Supplementary Table 1.

Sequences of OPEN-selected two-finger modules, B2H *lacZ* activity, expected binding sites and B1H determined specificity. – See attached Excel sheet.

Supplementary Table 2.

Archive of 1209 one-finger and 678 two-finger modules used to train our RF models. Each motif is represented as a PFM with the name of the clone, the amino acids present at positions -1, +2, +3 and +6, and the amino acids at positions -1 through 6 in the recognition helix. For two finger modules the amino acids in the N-terminal finger are listed first. – See attached txt file.

Supplementary Table 3.

Comparison of the mean and median mean squared error (MSE) values for the prediction of the ZFP specificities in Supplementary Figure 6.

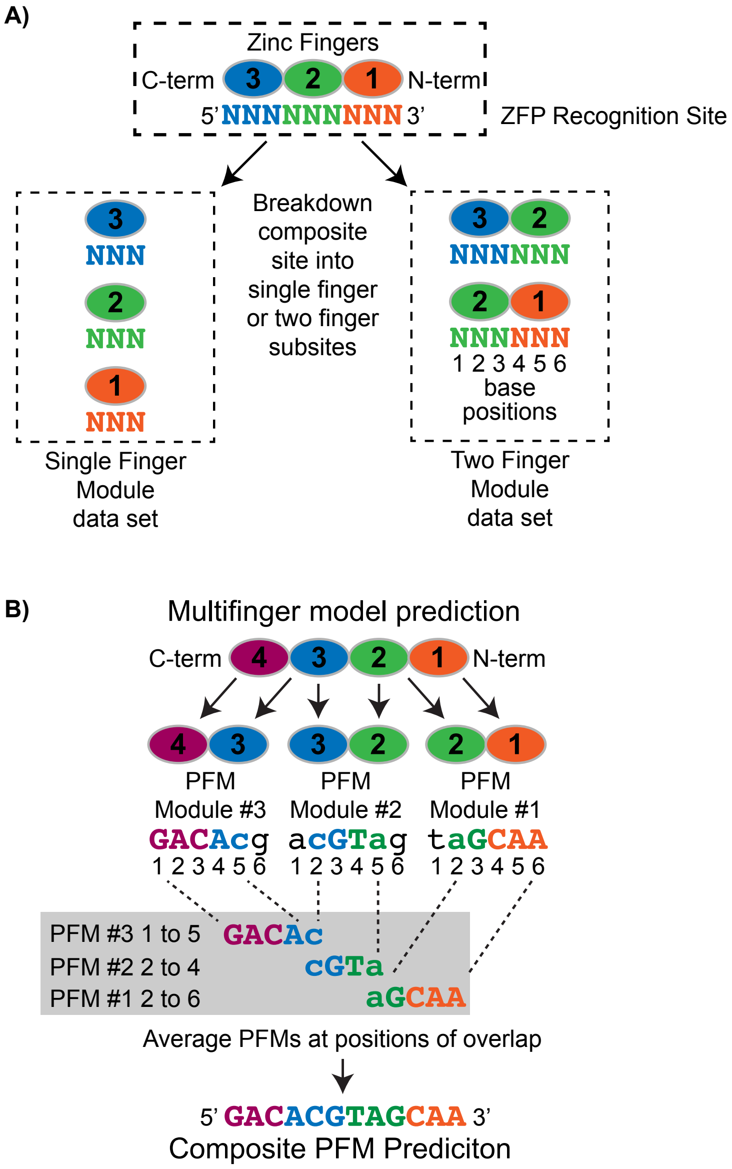
Model	**Mean MSE**	**Median MSE**
Average One-Finger & Multi-Finger	0.045	0.044
Multi-Finger	0.047	0.050
One-Finger	0.048	0.047

Supplementary Table 4.

Comparison of the mean squared error (MSE) values for the motifs for our average ZFModel prediction or the ZF_Princeton prediction (79) to the determined SELEX data for each ZFP (105).

ZFP IDs	Avg Model	ZF_Princeton
AAVS1_ZFN-L	0.063	0.073
AAVS1_ZFN-R	0.059	0.132
OCT4_ZFN1-L	0.044	0.064
OCT4_ZFN1-R	0.051	0.088
OCT4_ZFN2-L	0.016	0.039
OCT4_ZFN2-R	0.044	0.097
Pitx3_ZFN-L	0.013	0.044
Pitx3_ZFN-R	0.056	0.158

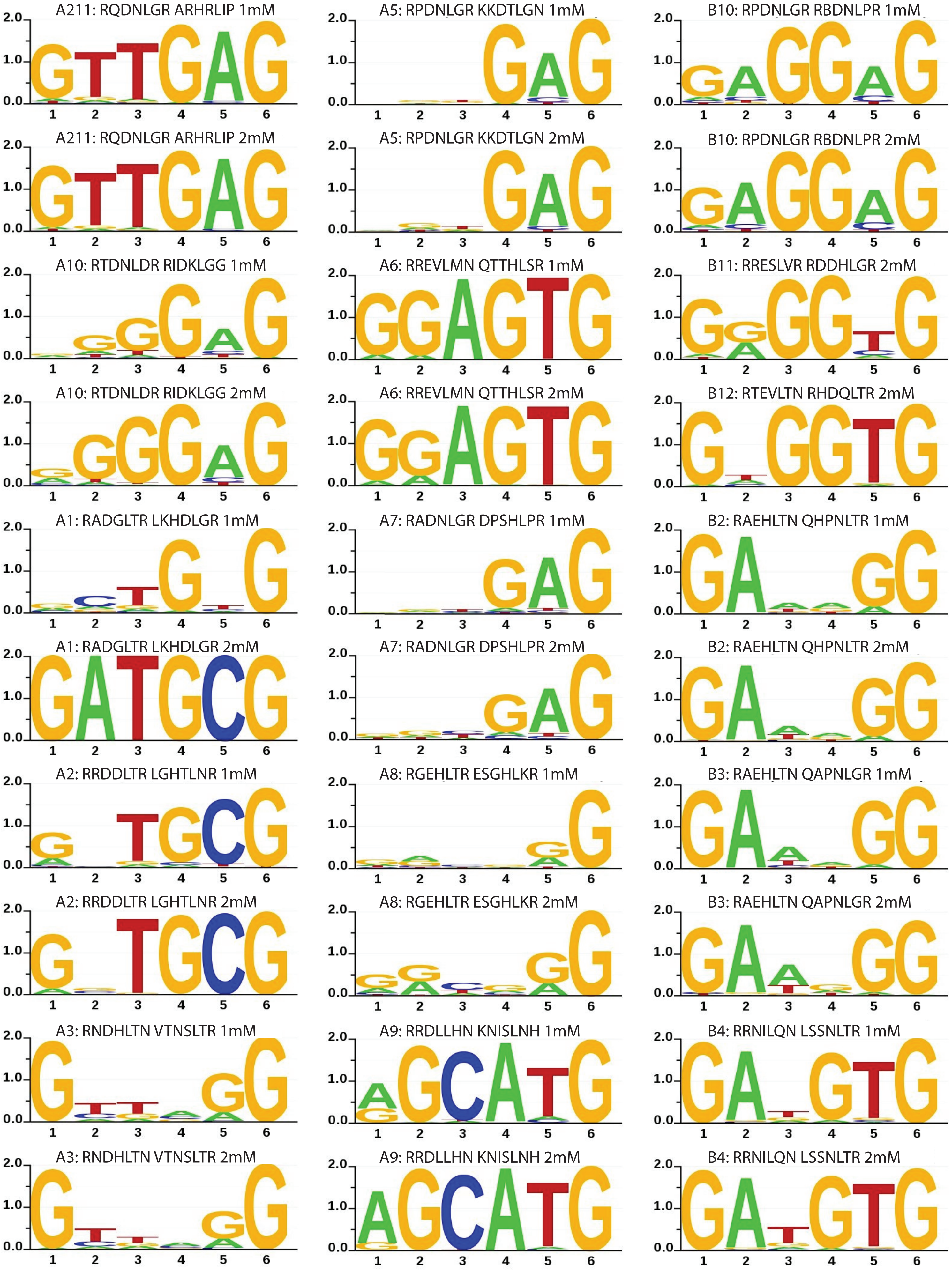
Supplementary Figure 1.



