



(51) International Patent Classification:

A61K 38/16 (2006.01) C12N 15/09 (2006.01)

C07K 19/00 (2006.01)

(21) International Application Number:

PCT/US2019/032937

(22) International Filing Date:

17 May 2019 (17.05.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/672,682 17 May 2018 (17.05.2018) US

62/828,277 02 April 2019 (02.04.2019) US

(71) Applicant: THE GENERAL HOSPITAL CORPO-

RATION [US/US]; 55 Fruit Street, Boston, Massachusetts 02114 (US).

(72) Inventors: COTTMAN, Rebecca Tayler; 11 Sixth Street, Apt. 3, Cambridge, Massachusetts 02141 (US). JOUNG, J. Keith; 1 Magnolia Way, Winchester, Massachusetts 01890 (US).

(74) Agent: WON, Doyon et al.; Fish & Richardson P.C., P.O. Box 1022, Minneapolis, Minnesota 55440-1022 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(54) Title: CCCTC-BINDING FACTOR VARIANTS

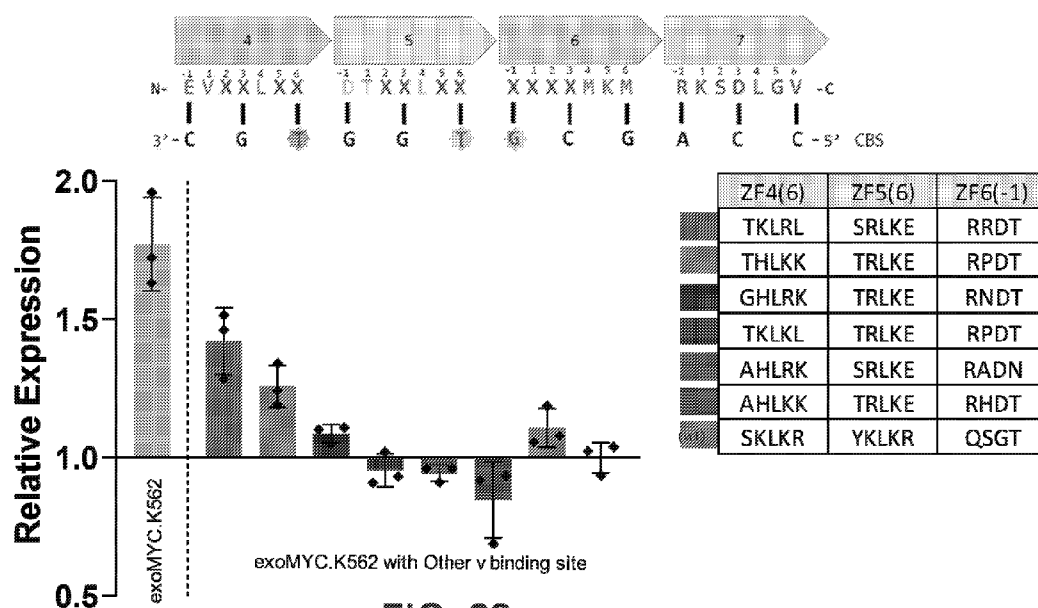


FIG. 28

(57) Abstract: The present invention relates to an engineered CCCTC-binding factor (CTCF) variant including at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF, where the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS) with a higher affinity than wild-type CTCF, the mutant CBS including at least one nucleotide base that differs in sequence from the nucleotide sequence of a consensus CBS, where the at least one amino acid residue that differs in sequence from the amino acid sequence of a wild-type CTCF is selected from the group consisting of the amino acid residues at the position(s) -1, +1, +2, +3, +5, and +6 of any of ZF7, ZF6, ZF5, ZF4, and ZF3 of the engineered CTCF variant.

(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

CCCTC-Binding Factor Variants

CLAIM OF PRIORITY

This application claims the benefit of U.S. Provisional Patent Application Serial No. 62/672,682, filed on May 17, 2018 and U.S. Provisional Patent Application Serial No. 62/828,277, filed on April 2, 2019. The entire contents of the foregoing are hereby incorporated by reference.

FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

This invention was made with Government support under Grant No. GM118158 awarded by the National Institutes of Health. The Government has certain rights in the invention.

TECHNICAL FIELD

The invention relates, at least in part, to engineered CCCTC-binding factor variants with altered DNA-binding specificities.

