR	QSGDLTR	HRQSLTR	**RSDTLKG**	**DRCNLTR**	QSGDLTR	QSGALTR
lrp8	GCCTGCCCCGACagcatgGAACTGGGCCCG	GTCGGGGCAGGC	**GAACTGGGCCCG**	6	2	2	EKSHLTR	QSGDLTR	RSDHLTR	DRSALAR	**RSDTLMV**	**DRSHLTR**	**RSDTLVE**	**QRGNLTR**
mc3r	CAGCTCCACGTCagcgtaGCTGCATGAAGG	**GACGTGGAGCTG**	GCTGCATGAAGG	6	3	3	**RSDTLKE**	**RSSNLTR**	**RSDTLER**	**ESGNLTR**	RSDHLTQ	QSSHLTQ	QSSHLTQ	HRQSLTR
apoeb	AACCTACGCACCtctctGGAGGGCCGTGT	**GGTGCG**TAGGTT	GGA**GGGCCG**TGT	5	3	3	TSGSLSR	RSDNLTQ	**RSDTLRR**	**IRFHLTR**	LRHHLVG	**RSDTLKE**	**RSSHLTR**	QRGHLTR
lepr	ATCCAATTCCTTgcttcaGCAGTGGAAGGT	AAG**GAATTG**GAT	GCAGTGGAAGGT	6	2	1	TSGNLTR	**RSDTLKG**	**QRCNLTR**	RSDNLTQ	CAHHLTR	QKCNLVR	RSDALTR	QRSTRKR
irs2	TACCATGCCCCTctgtatCAGGAAAAGGTA	AGG**GGCATG**GTA	CAG**GAAAAG**GTA	6	4	2	QSGALTR	**RSDTLKE**	**ARRNLTR**	RSDHLTQ	QSGALTR	**RSDNLTQ**	**QRGNLTR**	RSDNLSE

For each ZFN target site, the ZFNL and ZFNR sites are shown in uppercase letters whereas the spacer sequences are shown in lowercase letters. The number of non-GNN and non-N-G junctions in each target site is provided. Also the recognition helix sequences (-1, 1, 2, 3, 4, 5, 6) for each ZFN are provided with the sequences of 2F-modules highlighted in bold.

Supplementary Table 4: Analysis of ZFN-induced lesions in zebrafish.

Gene	5p ZFP binding site	3p ZFP binding site	Spacer length (bp)	Number of Sequences with Indels	Number of wild type sequences	Lesion Frequency (%)	Most frequent Deletion	Most Frequent Insertion
dab2ip	**GACTCG**GAC	**GACATG**GAC	6	26703	334851	8.0	9bp (9198)	4bp (1471)
hey2	**GGTATG**GTT	GGG**GAACTG**	6	3438	552924	0.6	4bp (706)	4bp (1234)
rock1	GCT**GGCACG**	**GACTCG**GCC	6	191	384243	0.0	3bp (182)	None
*zgc77041**	**GAACTGGAGTTG**	**GACATGGGTACG**	6	49640	317017	15.7	9bp (7255)	2bp (8403)
*dclk2**	**GAGCCGGAATGG**	**GAAACGGGATCG**	5	2370	212738	1.1	2bp (656)	4bp (164)
mc4r	GCTGCAGATGAA	GTAGCA**GACTTG**	6	128638	998060	12.9	5bp (31856)	4bp (12193)
lrp8	GTCGGGGCAGGC	**GAACTGGGCCCG**	6	53297	732947	7.3	9bp (6780)	4bp (3534)
mc3r	**GACGTGGAGCTG**	GCTGCATGAAGG	6	24520	792371	3.1	5bp (5209)	4bp (5012)
apoeb	**GGTGCG**TAGGTT	GGA**GGGCCG**TGT	5	11180	396708	2.8	2bp (3507)	2bp (185)
*lepr**	AAG**GAATTG**GAT	GCAGTGGAAGGT	6	12264	1412846	0.9	4bp (8617)	4bp (266)
*irs2**	AGG**GGCATG**GTA	CAG**GAAAAG**GTA	6	2634	742945	0.4	6bp (969)	4bp (235)

ZFN target sites and the genes are shown. ZFNL and ZFNR sites are given wherein the 6bp subsites for the 2F-modules are represented in bold. Lesion frequencies and the most frequent insertion and deletion are shown where the number in parentheses shows their frequency. An asterisk indicates targets where a non-canonical linker (TG**S**QKP) between the second and the third finger was employed to increase ZFN activity, where the position of the non-canonical linker is underlined in each half-site where it is present.