A) Schematic overview of the motif breakdown used to construct the one finger and two finger data sets for ZFModels training, where a three finger ZFP is broken down into 3 one finger modules, each with a 3 bp motif, or 2 two finger modules, each with a 6 bp motif. B) Overview of Multifinger motif construction method. Each ZFP is deconstructed into a set of overlapping two finger modules. Based on the specificity determinants that are present in each two finger module, ZFModels predicts a PFM. These PFMs are merged together at base position 2 and/or 5 of each motif, which is the base position contacted by the specificity determinant at position +3 of the recognition helix. The specificity at the positions of overlap (5 and 2 of overlapping motifs is generated by averaging the PFMs at this position.

Supplementary Figure 2.

Organized B1H motifs of the OPEN 2FMs. See following motifs.
DNA-binding specificities of the OPEN 2 Finger modules were determined using CV-B1H method. The clone ID, recognition helix sequence and stringency of the selection used to recover the binding sites that were incorporate into each GRaMS motif are displayed at the top of each image. Amino acid sequences are listed for the N-terminal finger followed by the C-terminal finger (positions -1 through 6). Most of these modules were characterized at both 1 mM and 2 mM 3-amino triazole in the B1H system (85). One clone, R27-2, is marked with "*", where and the two motifs are very different at each stringency. Because of this discrepancy neither were included in the training data. R27-2 has the same sequence as R55-1, which we used for training, but both of these motifs are quite different.