BACKGROUND

CCCTC-binding factor (CTCF) is a multi-domain protein that acts as an essential genome organizer by maintaining higher-order chromatin structure while also having a role in cell differentiation and the promotion or repression of gene expression (Ong and Corces, *Nature Reviews Genetics* (2014); Phillips and Corces, *Cell* (2009)). CTCF maintains topologically associated domains (TADs) spanning MBs of the genome as well as smaller scale Sub-TADs leading to fine-tuned gene insulation or gene activation within gene clusters (Ali et al., *Current Opinion in Genetics & Development* (2016); Nora et al., *Nature* (2012); Rao et al., *Cell* (2014)). In addition, CTCF has been found to regulate mRNA splicing by influencing the rate of transcription and more recently been implicated in promoting homologous recombination repair at double-strand breaks (Shukla et al., *Nature* (2011); Hilmi et al., *Science Advances* (2017); Han et al., *Scientific Reports* (2016)). CTCF binds throughout the genome via an 11 finger zinc finger (ZF) array that recognizes CTCF binding sites (CBSs). The CBS is typically 40bp in length with a highly conserved 15bp core sequence.

SUMMARY

The present invention is based, at least in part, on the development of engineered CTCF variants that can bind to mutant CBSs with higher affinity than a wild-type CTCF.

The present invention relates to an engineered CCCTC-binding factor (CTCF) variant including at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF, where the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS) with a higher affinity than wild-type CTCF, the mutant CBS including at least one nucleotide base that differs in sequence from the nucleotide sequence of a consensus CBS, where the at least one amino acid residue that differs in sequence from the amino acid sequence of a wild-type CTCF is selected from the group consisting of the amino acid residues at the position(s) -1, +1, +2, +3, +5, and +6 of any of ZF7, ZF6, ZF5, ZF4, and ZF3 of the engineered CTCF variant.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CTCF binding sequence (CBS) that has a Thymine (T), Adenine (A), or Guanine (G) residue at position 2 of the consensus CBS motif, the engineered CTCF including an amino acid residue threonine, asparagine, or histidine at ZF7 position +3.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a G residue at position 2 of the consensus CBS motif, the engineered CTCF including the amino acid sequence DHLQT, EHLNV, AHLQV, EHLRE, DHLQV, EHLKV, EHLVV, DHLRT, or DHLAT at ZF7 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a T, G, or C residue at position 3 of the consensus CBS motif, the engineered CTCF at ZF7 positions -1 to +3 including: the amino acid sequence RKHD or RRSD, where the mutant CBS has a T residue at position 3 of the consensus CBS motif; the amino acid sequence RKAD, IPRI, RKHD, or RKDD, where the mutant CBS has a G residue at position 3 of the consensus CBS motif; or the amino acid sequence GIVN, ELLN, QALL or PHRM, where the mutant CBS has a C residue at position 3 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a T, G, or A residue at position 5 of the consensus CBS motif, the engineered CTCF at ZF6 positions +2 to +6 including: the amino acid sequence

NAMKR, GNMAR, EGMTR, SNMVR, or NAMRG, where the mutant CBS has a T residue at position 5 of the consensus CBS motif; or the amino acid sequence EHMGR, DHMNR, THMKR, EHMRR, or THMNR, where the mutant CBS has a G residue at position 5 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a T, G, or C residue at position 6 of the consensus CBS motif, the engineered CTCF at ZF6 positions -1 to +3 including: the amino acid sequence MNES or HRES, where the mutant CBS has a T residue at position 6 of the consensus CBS motif; or the amino acid sequence RPDY, RTDI, or RHDT, where the mutant CBS has a G residue at position 6 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a C, A, or T residue at position 7 of the consensus CBS motif, the engineered CTCF at ZF5 positions +2 to +6 including: the amino acid sequence HGLKV, HRLKE, HALKV, SRLKE, or DGLRV, where the mutant CBS has a T residue at position 7 of the consensus CBS motif; the amino acid sequence HTLKV, or HGLKV, where the mutant CBS has an A residue at position 7 of the consensus CBS motif; or the amino acid sequence SRLKE, HRLKE or NRLKE, where the mutant CBS has a C residue at position 7 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a C, A, or T residue at position 8 of the consensus CBS motif, the engineered CTCF at ZF5 positions +2 to +6 including: the amino acid sequence ATLKR, QALRR, GGLVR, or HGLIR, where the mutant CBS has a T residue at position 8 of the consensus CBS motif; the amino acid sequence ANLSR, TGLTR, HGLVR, or GGLTR, where the mutant CBS has an A residue at position 8 of the consensus CBS motif; the amino acid sequence HTLRR, TVLKR, ADLKR, or HGLRR, where the mutant CBS has a C residue at position 8 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a T, A, or C residue at position 10 of the consensus CBS motif, the engineered CTCF at ZF4 positions +2 to +6 including: the amino acid sequence AHLRK, wherein the mutant CBS has a T residue at position 10 of the consensus CBS motif; the amino acid sequence AKLRV, EKLRI, or AKLRI, where the mutant CBS has an A residue at