Supplementary Table 5: Influence of non-canonical linker on ZFN activity in zebrafish.

	Lesion Frequency (%)				
	TGQKP	TGQKP	TG***S***QKP	TG***S***QKP	ZFNL Linker
Target gene	TGQKP	TG***S***QKP	TGQKP	TG***S***QKP	ZFNR Linker
zgc77041	4.4	4.1	12.3	15.7	
dclk2	0.1	0.5	0.3	1.1	

The lesion frequencies (in %) in zebrafish are shown for different combinations of ZFNL and ZFNR with canonical (TGQKP) and non-canonical (TGSQKP) of *zgc77041* and *dclk2* ZFNs.

Supplementary Table 6: Founder rates for four ZFN target genes.

Gene Name	Number of ZFN injected Fish Screened	Number of Founders Identified	Size of insertions or deletions at target site (+/-bp)	ZFN Lesion Frequency in embryos
mc4r	9	2	-5, -5	**12.9**
lrp8	17	8	+5, -3, -7, -8, -10,-12, -21	**7.3**
mc3r	5	2	+4, -11	**3.1**
apoeb	11	3	-4, -37, +9	**2.8**

Supplementary Table 7: Metrics for comparison of different ZFN assembly systems.

	Gupta 1/2FM	**CoDA 2FM**	**Kim 1/2FM**	**Zhu 1FM**	**Kim 1FM**
Archive Reference	A	B	C	D	E
Number of **Unique** ZFN sites in zebrafish protein-coding exons (25090 unique genes Zv9.64)	608,081	110,629	8,645,342	182,698	n.d.
Fraction of zebrafish protein-coding genes containing ZFN site	95.0%	79.2%	98.8%	85.9%	n.d.
Average density of ZFN sites (# bp/site)	132	722	10	438	n.d.
Number of **Unique** ZFN sites in human protein-coding exons (20236 unique genes GRCh37.p5)	1,384,075	269,242	14,669,536	444,163	n.d.
Fraction of human protein-coding genes containing ZFN site	96.7%	92.2%	97.8	94.5%	n.d.
Average density of ZFN sites (# bp/site)	123	633	12	383	n.d.
	Tested ZFNs				
Number of ZFN pairs tested in Archive Reference	11	38	13	29	315 ^
Number "active" ZFNs	9	19	3	8	23
Percent active ZFNs	82%	50%	23%	28%	7%
Percent GNN modules in ZFNs	**64%**	**99%**	**63% ***	**86%**	**40%**
ZFNs sites with non-GNN finger (active)	11 (9)	2 (1)	(3) *	17 (2)	33 (8) ^
ZFN sites with non-N-G junctions (active)	11 (9)	1 (0)	(3) *	10 (1)	33 (8) ^

A = this manuscript
B = Sander, J. D. et al. *Nature methods* **8**, 67-69 (2011)
C = Kim, S. el al. *Nature methods* **8**, 7 (2011)
D = Zhu, C. *et al. Development* **138**, 4555-4564 (2011)
E = Kim, H. J., et al. *Genome Res* **19**, 1279-1288 (2009)
n.d. = not determined
* = only target sequences for successful ZFNs reported
^ = multiple zfn pairs were tested at each target site

Supplementary Table 8: Primer sequences for 2F-module library construction and specificity analysis.