A211: RQDNLGR ARHRLIP 1mM

A211: RQDNLGR ARHRLIP 2mM

A10: RTDNLDR RIDKLGG 1mM

A10: RTDNLDR RIDKLGG 2mM

A1: RADGLTR LKHDLGR 1mM

A1: RADGLTR LKHDLGR 2mM

A2: RRDDLTR LGHTLNR 1mM

A2: RRDDLTR LGHTLNR 2mM

A3: RNDHLTN VTNSLTR 1mM

A3: RNDHLTN VTNSLTR 2mM

A5: RPDNLGR KKDTLGN 1mM

A5: RPDNLGR KKDTLGN 2mM

A6: RREVLMN QTTHLSR 1mM

A6: RREVLMN QTTHLSR 2mM

A7: RADNLGR DPSHLPR 1mM

A7: RADNLGR DPSHLPR 2mM

A8: RGEHLTR ESGHLKR 1mM

A8: RGEHLTR ESGHLKR 2mM

A9: RRDLLHN KNISLNH 1mM

A9: RRDLLHN KNISLNH 2mM

B10: RPDNLGR RBDNLPR 1mM

B10: RPDNLGR RBDNLPR 2mM

B11: RRESLVR RDDHLGR 2mM

B12: RTEVLTN RHDQLTR 2mM

B2: RAEHLTN QHPNLTR 1mM

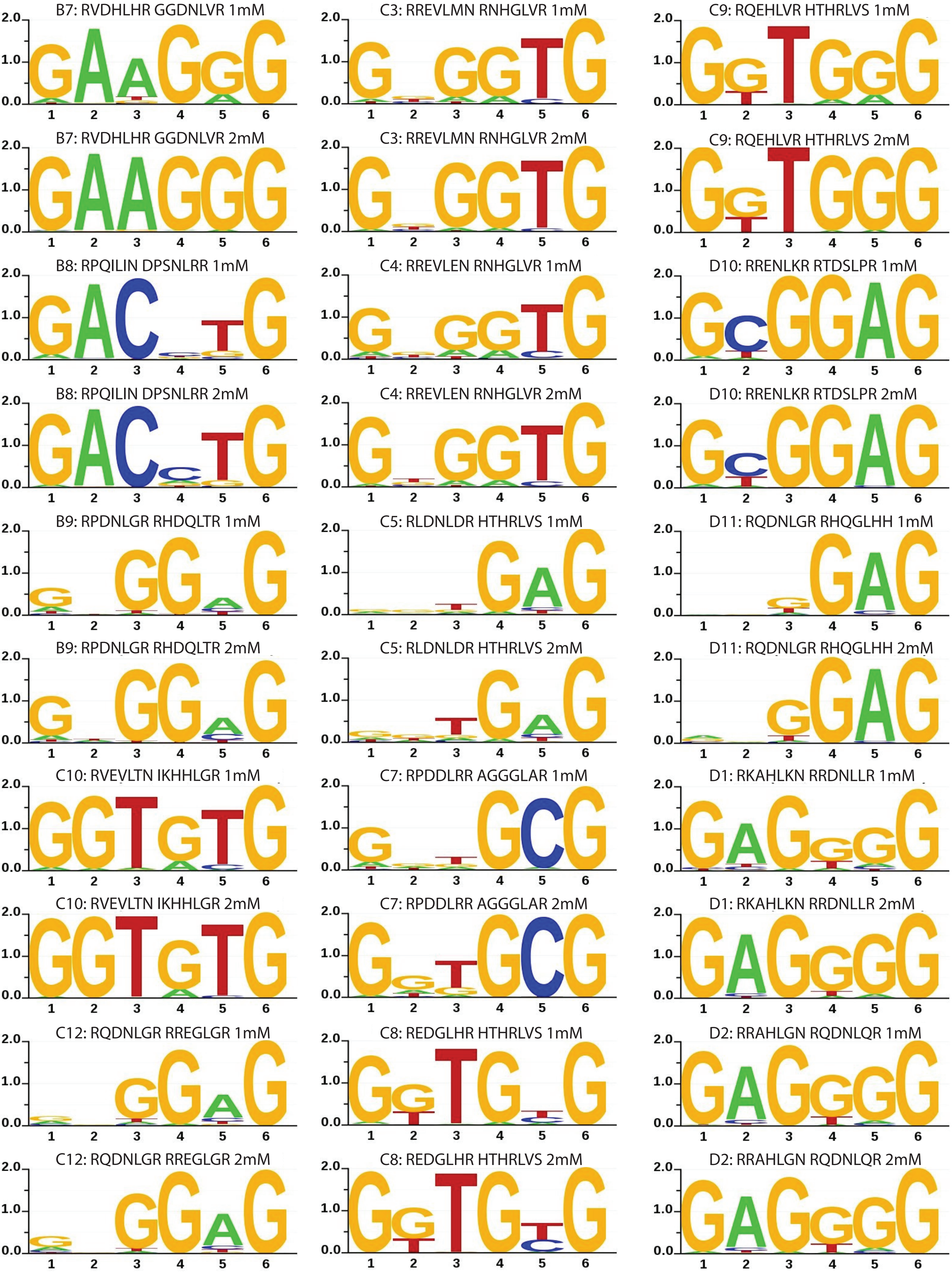
B2: RAEHLTN QHPNLTR 2mM

B3: RAEHLTN QAPNLGR 1mM

B3: RAEHLTN QAPNLGR 2mM

B4: RRNILQN LSSNLTR 1mM

B4: RRNILQN LSSNLTR 2mM



B7: RVDHLHR GGDNLVR 1mM

C3: RREVLMN RNHGLVR 1mM

C9: RQEHLVR HTHRLVS 1mM

B7: RVDHLHR GGDNLVR 2mM

C3: RREVLMN RNHGLVR 2mM

C9: RQEHLVR HTHRLVS 2mM

B8: RPQILIN DPSNLRR 1mM

C4: RREVLEN RNHGLVR 1mM

D10: RRENLKR RTDSLPR 1mM

B8: RPQILIN DPSNLRR 2mM

C4: RREVLEN RNHGLVR 2mM

D10: RRENLKR RTDSLPR 2mM

B9: RPDNLGR RHDQLTR 1mM

C5: RLDNLDR HTHRLVS 1mM

D11: RQDNLGR RHQGLHH 1mM

B9: RPDNLGR RHDQLTR 2mM