position 10 of the consensus CBS motif; or the amino acid sequence TKLKV, wherein the mutant CBS has a C residue at position 10 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a T, A, or C residue at position 11 of the consensus CBS motif, the engineered CTCF at ZF4 positions +2 to +6 including: the amino acid sequence ATLRR or RRLDR, where the mutant CBS has a T residue at position 11 of the consensus CBS motif; the amino acid sequence TNLRR, ANLRR, or GNLTR, where the mutant CBS has an A residue at position 11 of the consensus CBS motif; or the amino acid sequence AMLKR, HMLTR, AMLRR, or TMLRR, where the mutant CBS has a C residue at position 11 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a T, A, or C residue at position 13 of the consensus CBS motif, the engineered CTCF at ZF3 positions +2 to +6 including: the amino acid sequence QQLIV, SQLIV, QQLLV, GELVV, or QQLLI, where the mutant CBS has a T residue at position 13 of the consensus CBS motif; the amino acid sequence GQLIV, GQLTV, GKLVT, TELII or QGLLV, where the mutant CBS has an A residue at position 13 of the consensus CBS motif; or the amino acid sequence QQLLT, GQLLT, GELLT, or QQLLI, where the mutant CBS has a C residue at position 13 of the consensus CBS motif.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has A, G, T, and T residues at positions 2, 6, 7, and 10 of the consensus CBS motif, respectively, the engineered CTCF including: (i) the amino acid sequence AKLKK, AKLRK, AHLRV, AKLRV, or SKLRL at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence ERLRV, NRLKV, SRLKE, or NRLKV at ZF5 positions +2 to +6 of the engineered CTCF; (iii) the amino acid sequence RPDT, RTET, or RADV at ZF6 positions -1 to +3 of the engineered CTCF; and (iv) the amino acid sequence DNLLA, SNLLV, DNLMA, or DNLRV at ZF7 positions +2 to +6 of the engineered CTCF.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has G, G, T, and T residues at positions 2, 6, 7, and 10 of the consensus CBS motif, respectively, the engineered CTCF including: (i) the amino acid sequence GHLKK, AHLRK, or GKLRI at ZF4 positions +2 to +6 of the engineered CTCF;

(ii) the amino acid sequence SRLKE, DALRR, DGLKR, or TRLRE at ZF5 positions +2

to +6 of the engineered CTCF; (iii) the amino acid sequence at RPD_{TM}KR or RTEN_{TM}KM at ZF6 positions -1 to +6 of the engineered CTCF; and (iv) the amino acid sequence EHLKV, DHLLA, or HHLDV at ZF7 positions +2 to +6 of the engineered CTCF.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has A, G, and A residues at positions 2, 5, and 11 of the consensus CBS motif, respectively, the engineered CTCF including: (i) the amino acid sequence SNLRR, GNLVR, GNLRR, GNLKR, ANLRR, NNLRR, or TNLRR at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence EHM_{TM}KR, EHM_{TM}RR, THM_{TM}KR, EHM_{TM}NR, or EHM_{TM}AR at ZF6 positions +2 to +6 of the engineered CTCF; and (iii) the amino acid sequence DNLLT, DNLLV, DNLQT, DNLLA, DNLAT, DNLQA, DNLMA, or DNLMT at ZF7 positions +2 to +6 of the engineered CTCF.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has G, G, and A residues at positions 2, 5, and 11 of the consensus CBS motif, respectively, the engineered CTCF including: (i) the amino acid sequence GNLVR, GNLRR, GNLAR, GNL_{TM}MR, ANLRR, SNLRR, or NNLRR at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence EHM_{TM}NR, EHM_{TM}KR, EHM_{TM}RR, SHM_{TM}NR, SHM_{TM}RR, THM_{TM}KR, or DHM_{TM}NR at ZF6 positions +2 to +6 of the engineered CTCF; and (iii) the amino acid sequence EHLKV, EHLAE, STLNE, DHLQV, EHLNV, DHLNT, EHLQA, or HHLMH at ZF7 positions +2 to +6 of the engineered CTCF.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has G, T, and T residues at positions 6, 7, and 10 of the consensus CBS motif, respectively, the engineered CTCF including: (i) the amino acid sequence GHLKK, AHLKK, TKLRL, TKLKL, GHLRK, THLKK, or AHLRK at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence TRLKE or SRLKE at ZF5 positions +2 to +6 of the engineered CTCF; and (iii) the amino acid sequence RADN, RHDT, RRDT, RPD_{TM}T, RTSS, or RNDT at ZF6 positions -1 to +3 of the engineered CTCF.