Library Construction Oligos:
F1 library top oligo: CCTGCGACCGCCGCTTCTCCAGATCTGAyACnCTnvnsvnsCATATACGTATTCACAC
F1 3' complement bottom oligo: GCCGGTGTGAATACGTATATG
F1 5' complement bottom oligo: AGATCTGGAGAAGCGGCGGTCG
F2 library top oligo (His+3F2): CTGCATGAAGGCCTTCTCTnnwnnwnnwCAyCTnACACGTCACATCAGGACCCACAC
F2 library top oligo (Asn+3F2): CTGCATGAAGGCCTTCTCTnnwnnwnnwAAyCTnACACGTCACATCAGGACCCACAC
F2 3' complement bottom oligo: GCCGGTGTGGGTCCTGATGTGACGTGT
F1 5' complement bottom oligo: AGAGAAGGCCTTCAT

Cloning B1H-selected 2F modules into 3F F1-GCG constructs:
GCG-for: CCATGGTACCTCTAGACCC
GCG-rev: GGGCAAGCATACGGTTTTTCACCGGTATGA
2F-module for: GTGAAAAACCGTATGCTTGCCCTGTCGAGTC
2F-module rev: TTACTGTGCAGAGGATCCCCTCAGGTGGGTCCTGATGTGACG

Supplementary Table 9: Primer sequences for ZFN assembly.

Primer Name	Sequence (5' to 3')
F0Fn	CCCAGTCACGACGTTGTAAAACGGTACCAAGCCCTATAAATGTCCTGAATG
F0Rn	ACACGCGTATGGCTTCTCACCGGTGTGCGTA
F1Fn	TGAGAAGCCATACGCGTGTCCTGTCGAGTCCTGT
F1Rn	GCATTGAAACGGTTTTTGCCCTGTGTGAATC
F2Fn	GCAAAAACCGTTTCAATGCCGCATCTGCATG
F2Rn	ACAGGCGAAGGGCTTTTCTCCTGTGTGGGTG
F3Fn	AGAAAAGCCCTTCGCCTGTGACATCTGCGG
F3RnLRGS	AGCGGATAACAATTTCACACAGGATCCACGGAGGTGGATCTTGGTGTG
F3RnTGPGAAGS	AGCGGATAACAATTTCACACAGGATCCTGCAGCACCAGGGCCAGTGTGGATCTTGGTGTG
F1(noF0)Fn	CCCAGTCACGACGTTGTAAAACGGTACCCGCCCATATGCTTGCCC
2FM-F0Fn	CCCAGTCACGACGTTGTAAAACGGTACCAAACCGTATGCTTGCCCTGTC
2FM-F1Rn	GCATTGAAACGGTTTTTGCCCTGTGTGGGTCCTGATGTG
2FM-F1Fn	TGAGAAGCCATACGCGTGTCCTGTCGAGTCCTGTGAC
2FM-F2Rn	ACAGGCGAAGGGCTTTTCTCCTGTGTGGGTCCTGATGTG
2FM-F2Fn	GCAAAAACCGTTTCAATGCCCTGTCGAGTCCTGCGAC
2FM-F3RnLRGS	AGCGGATAACAATTTCACACAGGATCCACGGAGGTGGGTCCTGATGTG
2FM-F3RnTGPGAAGS	AGCGGATAACAATTTCACACAGGATCCTGCAGCACCAGGGCCAGTGTGGGTCCTGATGTG
2FM-F1(noF0)Fn	CCCAGTCACGACGTTGTAAAACGGTACCAAACCGTATGCTTGCCCTG
2FM-F0-QRG(**X**)Fn	CCGTATGCTTGCCCTGTCGAGTCCTGCGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-F1-QRG(**X**)Fn	TGAGAAGCCATACGCGTGTCCTGTCGAGTCCTGTGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-F2-QRG(**X**)Fn	GCAAAAACCGTTTCAATGCCCTGTCGAGTCCTGCGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-F1(noF0)-QRG(**X**)Fn	TTGTAAAACGGTACCAAACCGTATGCTTGCCCTGTCGAGTCCTGCGACCGCCGCTTCTCCcagcgcggc**NNN**CT
2FM-NT-in-Fn	CGTTGTAAAACGGTACCAAACCTTATGCTTGCCCTGTC
2FM-NT-out-Fn	ACGTTGTAAAACGGTACCAAACCT
2FM-CT-out-Rn	AACAATTTCACACAGGATCCACG

NOTE: For QRG(X) primers in place of NNN use ACN if X (F1 position 3) is Thr, use AAY if X (F1 position 3) is Asn, use CAC if X (F1 position 3) is His.