C5: RLDNLDR HTHRLVS 2mM

D11: RQDNLGR RHQGLHH 2mM

C10: RVEVLTN IKHHLGR 1mM

C7: RPDDLRR AGGGLAR 1mM

D1: RKAHLKN RRDNLLR 1mM

C10: RVEVLTN IKHHLGR 2mM

C7: RPDDLRR AGGGLAR 2mM

D1: RKAHLKN RRDNLLR 2mM

C12: RQDNLGR RREGLGR 1mM

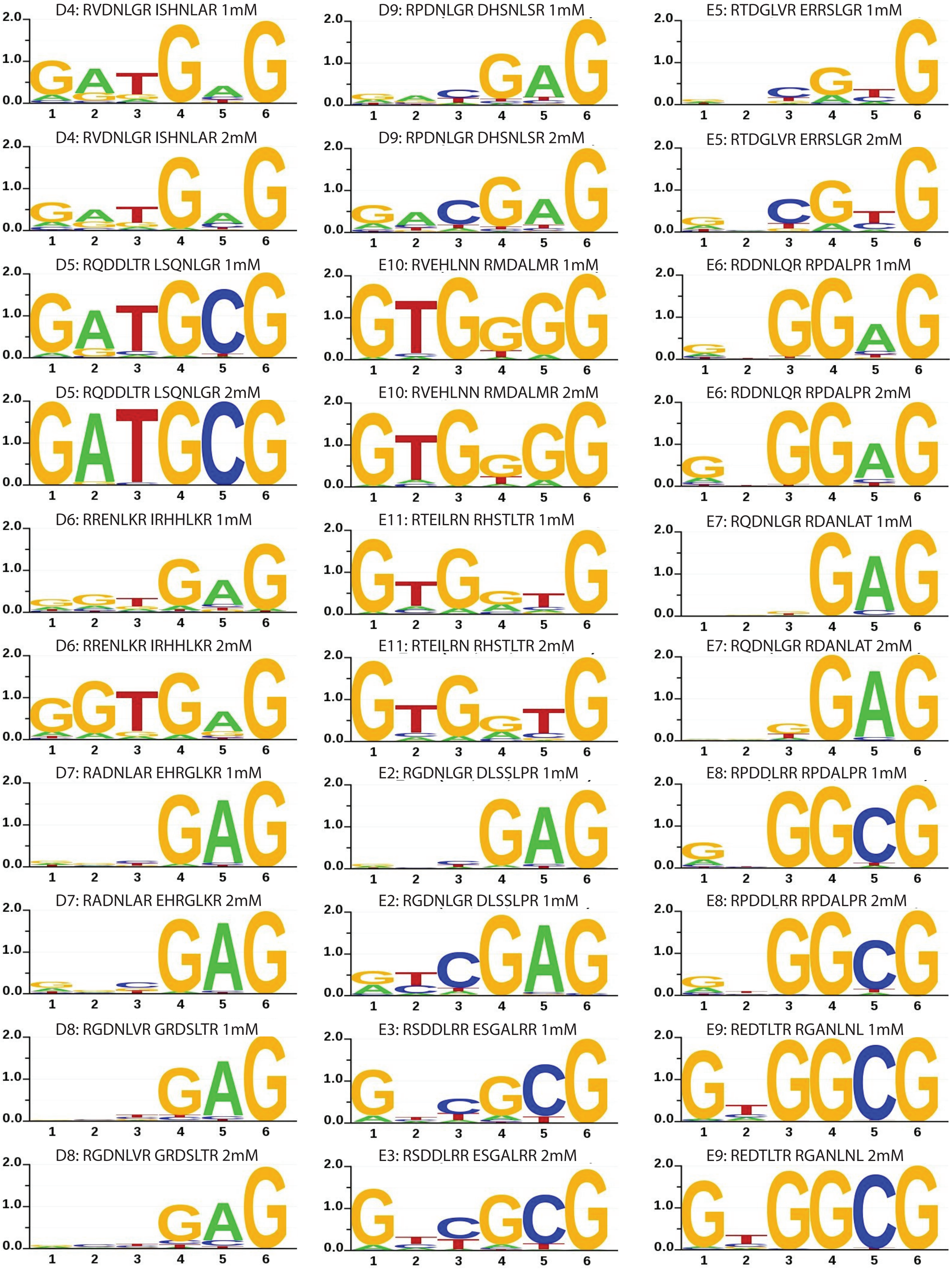
C8: REDGLHR HTHRLVS 1mM

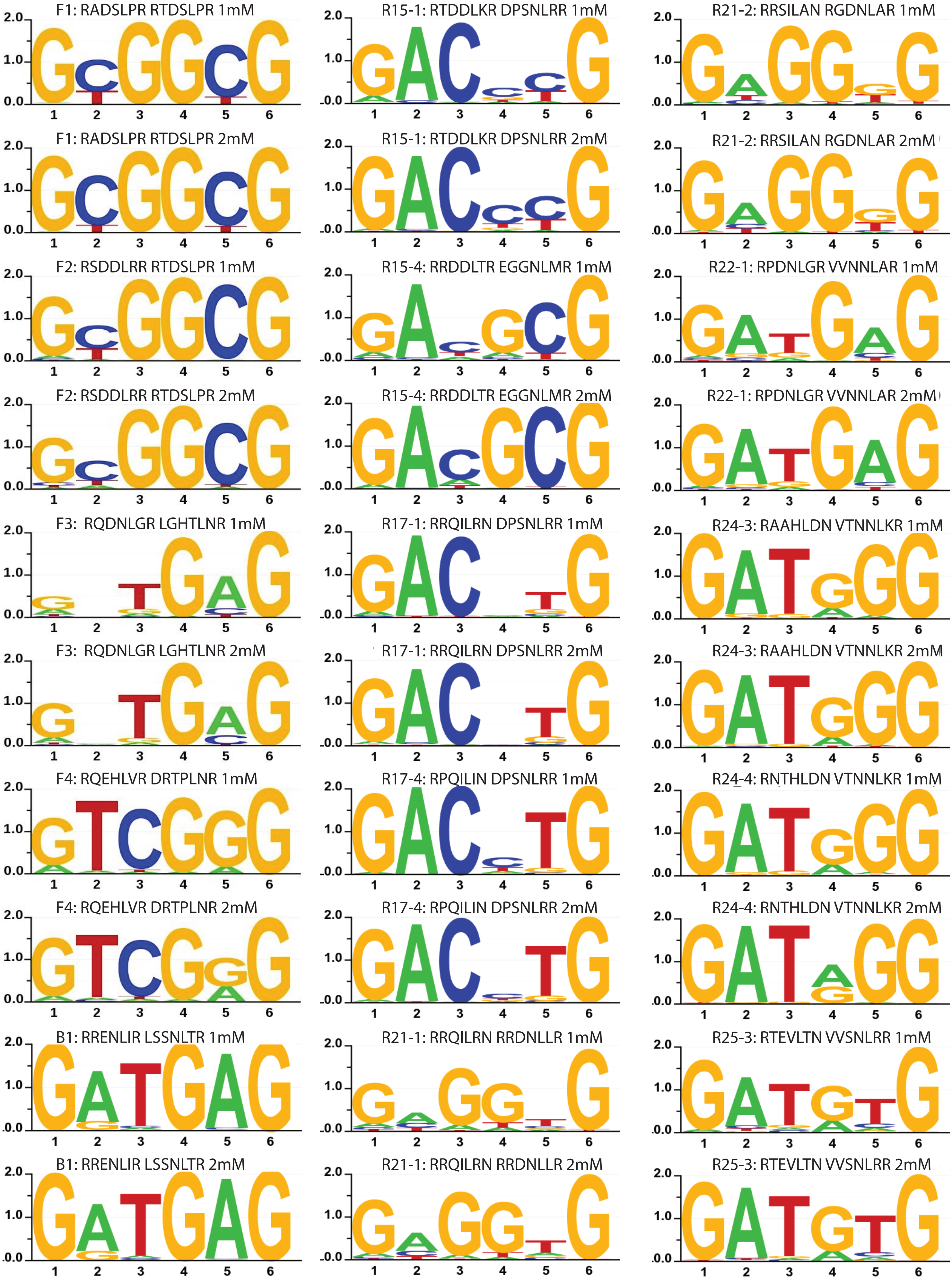
D2: RRAHLGN RQDNLQR 1mM

C12: RQDNLGR RREGLGR 2mM

C8: REDGLHR HTHRLVS 2mM

D2: RRAHLGN RQDNLQR 2mM


D4: RVDNLGR ISHNLAR 1mM
D9: RPDNLGR DHSNLSR 1mM
E5: RTDGLVR ERRSLGR 1mM
D4: RVDNLGR ISHNLAR 2mM
D9: RPDNLGR DHSNLSR 2mM
E5: RTDGLVR ERRSLGR 2mM
D5: RQDDLTR LSQNLGR 1mM
E10: RVEHLNN RMDALMR 1mM
E6: RDDNLQR RPDALPR 1mM
D5: RQDDLTR LSQNLGR 2mM
E10: RVEHLNN RMDALMR 2mM
E6: RDDNLQR RPDALPR 2mM
D6: RRENLKR IRHHLKR 1mM
E11: RTEILRN RHSTLTR 1mM
E7: RQDNLGR RDANLAT 1mM
D6: RRENLKR IRHHLKR 2mM
E11: RTEILRN RHSTLTR 2mM
E7: RQDNLGR RDANLAT 2mM
D7: RADNLAR EHRGLKR 1mM
E2: RGDNLGR DLSSLPR 1mM
E8: RPDDLRR RPDALPR 1mM
D7: RADNLAR EHRGLKR 2mM
E2: RGDNLGR DLSSLPR 1mM
E8: RPDDLRR RPDALPR 2mM
D8: RGDNLVR GRDSLTR 1mM
E3: RSDDLRR ESGALRR 1mM
E9: REDTLTR RGANLNL 1mM
D8: RGDNLVR GRDSLTR 2mM
E3: RSDDLRR ESGALRR 2mM
E9: REDTLTR RGANLNL 2mM
2.0
1.0
0.0
1
2
3
4
5
6