In some embodiments, the engineered CTCF variant includes at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF, where the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS) with a higher affinity than wild-type CTCF, the mutant CBS including at least one nucleotide base that differs in sequence from the nucleotide sequence of a consensus CBS, where

the at least one amino acid residue that differs in sequence from the amino acid sequence of a wild-type CTCF is selected from the group consisting of the amino acid residues at the position(s) -1, +1, +2, +3, +5, and +6 of any of ZF7, ZF6, ZF5, ZF4, and ZF3 of the engineered CTCF variant.

In some embodiments, the engineered CCCTC-binding factor (CTCF) variant that binds with a higher affinity than a wild-type CTCF to a mutant CTCF binding sequence (CBS) that differs from a consensus CBS at position 2 of the consensus CBS motif, the engineered CTCF including an amino acid residue threonine, asparagine, or histidine at ZF7 +3 position.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that has a C-to-G mutation at position 2 of the consensus CBS motif, the engineered CTCF including the amino acid sequence DHLQT, EHLNV, AHLQV, EHLRE, DHLQV, EHLKV, DHLQV, EHLVV, DHLRT, DHLAT, or DHLQT at ZF7 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 3 of the consensus CBS motif, the engineered CTCF including the amino acid sequence RKHD, RRSD, GIVN, ELLN, or PHRM at ZF7 positions -1 to +3.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 5 of the consensus CBS motif, the engineered CTCF including the amino acid sequence NAMKR, EHMGR, DHMNR, THMKR, EHMRR, or THMNR at ZF6 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 6 of the consensus CBS motif, the engineered CTCF including the amino acid sequence MNES, HRES, RPDT, RTDI, or RHDT at ZF6 positions -1 to +3.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 7 of the consensus CBS motif, the engineered CTCF including the amino acid sequence HGLKV, HRLKE, HALKV, SRLKE, DGLRV, HTLVK, or NRLKE at ZF5 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 8 of the

consensus CBS motif, the engineered CTCF including the amino acid sequence ATLKR, QALRR, GGLVR, HGLIR, ANLSR, TGLTR, HGLVR, GGLTR, HTLRR, TVLKR, ADLKR, or HGLRR at ZF5 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 10 of the consensus CBS motif, the engineered CTCF including the amino acid sequence AHLRK, AKLRV, GGLGL, AKLRI, TKLKV, or SKLRV at ZF4 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 11 of the consensus CBS motif, the engineered CTCF including the amino acid sequence ATLRR, RRLDR, TNLRR, ANLRR, GNLTR, AMLKR, HMLTR, AMLRR, or TMLRR at ZF4 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at position 13 of the consensus CBS motif, the engineered CTCF including the amino acid sequence QQLIV, SQLIV, QQLLV, GELVV, QQLLI, GQLIV, GQLTV, TELII, QGLLV, QQLLT, GQLLT, GELLT, or QQLLI at ZF3 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at positions 2, 6, 7, and 10 of the consensus CBS motif, the engineered CTCF including:

- (i) the amino acid sequence AKLKK, AKLRK, AHLRV, AKLRV, or SKLRL at ZF4 positions +2 to +6;
- (ii) the amino acid sequence ERLRV, NRLKV, SRLKE, or NRLKV at ZF5 positions +2 to +6;
- (iii) the amino acid sequence RPDV, RTET, or RADV at ZF6 positions -1 to +3; and
- (iv) the amino acid sequence DNLLA, SNLLV, DNLMA, or DNLRV at ZF7 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at positions 2, 6, 7, and 10 of the consensus CBS motif, the engineered CTCF including:

- (i) the amino acid sequence GHLKK, AHLRK, or GKLRI at ZF4 positions +2 to +6;

(ii) the amino acid sequence SRLKE, DALRR, DGLKR, or TRLRE at ZF5 positions +2 to +6;

(iii) the amino acid sequence at RPDTRMKR or RTENMKM at ZF6 positions -1 to +36; and

(iv) the amino acid sequence EHLKV, DHLLA, or HHLDV at ZF7 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at positions 2, 5, and 11 of the consensus CBS motif, the engineered CTCF including:

(i) the amino acid sequence SNLRR, GNLVR, GNLRR, GNLKR, ANLRR,>NNLRR, or TNLRR at ZF4 positions +2 to +6;

(ii) the amino acid sequence EHMKR, EHMRR, THMKR, EHMNR, or EHMAR at ZF6 positions +2 to +6; and

(iii) the amino acid sequence DNLLT, DNLLV, DNLQT, DNLLA, DNLAT, DNLQA, DNLMMA, or DNLMT at ZF7 positions +2 to +6.