Supplementary Table 10: Sequences of the genotyping primers used for lesions detection in zebrafish embryos.

Gene	Genotyping assay used	Genotyping Forward primer for RFLP analysis or Cell assay (5' to 3')	Genotyping Reverse primer for RFLP analysis or Cell assay (5' to 3')	Forward Primer for Illumina Sequencing (5' to 3')	Reverse Primer for Illumina Sequencing (5' to 3')	Restriction Enzyme site used for Illumina sample preparation	5' Tag for Counting InDels	3' Tag for Counting InDels
dab2ip	RFLP - *SfcI*	CAGGGTACCACTTCTCCAC	CAGCCTATATGCCCGCAC	CGGCATACGAGCTCTTCCGATCTCCACTTCTCCACCAGCTGC	GCGGTCCAGAGCGGTACCGTCC	*Hpy188I*	GTACCGTCCAT	TCGGAC
hey2	RFLP - *XcmI*	CAGCCCCAGCGTTACAGC	CTGCTGACCGAAGCAGGC	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTCTGCTGACCGAAGCAGGC	CTGCTGACCGAAGCAGGC	*Hpy166II*	AACCAT	GAACTG
rock1	RFLP - *Hpy188I*	GAGATGGTGGAGTCTTTCTC	GTATTGTCTGCAGGGAGTCTC	CAAGCAGAAGACGGCATACGAGCTCTCCGATCTGAGATGGTGGAGTCTTTCTC	GTATTGTCTGCAGGGAGTCTC	*StyI* HF	CAAGGCCGA	GGCACG
zgc77041	Cel1	GGAGCAAATGTAAGGCAAACC	ATTGTTACATTTTCAAAGATGCTG	GGAGCAAATGTAAGGCAAACC	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTATTGTTACATTTTCAAAGATGCTG	*SnaBI*	GTACCCAT	CTGGAG
dclk2	RFLP - *AvaI*	GACACGGCGTACACAAGCC	GAACCAGCGCTATCACTTAAG	CAAGCAGAAGACGGCATACGAGCTCTCCGATCTGACACGGCGTACACAAGCC	GGCAGCGGCCGGCTCCC	*NaeI*	GGCTCCCATTCCGG	ACGGGA
mc4r	Cel1	CAGCCTCCTGGAGAACATCC	TCACGGTTGGTCAGGTTGC	CAAGAACCTACATTCCCCTATGAACTTCTTC	CGGCATACGAGCTCTTCCGATCTCATAGAGTCAAACACGTTGTC	*XmnI*	TCTTCATCTGC	GCAGAC
lrp8	RFLP - *MwoI*	GAGGCTGTGAGTATCTGTGC	GAAAGTGTGCAGTATGAGTAAAC	CACTCACCCAAATACACCGGTACCTGCC	CGGCATACGAGCTCTTCCGATCTCCAAATTTTACTCACAACAATG	*Acc65I*	GTACCTGCCCC	CTGGGC
mc3r	Cel1	TTCTTCTCGCCAGACTTCAC	CACCAGTAGAATGAGGTGGAG	CCCGGCGGCTCCTGGTGCTGGGTACCCAGCTC	CGGCATACGAGCTCTTCCGATCTGCAGAGGCAGAGCGGATG	*Acc65I*	GTACCCAGCTCCAC	GCATGA
apoeb	RFLP - *Hpy188III*	CCACCCAGAAACTGGGCGC	GGTAAGTGTGGAGCTCTTAAGC	GAAGCTGGAGGAGACAGCCGGGTACCTAC	CGGCATACGAGCTCTTCCGATCTGGTAAGTGTGGAGCTCTTAAGC	*Acc65I*	GTACCTACGC	GGGCCG
lepr	Cel1	AGGTGGACCGGCACACAAC	CACAATTCTTACAAACATCAC	CGGCATACGAGCTCTTCCGATCTGGCGCACCTGTCAATCTGC	CATTACACCAACAAAAGAGACCAGGTACCTTCC	*Acc65I*	GTACCTTCCAC	GAATTG
irs2	RFLP - *Hpy188III*	GTTCACACTCTTCTAAACTGTG	CCTTTTGAAACCCCCTGGTTG	GTTTTCTCAACGAACAGAGAAAGGTACCATG	CGGCATACGAGCTCTTCCGATCTCCTTTTGAAACCCCCTGGTTG	*Acc65I*	GAATGTACCATGCC	GAAAAG