F1: RADSLPR RTDSLPR 1mM

F1: RADSLPR RTDSLPR 2mM

F2: RSDDLRR RTDSLPR 1mM

F2: RSDDLRR RTDSLPR 2mM

F3: RQDNLGR LGHTLNR 1mM

F3: RQDNLGR LGHTLNR 2mM

F4: RQEHLVR DRTPLNR 1mM

F4: RQEHLVR DRTPLNR 2mM

B1: RRENLIR LSSNLTR 1mM

B1: RRENLIR LSSNLTR 2mM

R15-1: RTDDLKR DPSNLRR 1mM

R15-1: RTDDLKR DPSNLRR 2mM

R15-4: RRDDLTR EGGNLMR 1mM

R15-4: RRDDLTR EGGNLMR 2mM

R17-1: RRQILRN DPSNLRR 1mM

R17-1: RRQILRN DPSNLRR 2mM

R17-4: RPQILIN DPSNLRR 1mM

R17-4: RPQILIN DPSNLRR 2mM

R21-1: RRQILRN RRDNLLR 1mM

R21-1: RRQILRN RRDNLLR 2mM

R21-2: RRSILAN RGDNLAR 1mM

R21-2: RRSILAN RGDNLAR 2mM

R22-1: RPDNLGR VVNNLAR 1mM

R22-1: RPDNLGR VVNNLAR 2mM

R24-3: RAAHLDN VTNNLKR 1mM

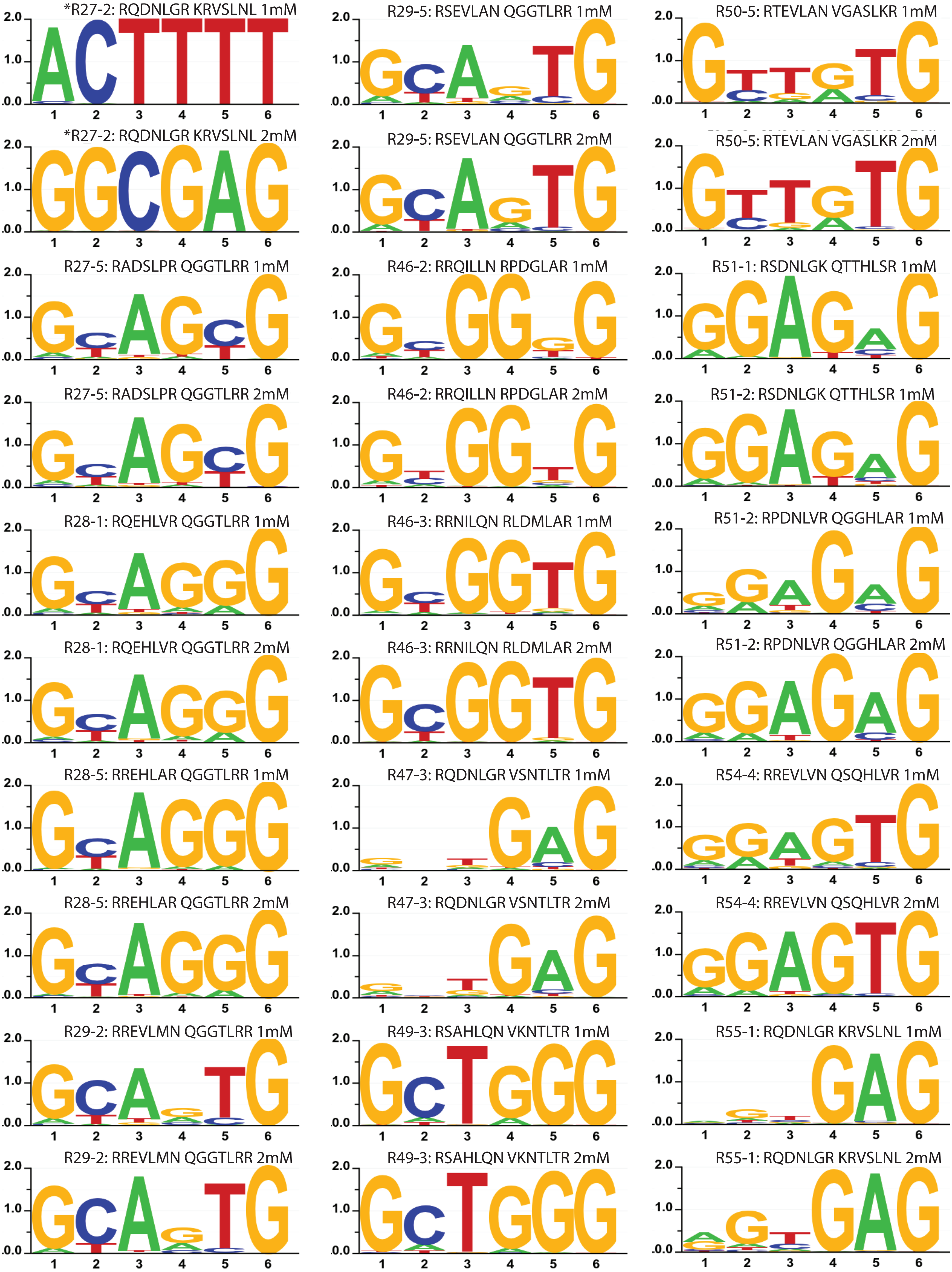
R24-3: RAAHLDN VTNNLKR 2mM

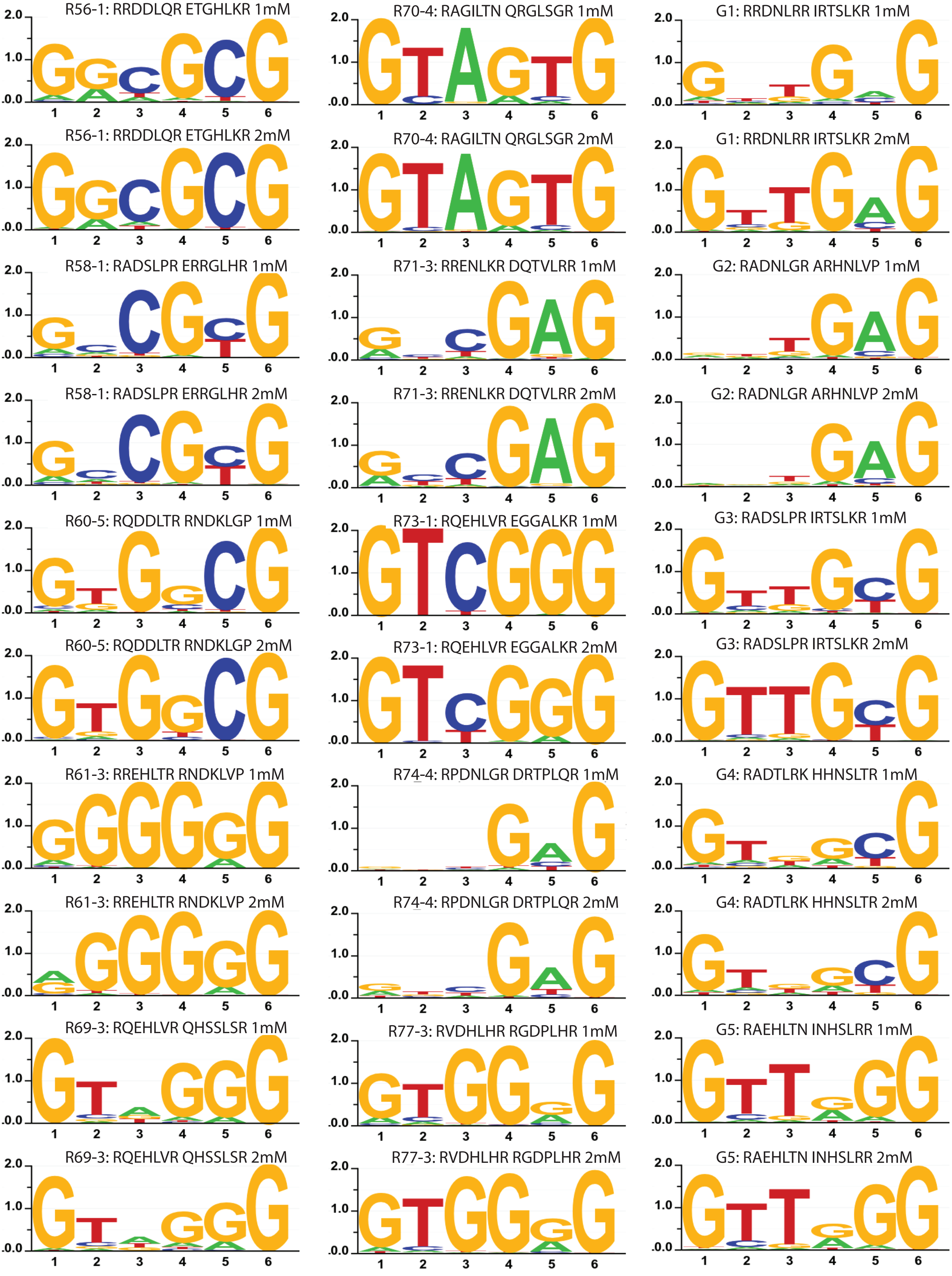
R24-4: RNTHLDN VTNNLKR 1mM

R24-4: RNTHLDN VTNNLKR 2mM

R25-3: RTEVLTN VVSNLRR 1mM

R25-3: RTEVLTN VVSNLRR 2mM


*R27-2: RQDNLGR KRVSLNL 1mM
R29-5: RSEVLAN QGGTLRR 1mM
R50-5: RTEVLAN VGASLKR 1mM
*R27-2: RQDNLGR KRVSLNL 2mM
R29-5: RSEVLAN QGGTLRR 2mM
R50-5: RTEVLAN VGASLKR 2mM
R27-5: RADSLPR QGGTLRR 1mM
R46-2: RRQILLN RPDGLAR 1mM
R51-1: RSDNLGK QTTHLSR 1mM
R27-5: RADSLPR QGGTLRR 2mM
R46-2: RRQILLN RPDGLAR 2mM
R51-2: RSDNLGK QTTHLSR 1mM
R28-1: RQEHLVR QGGTLRR 1mM
R46-3: RRNILQN RLDMLAR 1mM
R51-2: RPDNLVR QGGHLAR 1mM
R28-1: RQEHLVR QGGTLRR 2mM
R46-3: RRNILQN RLDMLAR 2mM
R51-2: RPDNLVR QGGHLAR 2mM
R28-5: RREHLAR QGGTLRR 1mM
R47-3: RQDNLGR VSNTLTR 1mM
R54-4: RREVLVN QSQHLVR 1mM
R28-5: RREHLAR QGGTLRR 2mM
R47-3: RQDNLGR VSNTLTR 2mM
R54-4: RREVLVN QSQHLVR 2mM
R29-2: RREVLMN QGGTLRR 1mM
R49-3: RSAHLQN VKNTLTR 1mM
R55-1: RQDNLGR KRVSLNL 1mM
R29-2: RREVLMN QGGTLRR 2mM
R49-3: RSAHLQN VKNTLTR 2mM
R55-1: RQDNLGR KRVSLNL 2mM
2.0
1.0
0.0
1
2
3
4
5
6