In some embodiments, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at positions 2, 5, and 11 of the consensus CBS motif, the engineered CTCF including:

(i) the amino acid sequence GNLVR, GNLRR, GNLAR, GNLMR, ANLRR, SNLRR, or>NNLRR at ZF4 positions +2 to +6;

(ii) the amino acid sequence EHMNR, EHMKR, EHMRR, SHMNR, SHMRR, THMKR, or DHMNR at ZF6 positions +2 to +6; and

(iii) the amino acid sequence EHLKV, EHLAE, STLNE, DHLQV, EHLNV, DHLNT, EHLQA, or HHLMH at ZF7 positions +2 to +6.

In one embodiment, the engineered CTCF variant binds with a higher affinity than a wild-type CTCF to a mutant CBS that differs from a consensus CBS at positions 6, 7, and 10 of the consensus CBS motif, the engineered CTCF including:

(i) the amino acid sequence GHLKK, AHLKK, TKLRL, TKLKL, GHLRK, THLKK, or AHLRK at ZF4 positions +2 to +6;

(ii) the amino acid sequence TRLKE or SRLKE at ZF5 positions +2 to +6; and

(iii) the amino acid sequence RADN, RHDT, RRDT, RPDTR, RTSS, or RNDTR at ZF6 positions -1 to +3.

In some embodiments, the engineered CTCF variant interacts with cohesion to mediate the formation of an enhancer-promoter loop to modulate gene expression.

In another aspect, the invention features a method of treating a subject in need thereof, the method including administering to the subject a therapeutically effective amount of an engineered CTCF variant described herein.

In some embodiments, the subject can have cancer.

In another aspect, the invention features a method of activating or repressing expression of a gene which is under the control of a CBS bearing one or more mutations, the method including contacting the engineered CTCF according to any one of claims 1-15 with a sequence of interest in the gene, such that the expression of the gene is regulated.

In another aspect, the invention features a pharmaceutical composition including an engineered CTCF variant described herein.

In another aspect, the invention features a gene expression system for regulation of a gene, the system including a nucleic acid encoding an engineered CTCF variant according described herein.

In another aspect, the invention features a method of altering the structure of chromatin including contacting an engineered CTCF variant according to any one of claims 1-15 with a sequence of interest to form a binding complex, such that the structure of the chromatin is altered.

In another aspect, the invention features a method of activating or repressing expression of a gene which is under the control of a CBS bearing one or more mutations, the method including contacting the CBS bearing one or more mutations with an engineered CTCF variant described herein.

In another aspect, the invention features a kit including an engineered CTCF variant described herein.

In this disclosure, “comprises,” “comprising,” “containing” and “having” and the like can have the meaning ascribed to them in U.S. Patent law and can mean “includes,” “including,” and the like; “consisting essentially of” or “consists essentially” likewise has the meaning ascribed in U.S. Patent law and the term is open-ended, allowing for the presence of more than that which is recited so long as basic or novel characteristics of that which is recited is not changed by the presence of more than that which is recited, but excludes prior art embodiments.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Methods and materials are described herein for use in the present invention; other, suitable methods and materials known in the art can also be used. The materials, methods, and examples are illustrative only and not intended to be limiting. All publications, patent applications, patents, sequences, database entries, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control.

Other features and advantages of the invention will be apparent from the following detailed description and figures, and from the claims.

DESCRIPTION OF DRAWINGS

The following Detailed Description, given by way of example, but not intended to limit the invention to specific embodiment described, may be understood in conjunction with the accompanying figures, incorporated herein by reference.

FIG. 1: Diagram of an exemplary 11-finger CTCF zinc finger array protein-DNA interactions at the CTCF binding site. Each zinc finger of the 11-finger array contained a recognition alpha-helix where protein-DNA base contacts were made by amino acids in position -1, 2, 3 and 6 of each alpha-helix. Here, position -1, 3, and 6 were only depicted as position 2 makes a cross strand contact with the opposite strand of the binding site that is not shown here. The sequence for the binding site was derived from ChIP-seq data (Nakahashi et al., 2013). The binding site was partitioned into three segments: 5' flanking (gray-line), core (black-line), and 3' flanking (light gray line). The position of each nucleotide within each segment are numbered. Dashes indicate known DNA-protein contacts (black) and theoretical DNA-protein contacts (gray) between the zinc finger array and the binding site. Zinc fingers 3-7 of the array (white) make protein-DNA contacts with the core sequence (bold, black lined). There was a possible 5-6 base pair gap (represented by horizontal dashed lines) between zinc finger 8 and zinc fingers 9-11 as suggested by ChIP-exo and DNase I footprinting of CTCF bound DNA fragments (Hashimoto, H. et al., 2017). Note CTCF binds to its target site in the 3'-5' direction with the N-terminal side of the protein binding to the 3' end of the binding site.