For the analysis by Illumina sequencing the restriction enzymes used for truncating the PCR product near the ZFN site for adaptor ligation are indicated. The unique 5’ and 3’ tags employed for distinguishing and counting sequences containing InDels for each target site are listed.

Supplementary Discussion 1:

Comparison to previously described Finger Archives:

A number of different systems have been described for assembling Zinc Finger Arrays (ZFAs) from one-finger (1F)[1-10] or two-finger (2F)[11, 12] archives. These archives display diversity in the number of fingers, the base composition of their recognition sequences and the strategies for their assembly. The quality of many of these archives have been assessed on a moderate to large scale through characterization of the constructed ZFAs[13-16] or assessment of the activity of ZFNs containing these ZFAs in cell lines or *in vivo*[9-12, 15] **(Supplementary Table 7)**. The likelihood of any given ZFN constructed from the different archives being active varies, where the rates for ZFNs derived from 1F archives are below 30% and those from two-finger archives are generally higher, ranging between 24 and 82% (**Supplementary Table 7**).

The finger archive that is most advantageous for a user to employ for constructing ZFNs may depend not only on its potential success rate, but also the availability of target sites near a specific genomic position for applications where site specific modification is desired. To assess the general utility of these archives for gene editing in vertebrates, we compared the number of potential target sites in protein-coding exons within the zebrafish (Zv9) and human (GRCh37.p5) genome, as well as the overlap of target sites between these different archives. We focused our comparisons on the two-finger module archives because they generally have higher success rates (Gupta 1/2FM, CoDA 2FM[11] and Kim 1/2FM[12]; **Supplementary Table 7)**. The combination of our 2FM archive with our previously described 1FM archive (Zhu 1FM[10]) expands the targeting density of our original archive by 3-fold, while creating ZFNs with promising activity. This combined archive has a ~5-fold higher density of ZFN sites than the CoDA archive, with an average of one unique ZFN site every ~140 bp. The Kim 1/2FM archive has the highest targeting density of the three archives due to the large number of 2F-modules it contains with an average rate of one unique ZFN site every 10 bp, albeit with a lower overall success rate.

While the targeting density provides one important reflection on the utility of an archive, its flexibility can be inferred from the composition of target sequences evaluated in studies validating its efficacy. While the CoDA archive contains a combination of GNN and non-GNN finger sets (61 non N-G junction 2F-modules), the ZFNs that were evaluated by Sander and colleagues were composed almost entirely of GNN finger sets (99%). This may reflect the fact that only 3 of 10 ZFAs containing non N-G junction 2F-modules were functional in their bacterial activity-assay[11], which was used as a prescreen for choosing modules employed in their ZFNs. Our characterized ZFNs contain a more diverse set of fingers where roughly two-thirds (64%) were GNN finger sets (**Supplementary Figure 1**). Of the 39 CoDA ZFNs that were evaluated, only one target contained a finger set recognizing a non-N-G junction between fingers, whereas all 11 of our evaluated ZFNs contained non-N-G junction, demonstrating the breadth of sequences that can be effectively targeted using

our system (**Supplementary Table 7**). For the ZFNs evaluated in the Kim 1/2FM archive analysis, only the sequences of the three active ZFNs were reported limiting the comparisons that can be drawn between it and the other archives[12].