R56-1: RRDDLQR ETGHLKR 1mM

R56-1: RRDDLQR ETGHLKR 2mM

R58-1: RADSLPR ERRGLHR 1mM

R58-1: RADSLPR ERRGLHR 2mM

R60-5: RQDDLTR RNDKLGP 1mM

R60-5: RQDDLTR RNDKLGP 2mM

R61-3: RREHLTR RNDKLVP 1mM

R61-3: RREHLTR RNDKLVP 2mM

R69-3: RQEHLVR QHSSLSR 1mM

R69-3: RQEHLVR QHSSLSR 2mM

R70-4: RAGILTN QRGLSGR 1mM

R70-4: RAGILTN QRGLSGR 2mM

R71-3: RRENLKR DQTVLRR 1mM

R71-3: RRENLKR DQTVLRR 2mM

R73-1: RQEHLVR EGGALKR 1mM

R73-1: RQEHLVR EGGALKR 2mM

R74-4: RPDNLGR DRTPLQR 1mM

R74-4: RPDNLGR DRTPLQR 2mM

R77-3: RVDHLHR RGDPLHR 1mM

R77-3: RVDHLHR RGDPLHR 2mM

G1: RRDNLRR IRTSLKR 1mM

G1: RRDNLRR IRTSLKR 2mM

G2: RADNLGR ARHNLVP 1mM

G2: RADNLGR ARHNLVP 2mM

G3: RADSLPR IRTSLKR 1mM

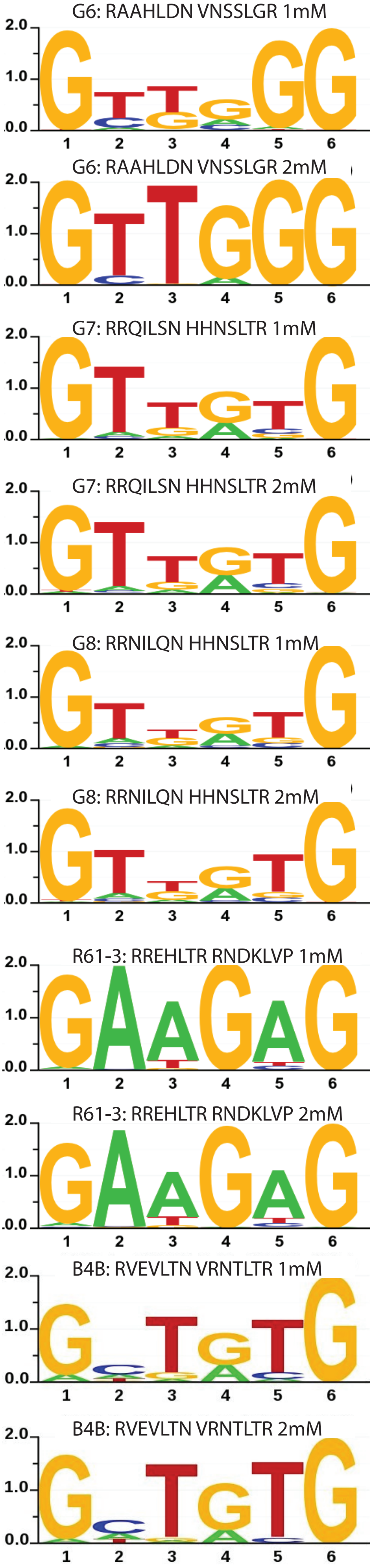
G3: RADSLPR IRTSLKR 2mM

G4: RADTLRK HHNSLTR 1mM

G4: RADTLRK HHNSLTR 2mM

G5: RAEHLTN INHSLRR 1mM

G5: RAEHLTN INHSLRR 2mM



G6: RAAHLDN VNSSLGR 1mM

G6: RAAHLDN VNSSLGR 2mM

G7: RRQILSN HHNSLTR 1mM

G7: RRQILSN HHNSLTR 2mM

G8: RRNILQN HHNSLTR 1mM

G8: RRNILQN HHNSLTR 2mM

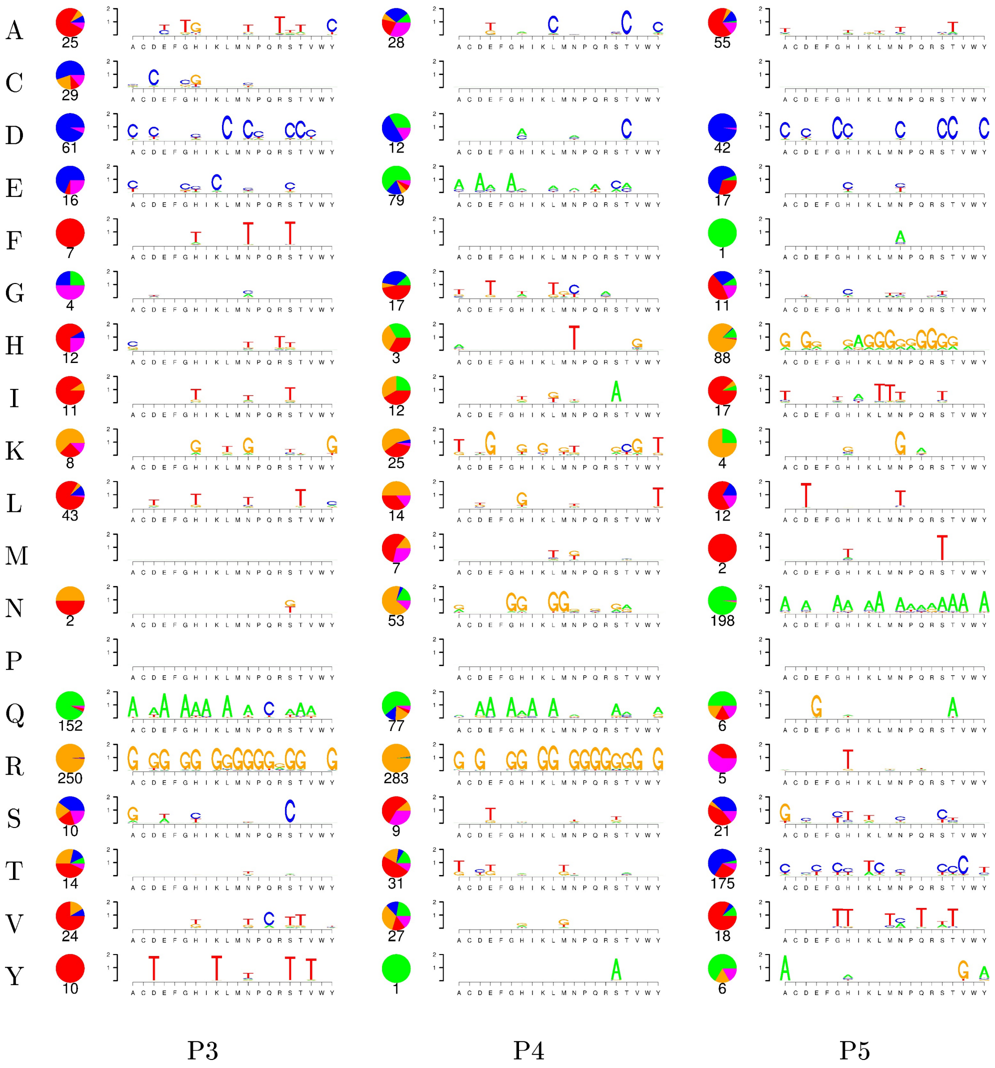
R61-3: RREHLTR RNDKLVP 1mM

R61-3: RREHLTR RNDKLVP 2mM

B4B: RVEVLTN VRNTLTR 1mM

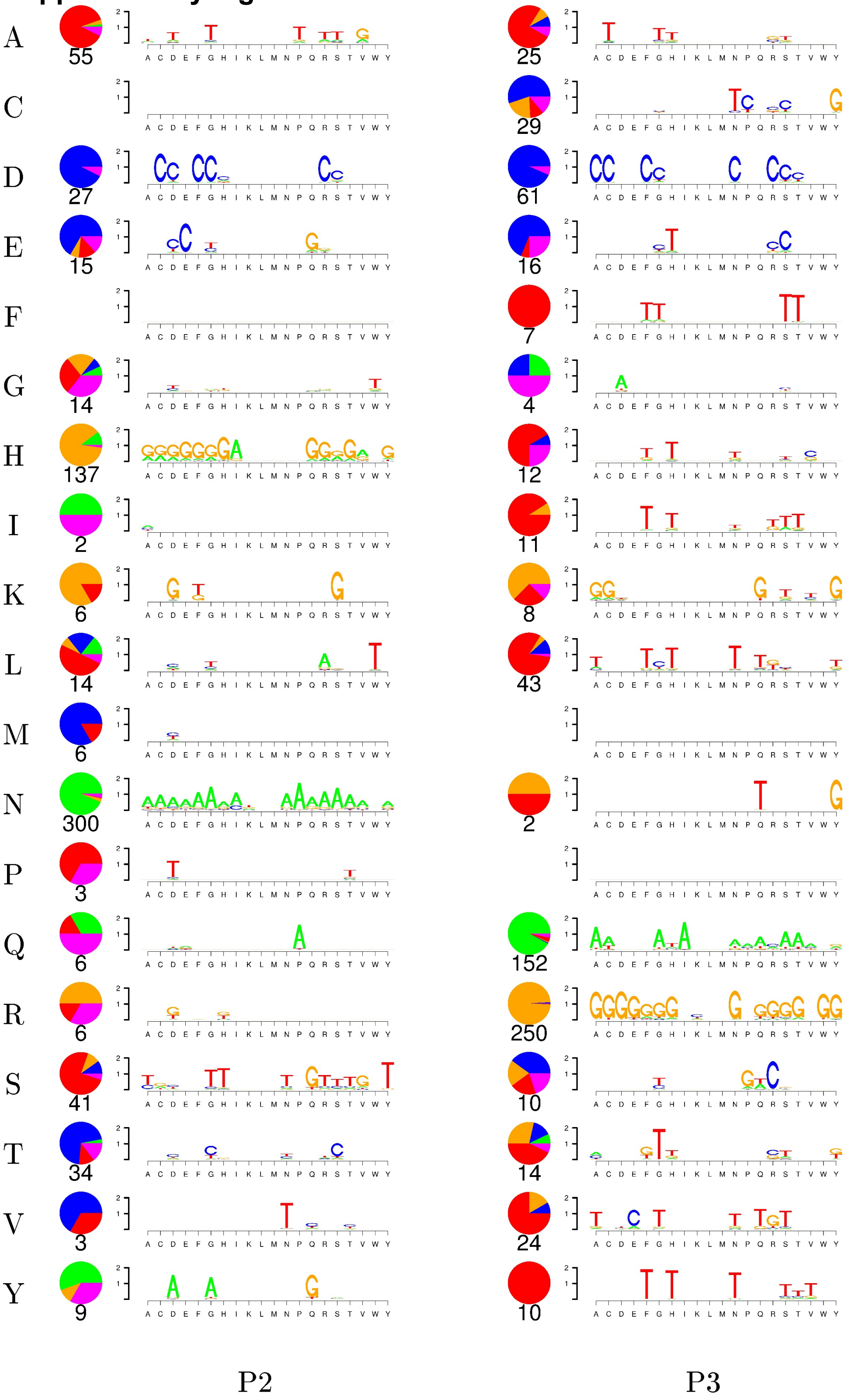
B4B: RVEVLTN VRNTLTR 2mM

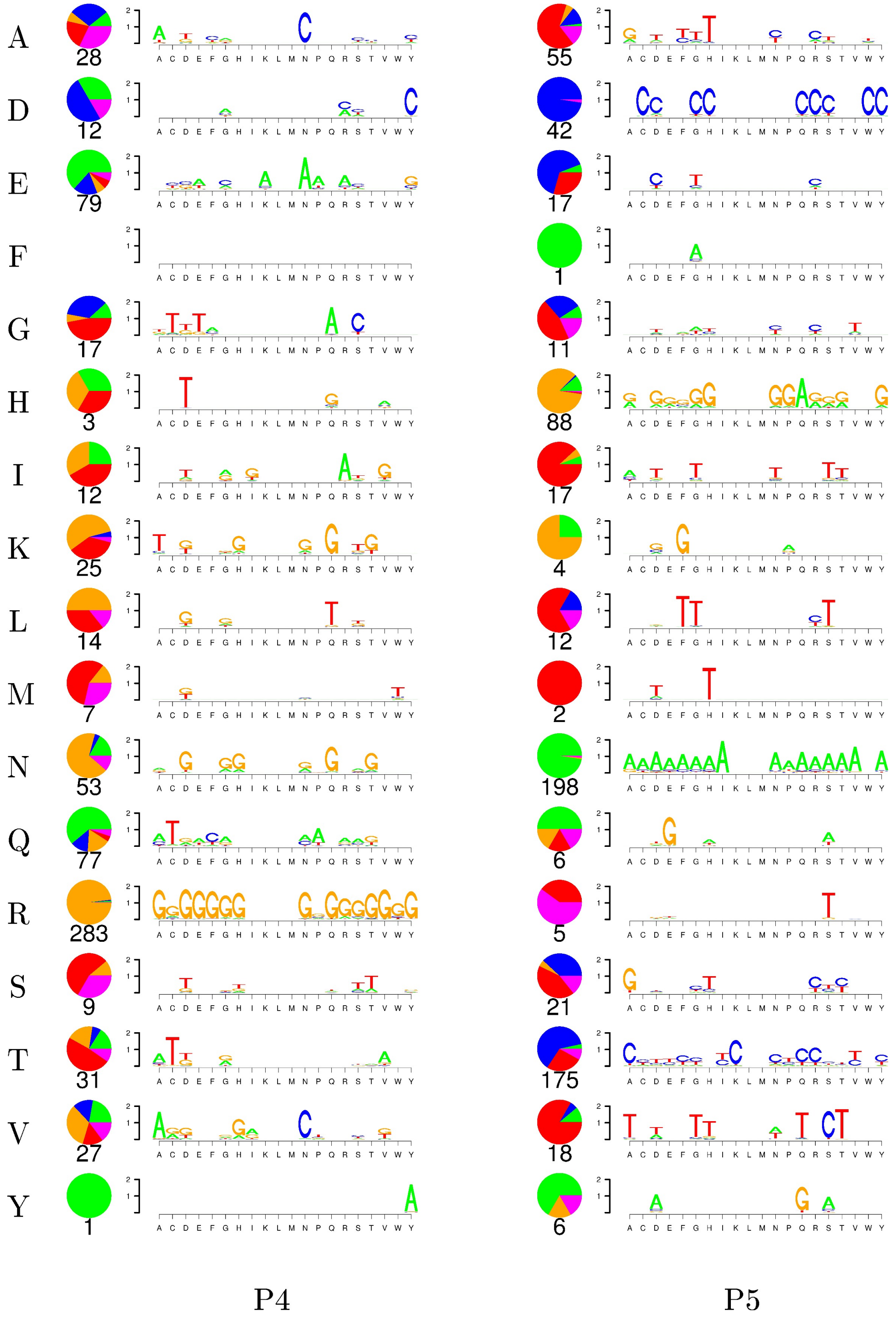
Supplementary Figure 3.



Influence of amino-acid at position 3 of C-terminal finger of a two finger module on base preferences at the positions 3 through 5 (P3-P5) of the six bp binding site. (Left column – P3) Influence of the amino acid at position +3 on the base preferred at position 3 (P3) of the binding site. In the canonical binding mode, the preference of the base at this position is influenced primarily by the amino acid