FIG. 2: Diagram of B2H Beta-galactosidase reporter assay. The B2H reporter assay used Gal11P-mediated recruitment of Gal4 to indicate binding. *E. coli* is transformed with two plasmids: one plasmid encoded for both a zinc finger-Gal11P fusion and an alpha N-terminal domain of RNA polymerase (α -NTD)-Gal4 fusion; the second plasmid contained a modifiable binding sequence upstream of a weak promoter that drives the expression of the *lacZ* gene, which encodes for β -galactosidase. A zinc finger-Gal11P fusion that was able to bind to the target sequence recruited the α -NTD-Gal4 fusion to the promoter, thereby inducing the expression of *lacZ*. This increase in β -galactosidase levels was detected by a simple colorimetric ONPG-based assay. The CTCF zinc finger array-gal11P fusion was bound to a CTCF binding site in this diagram, recruiting the α -NTD-Gal4 fusion to the promoter region upstream of *lacZ*, leading to expression.

FIG. 3: Fold activation in the B2H B-gal assay was greatest when CTCF zinc fingers 1-11 of 11 finger array interacts with full length target site. Five target sites (sequence indicated in the legend) were tested along with the full CTCF zinc finger array and four different subsets (indicated on the x-axis). The core sequence (black and bolded) which is the most highly conserved sequence of CTCF binding sites was tested independently and with different quantities of flanking sequence as derived from Hashimoto, H. et al. Mol. Cell. 2017 (black and light gray); Persikov, A and Singh, M. NAR. 2014 (medium gray); and Nakahashi, H. et al., Cell Rep. 2013 (very light gray and dark gray). Positive control reflects binding activity of a known 3-finger zinc finger that binds strongly in bacterial and human contexts to a known sequence. The negative control reflects baseline beta-galactosidase levels when the alpha N-terminal domain of RNA polymerase (α -NTD)-Gal4 fusion is not directly recruited to the promoter of *lacZ*. This baseline was used to calculate fold activation when the CTCF zinc finger array is fused to gal11P.

FIG. 4: CTCF zinc finger array is sensitive to sequence changes at certain positions of the core region within the CTCF binding site. Each of the four possible nucleotides at each position of the 40bp reference CBS were tested for ability to bind the CTCF zinc finger array in the B2H y. Fold activation reflects binding activity above background β -galactosidase levels (Background β -gal levels are obtained from the levels of β -gal from samples with each binding site in the presence of the gal4-RNA polymerase fusion with no zinc finger array fused to gal11P). The reference sequence above is partitioned into three segments: 5' flanking (dark gray lined), core (black lined), and 3' flanking (gray lined). The position of each nucleotide within each segment

are numbered. Dashes indicate known DNA-protein contacts (black) and theoretical DNA-protein contacts (gray) between the zinc finger array and the binding site. Core sequence 1-15 of the binding site (black, bold) interacts with zinc finger 3-7 of the array (white, black outline) and appear to be most sensitive to changes in the binding sequence. Alterations to the 5' flanking sequence as well as the 3' flanking sequence did not negatively impact binding.

FIG. 5: Maximizing binding potential of the CTCF binding site. Modifications were made to the reference binding site (bottom) to combine nucleotide changes that, individually, showed increased binding activity of the CTCF zinc finger array. The core sequence motif is bold while changes made are underlined. Binding activity of the 11-finger CTCF zinc finger array was quantified in the B2H Beta-galactosidase reporter assay in triplicate. Fold activation reflects binding activity above background levels when no DNA binding protein is present.

FIG. 6: Diagram of B2H Beta-lactamase inhibitor selection. The selection system contained the same components as the reporter system except successful binding of the zinc finger array to the CBS drove BlaC expression, an inhibitor of the beta-lactamase class of antibiotics, instead of lacZ. Expression of BlaC allowed for growth on Carbenicillin plates. The selection was driven by the addition of Clavulanic acid, an inhibitor of beta lactamase inhibitors. Low level expression of BlaC can result in growth on Carbenicillin plates, but the addition of clavulanic acid inhibits BlaC activity and results in the depletion of false positives and further enrichment of strong binders to any modification made to the binding site. Libraries of mutations in the zinc finger array fused to gal11P were selected for binders to an altered binding sequence through low stringency conditions followed by selection on a gradient of clavulanic acid. Growth on the highest stringency end of the gradient indicated variants in the zinc finger array that are strong binders to the new binding sequence.