The ZFNs evaluated in our study were chosen to serve a number of different goals. Foremost, ZFNs were chosen to assay different numbers of fingers per ZFN and different mixtures of 2F- and 1F-modules, where all of the ZFN pairs contain at least one non-GNN finger and one non-N-G interface. While there is some bias in the composition of the fingers comprising the ZFNs that were evaluated, many of the choices were driven by the desire to inactivate specific target genes in zebrafish that if successful could potentially yield useful disease models (atherosclerosis (*apoeb*, *lrp8*), obesity (*lepr*, *mc3r*, *mc4r*), and diabetes (*irs2*)). Nonetheless, we believe that the 82% success rate achieved in this sample set will not be completely representative of ZFNs constructed from this archive. For example, this archive is a mixture of 2F-modules and 1F-modules, where about 30% of the identified ZFNs are composed of only 1F-modules. Based on our prior evaluation, we would anticipate only about one-fourth of these ZFNs composed entirely of 1F-modules to be active[10]. To aid the user in the choice of ZFNs for specific target genes we have constructed a scoring function that weights the 2F-modules based on their specificity in the B1H system. This has been integrated with our previously described 1F-module scoring function[10] in the web-based tool described below.

Supplementary Discussion 2:

Modifications to our ZFN sets to increase their activity *in vivo*:

In the course of these studies we have evaluated four finger ZFNs with and without a disrupted linker (TGSQKP) between pairs of fingers to assess the effects of this modification on ZFN activity. Pioneering work by Choo and colleagues[17-19] demonstrated the potential utility of contemplating recognition by zinc fingers as two finger units. The effective use of disrupted linkers can be found in the ZFNs employed in a large number of publications from Sangamo BioSciences[20-22]. Based on their preferential use of disrupted linkers in the majority of their ZFNs, we infer that this modification provides mechanistic advantages. We have independently examined the effect of utilizing the TGSQKP linker between pairs of fingers, and in two instances (*zgc77041* & *dclk2*) found clear benefits when using this modified linker (**Supplementary Table 5**). This linker modification is not required for ZFN function, but may be beneficial for function in some instances. Consequently we have used this modification in two other ZFNs (*lepr* & *irs2*) that displayed moderate activity in our initial analysis of somatic lesion frequency (**Supplementary Figure** 10).

In our ZFNs we have employed a four amino acid linker (LRGS) between the ZFA and the FokI domain described by Cathomen and colleagues[23] for ZFN sites with a five or six bp gap between the half-sites. Our website also discovers

potential ZFN sites that contain a seven bp gap. For these sites we recommend the use of a eight amino acid linker (TGPGAAGS), as Cathomen and colleagues[23] have demonstrated that longer linkers provide improved efficiency for the seven bp gap between ZFN recognition sites.

Supplementary Discussion 3:

Description of web interface for identification of ZFN sites within query sequences:

Our website (http://pgfe.umassmed.edu/ZFPmodularsearchV2.html) allows a user to input a single sequence or multiple sequences in FASTA format for the identification of sites that can be targeted with ZFN constructed from our single finger[10] and two finger archives. This website is completely anonymous; no login is required to use the interface and no user information is saved from a submitted query. Users can choose from multiple formats (browser, text file, word document or excel file) for the output from the initial analysis. Potential ZFN sites are ranked based on their overall score (the scoring metric is described in Methods). Additional information is provided regarding the position of the site within each input sequence, the target sequence for each ZFN monomer, the gap separating these sites, whether there is a restriction enzyme (RE) site within this spacer, and the identity of the finger modules that comprise each ZFA monomer. Each ZFA has four potential fingers (F3, F2, F1, & F0), where the fourth finger (F0 in our nomenclature) if absent is indicated by 'XXX'. Modules appearing in UPPERCASE are from the single finger module archive[10], while modules appearing in lowercase are from the archive described in this study and will occur in pairs (e.g. grn & nyr pairs). Within the browser output, more detailed information on each ZFN can be output using a button at the end of each column. Again there is a choice of output formats, where for each ZFN additional pertinent information is provided: the ZFA amino acid and DNA sequences for gene synthesis, modules IDs within our archive for PCR-based construction, recognition helix sequences, and information on RE sites that overlap with the spacer region for genotyping. The DNA sequences that are provided include Acc65I and BamHI sites at their termini for cloning into our pCS2 vectors (DD/RR or EL/KK versions) that are available from Addgene. A detailed protocol for the assembly of the ZFA can be downloaded from the home page of the website, but we recommend gene synthesis for the construction of ZFAs due to its affordability and ease.