present at position -1 of C-terminal finger (identity indicated to the left of each chart). Numbers of mutants in our database with that amino acid at position -1 of C-terminal finger are given at the bottom of individual pie chart. The different colored regions in pie charts indicate the frequency with which the bases were recovered as a preferred base at P3. Frequencies of occurrences of Ade, Cyt, Gua and Thy are rendered by magenta, yellow, blue and red, respectively. In the complete absence of context dependence, only one color should be observed in pie chart. Amino acids present at position +3 of C-terminal finger are given in X-axis tick labels for each logo. If there is more than one mutant with the same amino acids at position -1 and position +3 of C-terminal finger, then the frequencies of bases are averaged over all mutants. In summary, each logo indicates relative preference of bases as a function of amino acids at position +3 (as in the X-axis labels) and -1 of C-terminal finger (given on the left most panel). (middle P4 column) Influence of the amino acid at position +3 on base P4. Each logo indicates relative preference of bases as a function of amino acids at position +3 (as in the X-axis labels) of C-terminal finger and at position +6 of N-terminal finger (given on the left most panel). (D) Influence of the amino acid at position +3 on base P5. Each logo indicates relative preference of bases as a function of amino acids at position +3 (as in the X-axis labels) of C-terminal finger and at position +3 of N-terminal finger (given on the left most panel).

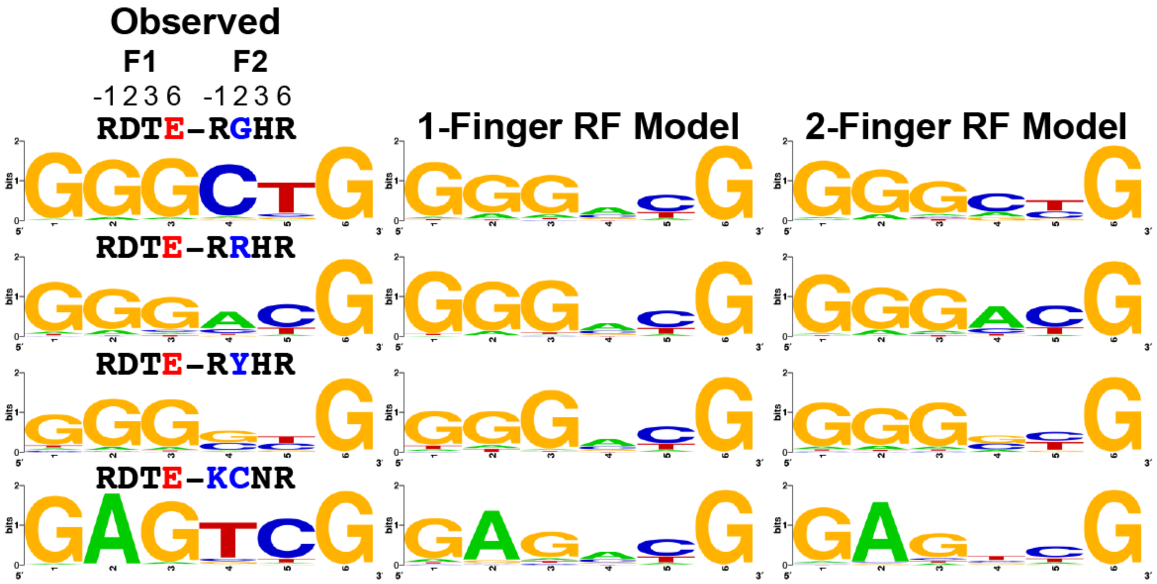
Supplementary Figure 4.