FIGS. 7A-C: Binding activity of variants on altered CTCF binding sites. Variants picked from the high stringency gradient of the selective plates were tested for binding activity on sequences representing all four possible nucleotides at position 2 of the core sequence (gray star). Amino acid sequence of variants pulled out of the selection were listed above the heat map and the nucleotide present at position 2 of the core sequence was indicated on the y-axis. FIG 7A: The nucleotide at position 2 is T. FIG 7B: The nucleotide at position 2 is A. FIG 7C: The nucleotide at Binding was quantified by the beta-galactosidase reporter system and colorimetric ONPG assay. Binding activity of wild-type CTCF zinc finger array on the wild-type binding site

sequence was indicated by the white dot. A diagram of the ZF7 alpha recognition helix for each nucleotide change is on the left. It included the amino acid residues interacting with the triplet in the binding sequence. The amino acid at position 3 of the alpha helix was varied in the library and is indicated by an 'X'.

FIG. 8: Increasing the variation within the recognition helix produced stronger binders. Four amino acids were targeted for variance in the library to allow for more flexibility in the selection and generate stronger binders to the modified binding site of choice. ZF7 targeting a C:G change at position 2 (gray star) of the core sequence was selected for variants using the expanded approach. Each amino acid codon was replaced with 'VNS' codons at the indicated sites ('X'). Twelve colonies were picked from the high-stringency end of the selection and tested for their ability to bind to the CTCF binding site when the indicated nucleotide is at position 2 of core sequence. Amino acid sequence of the variants selected are listed on the x-axis and the nucleotide at position two of the core sequence is on the y-axis. Wild-type zinc finger array binding activity on wild-type binding sequence is indicated by the white dot.

FIGS. 9A-C: Selected variants binding altered binding sites sequence at position 3 of core motif in CBS. Selections performed on library of variants centered around alterations in position -1 to 3 of recognition helix in ZF7 of the 11 finger CTCF zinc finger array. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 3 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated by dashed lines. (A) Selections performed on A:T change in the binding site, (B) A:G change, (C) A:C change. Most variants pulled out had relaxed binding specificity instead of altered specificity.

FIGS. 10A-B: Selections performed targeting sequence changes at position 5 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position 2 to 6 of the ZF6 recognition helix, leaving the 4th position unchanged. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 5 of the core motif in the core motif of the CBS (gray star). Direct protein-DNA contacts were indicated by dashed lines. (A) Selections performed on C:T change in the binding site, (B) C:G change. No variants grew beyond the low stringency end of the gradient on selection plates for C:A change and were considered weak/insufficient binders. Most variants

pulled out had relaxed binding specificity instead of altered specificity with the exception of 'THMKR' targeting C:G change in the binding sequence.

FIGS. 11A-C: Selections performed targeting sequence changes at position 6 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position -1 to 3 of ZF6 recognition helix. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 6 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated by dashed lines. (A) Selections performed on A:T change in the binding site, (B) A:G change, (C) A:C change. Variants analyzed from the A:T selection had relaxed binding profile while variants from A:G selection showed strong binding for only the changed nucleotide. No good binders were identified in the A:C selection.

FIGS. 12A-C: Selections performed targeting sequence changes at position 7 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position 4 to 6 of ZF5 recognition helix, leaving the 4th position unchanged. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 7 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated by a line. (A) Selections performed on G:T change in the binding site, (B) G:A change, (C) G:C change.

FIGS. 13A-C: Selections performed targeting sequence changes at position 8 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position 2 to 6 of ZF5 recognition helix, leaving the 4th position unchanged. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 8 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated by a line. (A) Selections performed on G:T change in the binding site, (B) G:A change, (C) G:C change. Note the different variants that appear with the same library being used to bind to the same changes in the sequence, but in a different position on the binding site.

FIGS. 14A-C: Selections performed targeting sequence changes at position 10 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position 2 to 6 of ZF4 recognition helix, leaving the 4th position unchanged. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 10 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated

by a line. (A) Selections performed on G:T change in the binding site, (B) G:A change, (C) G:C change. G:C selection did not produce any growth at the high stringency end of the gradient selective plates. Binding data reflects colonies picked from mid-tier region, which is why they did not perform well as binders. White dot indicates wild-type CTCF zinc finger array binding activity on wild-type binding sequence

FIGS. 15A-C: Selections performed targeting sequence changes at position 11 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position 2 to 6 of ZF4 recognition helix, leaving the 4th position unchanged. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 11 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated by a line. (A) Selections performed on G:T change in the binding site, (B) G:A change, (C) G:C change.