Supplementary Methods:

Description of $GRaMS_c$ and W log-odds motif construction algorithms used in this study.

$GRaMS_c$: In the original implementation of GRaMS[24], nonlinear regression was employed to parameterize a model consisting of a PWM and a parameter, μ, which describes the degree of saturation of each binding site due to the free concentration of the TF. We used the same model for this study, but re-arranged the objective function. Instead of fitting to the observed growth rate of each site, we fit to the observed counts per site. We call this version of the program $GRaMS_c$.

Many of the Zif268 mutants in this study are more specific than wild type Zif268 for their preferred sequences. We found in practice that for very specific proteins that resulted in only a handful of sites with growth rates significantly larger than the median growth rate, the most accurate recognition model was generated by re-arranging the GRaMS objective function to fit directly to the observed counts per site. Otherwise, when there were very few appreciably enriched sites (few informative data points), there was a tendency to over fit to the noise in the growth rate data. We found it also helped to adjust M, the maximum observed growth rate by a factor of 1.02. This prevented a single site from dominating the motif completely when only very few sites had growth rates appreciably larger than the median growth rate and one site clearly had a much higher growth rate than the other enriched sites. The following equations describe the adjusted model. The observed growth rate (r_i), or enrichment, of each site is given by:

$$r_i = log_2\left(\frac{f_i(t)}{f_i(0)}\right)/t \qquad (1)$$

where t indicates the duration of the selection experiment, i is an index over all 4^6 6mers, $f_i(t)$ is the frequency of site i at time t, and $f_i(0)$ is the initial frequency of site i at time 0. The growth rate of a site, S_i, is a sigmoid function of μ, the chemical potential of the TF, and the Gibbs free energy of the TF binding to the site as well as the maximum and minimum possible growth rates:

$$r_i = \frac{M-D}{1+e^{S_iW-\mu}} + D \qquad (2)$$

where W is the PWM and S_i*W yields the Gibbs free energy of binding to S_i. The variables M and D determine the upper and lower plateaus of the sigmoid curve. M is set to the maximum observed growth rate times a scalar of 1.02, and D is set to the median observed growth rate. The total number of times each site was observed is modeled by the following equation:

$$c_i = N_F f_i(0) 2^{\left(\frac{M-D}{1+e^{S_iW-\mu}}+D\right)t} \qquad (3)$$

where N_F is the total number of sequenced sites. The Levenberg-Marquardt algorithm was used to fit the parameters of the PWM and the μ parameter. Regularization was used to prevent over fitting.

W log-odds: The W log-odds ('W' stands for 'word based' log-odds) method more accurately reflects our knowledge of the initial frequency of each 6mer in the library than a simple log-odds weight matrix. Generally, the following formula is used to compute log-odds PWMs:

$$W_{bj} = -\log\left(\frac{P_{bj}}{P_b}\right) \qquad (4)$$

where W_{bj} is the log-odds matrix, P_{bj} is the probability after selection of observing base b at position j in the binding site, and P_b is the initial probability of observing base b before selection. Because the initial frequency of each 6mer binding site prior to selection was known from deep sequencing of the initial counter selected library, the enrichment of each site after selection was calculated directly. The enrichment ratio of the i^{th} site is given by the equation

$$\frac{f_i(t)}{f_i(0)} \qquad (5)$$

where t is the final time or duration of the selection experiment, $f_i(t)$ is the final frequency at time t, $f_i(0)$ is the initial frequency at time 0, and i is an index over all 4^6 6mers. A site's enrichment ratio can be thought of as the K_a of that site. A pseudo count of one was added to all final and initial counts when calculating the initial and final frequencies. The sum of all the enrichment ratios for all 6mers containing base b at position j was used to calculate each element of the W log-odds matrix:

$$W_{bj} = -\log\left(\sum_{i=1}^{4096} \delta_{S_{ij},B_b} \frac{f_i(t)}{f_i(0)}\right) \qquad (6)$$

where S_{ij} indicates the base at position j of site i, b is an index over the four nucleotide bases, B_b returns base b and $\delta_{x,y}$ is the Kronecker delta function which returns 1 if the bases x and y are identical and zero otherwise. For example, to determine the energy contribution of an A in the first position of the binding site ($W_{1,1}$) the set of all 1024 6mers that have an A at position 1 was determined and the enrichment ratios for all of these sites were summed and the negative of the log of this value was taken.

Supplementary References:

1. Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D. & Barbas, C.F., 3rd *J Biol Chem* **276**, 29466-29478 (2001).
2. Segal, D.J., Dreier, B., Beerli, R.R. & Barbas, C.F., 3rd *Proc Natl Acad Sci U S A* **96**, 2758-2763 (1999).
3. Dreier, B., Segal, D.J. & Barbas, C.F., 3rd *J Mol Biol* **303**, 489-502 (2000).
4. Liu, Q., Xia, Z., Zhong, X. & Case, C.C. *J Biol Chem* **277**, 3850-3856 (2002).
5. Bae, K.H. *et al. Nature biotechnology* **21**, 275-280 (2003).
6. Dreier, B. *et al. J Biol Chem* **280**, 35588-35597 (2005).
7. Mandell, J.G. & Barbas, C.F., 3rd *Nucleic Acids Res* **34**, W516-523 (2006).
8. Wright, D.A. *et al. Nature protocols* **1**, 1637-1652 (2006).
9. Kim, H.J., Lee, H.J., Kim, H., Cho, S.W. & Kim, J.S. *Genome Res* **19**, 1279-1288 (2009).
10. Zhu, C. *et al. Development* **138**, 4555-4564 (2011).
11. Sander, J.D. *et al. Nature methods* **8**, 67-69 (2011).
12. Kim, S., Lee, M.J., Kim, H., Kang, M. & Kim, J.S. *Nature methods* **8**, 7 (2011).
13. Segal, D.J. *et al. Biochemistry* **42**, 2137-2148 (2003).
14. Carroll, D., Morton, J.J., Beumer, K.J. & Segal, D.J. *Nature protocols* **1**, 1329-1341 (2006).
15. Ramirez, C.L. *et al. Nature methods* **5**, 374-375 (2008).
16. Sander, J.D., Zaback, P., Joung, J.K., Voytas, D.F. & Dobbs, D. *Nucleic Acids Res* **37**, 506-515 (2009).
17. Isalan, M., Klug, A. & Choo, Y. *Biochemistry* **37**, 12026-12033 (1998).
18. Isalan, M., Klug, A. & Choo, Y. *Nature biotechnology* **19**, 656-660 (2001).
19. Moore, M., Klug, A. & Choo, Y. *Proc Natl Acad Sci U S A* **98**, 1437-1441 (2001).
20. Perez, E.E. *et al. Nature biotechnology* **26**, 808-816 (2008).
21. Doyon, Y. *et al. Nature biotechnology* **26**, 702-708 (2008).
22. Hockemeyer, D. *et al. Nature biotechnology* **27**, 851-857 (2009).
23. Handel, E.M., Alwin, S. & Cathomen, T. *Mol Ther* **17**, 104-111 (2009).
24. Christensen, R.G. *et al. Nucleic Acids Res* **39**, e83 (2011).
25. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. *Genome Res* **14**, 1188-1190 (2004).
26. Zhao, Y., Granas, D. & Stormo, G.D. *PLoS Comput Biol* **5**, e1000590 (2009).
27. Miller, J.C. & Pabo, C.O. *J Mol Biol* **313**, 309-315 (2001).
28. Gupta, A., Meng, X., Zhu, L.J., Lawson, N.D. & Wolfe, S.A. *Nucleic Acids Res* **39**, 381-392 (2010).