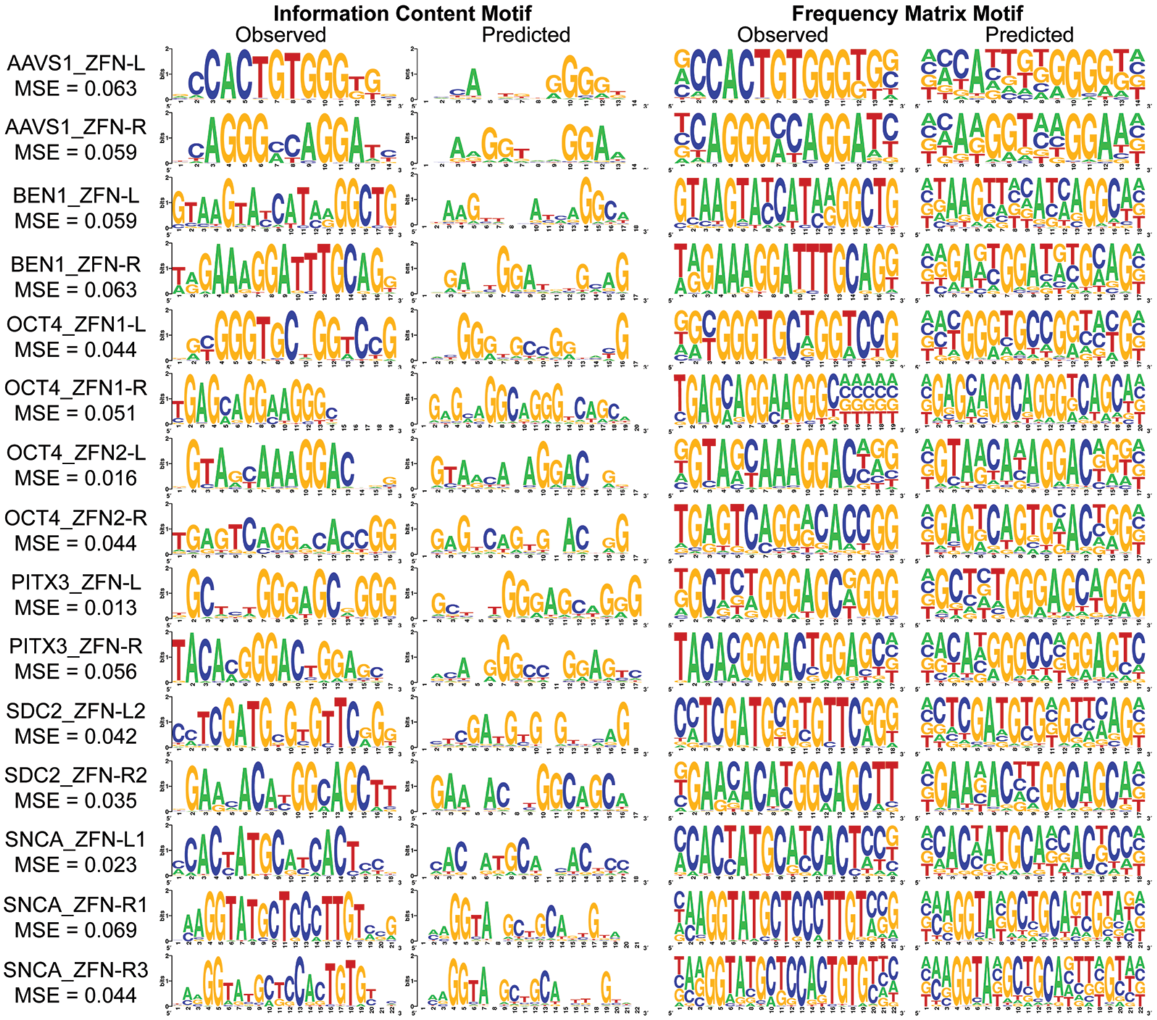
Influence of the amino acid at position +2 of C-terminal finger of a two finger module on base preferences at the positions 2 through 5 (P2-P5) of the six bp binding site. (A) Influence of the amino acid at position +2 on the base preferred at position 2 (P2) of the binding site. In the canonical binding mode, the preference of the base at this position is influenced primarily by the amino acid present at position +3 of C-terminal finger (indicated to the left of each chart). Numbers of mutants in our database with that amino acid at position +3 of C-terminal finger are given at the bottom of individual pie chart. The different colored regions in pie charts indicate the frequency with which the bases were recovered as a preferred base at P2. Frequencies of occurrences of Ade, Cyt, Gua and Thy are rendered by magenta, yellow, blue and red, respectively. In the complete absence of context dependence, only one color should be observed in pie chart. Amino acids present at position +2 of C-terminal finger are given in X-axis tick labels for each logo. If there is more than one mutant with the same amino acids at position +3 and position +2 of C-terminal finger, then the frequencies of bases are averaged over all mutants. In summary, each logo indicates relative preference of bases as a function of amino acids at position +2 (as in the X-axis labels) and 3 of C-terminal finger (given on the left most panel). (B) Influence of the amino acid at position +2 on base P3. Each logo indicates relative preference of bases as a function of amino acids at position +2 (as in the X-axis labels) and -1 of C-terminal finger (given on the left most panel) (C) Influence of the amino acid at position +2 on base P4. Each logo indicates relative preference of bases as a function of amino acids at position +2 (as in the X-axis labels) of C-terminal finger and at position +6 of N-terminal finger (given on the left most panel). (D) Influence of the amino acid at position 2 on base P5. Each logo indicates relative preference of bases as a function of amino acids at position +2 (as in the X-axis labels) of C-terminal finger and at position +3 of N-terminal finger (given on the left most panel).

Supplementary Figure 5.



Prediction of context dependent recognition preferences in Figure 3B. The one finger and two finger RF prediction models were used to generate a motif for the four two finger modules characterized in the B1H system (observed). The one finger RF model, which predicts each 3 bp subsite based on the recognition residues at positions -1, +2, +3 & +6, fails to capture the context dependence of recognition at the finger-finger interface. The two finger RF model effectively captures the context dependence. The successful prediction of the recognition motifs by the two finger model is expected since these the specificities of these four two-finger modules were included in the training set for the final RF model.

Supplementary Figure 6.



Comparison of the SELEX motifs for various ZFPs (Observed) (26,104-106) and their predicted motifs based on ZFModels (Predicted). The left columns display the motifs as information content, whereas the right columns display the motifs as position frequency plots. One base pair gaps were inserted into a subset of the predicted ZFP motifs based on the presence of non-canonical linkers connecting the fingers (BEN1-ZFN-L F3-F4; Pitx3-ZFN-R F2-F3; SDC2-ZFN-L2 F2-F3; SNCA-L1 F2-F3; SNCA-R1 F5-F6; SNCA-R3 F2-F3 & F5-F6). MSE values for the comparison of the SELEX and predicted motif are displayed above each predicted motif.