FIGS. 16A-C: Selections performed targeting sequence changes at position 13 of the core motif in the CBS. Selections performed on library of variants centered around alterations in position 2 to 6 of ZF3 recognition helix, leaving the 4th position unchanged. 'VNS' codons were introduced at positions indicated by 'X' and selected against three different nucleotide changes at position 13 of the core motif in the CBS (gray star). Direct protein-DNA contacts are indicated by a line. (A) Selections performed on G:T change in the binding site, (B) G:A change, (C) G:C change.

FIG. 17: Binding activity of multi-finger variants on multiple sequence changes to the CBS. Diagram of the recognition helices of zinc finger 4-7 out of the 11 finger array, binding to their respective triplets in the core motif of the CBS. Altered amino acids are indicated by 'X' and nucleotide changes to the wild-type CBS are indicated by a gray star in the diagram and by bolded letters. ZF1-3 and ZF8-11 were unmodified in this library. Protein-DNA contacts are indicated by lines between the ZF recognition helices and the CBS sequence. Wild-type CTCF 11-finger zinc finger array binding strength to wild-type CBS is indicated by a white dot. The amino acid sequence of each variant recognition helix in ZF4-7 are listed on the y-axis and binding activity on the modified CBS (changes in red) or the wild-type CBS are reflected by B2H β -gal reporter assay.

FIG. 18: Binding activity of multi-finger variants on multiple sequence changes to the CBS. The same selection as before except now there is a C:G change at position 2 of the CBS,

where previously there was a C:A change. Variants pulled out of this selection had binding activity on the modified CBS without binding to the wild-type CBS. Wild-type 11-finger ZF array only showed binding activity on wild-type CBS (white dot) and no ability to bind to the modified CBS. Interestingly, the dominant variant selected for in the library contained a mutation that occurs at position 9 of the recognition helix that was either introduced during oligo synthesis (0.05% chance of the wrong nucleotide at each position) or through PCR while constructing these libraries.

FIG. 19: Binding activity of multi-finger variants on multiple sequence changes to the CBS. Variants from individual pooled high stringency selections were stitched together and selected against three changes to the wild-type CBS (indicated by gray stars or bolded). Variants were assayed for binding on the modified CBS and the wild-type CBS alongside wild-type CTCF zinc finger array. The variants picked out of the selection were able to bind to the modified CBS, but not the wild-type sequence. Inversely, the wild-type zinc finger array was able to bind to the wild-type CBS, but not the modified one.

FIG. 20: Binding activity of multi-finger variants on multiple sequence changes to the CBS. Variants from individual pooled high stringency selections were stitched together and selected against three changes to the wild-type CBS (indicated by gray stars or bolded). Variants were assayed for binding on the modified CBS and the wild-type CBS alongside wild-type CTCF zinc finger array. The variants picked out of the selection were able to bind to the modified CBS, but not the wild-type sequence. Inversely, the wild-type zinc finger array was able to bind to the wild-type CBS, but not the modified one.

FIG. 21: Binding activity of multi-finger variants on multiple sequence changes to the CBS. Variants from individual pooled high stringency selections were stitched together and selected against three changes to the wild-type CBS (indicated by gray stars or bolded). Variants were assayed for binding on the modified CBS and the wild-type CBS alongside wild-type CTCF zinc finger array. The variants picked out of the selection were able to bind to the modified CBS, but not the wild-type sequence. Inversely, the wild-type zinc finger array was able to bind to the wild-type CBS (white dot), but not the modified one.

FIG. 22: Wild-type CTCF has binding activity to wild-type CTCF target site and no binding activity to two variant target sites. To confirm endogenous CTCF binds to the wild-type CBSs and not the variant binding sites, as seen in the B2H assay, in a human cell context, we

harvested K562 cells, a human erythroleukemia cell line, and examined CTCF binding through ChIP-qPCR. CTCF was assayed for binding to a known CTCF target site and to two endogenous variant binding site sequences using a CTCF specific antibody to enrich for genomic DNA crosslinked to CTCF. Two sets of qPCR primers were designed for each binding site (indicated by 1.1, 1.2, etc). Binding was determined by enrichment of target site above 1% input of crosslinked and sonicated sample not treated with antibody, which is to represent the levels of the site of interest as a fold increase over the frequency of the site of interest in a sample unenriched with antibody. Antibody based enrichment of each sample is quantified by fold enrichment above untreated, and therefore unenriched, input. The negative control reflects background qPCR amplification levels of a target site that CTCF does not bind to. Anything above this negative level is considered enriched indicating CTCF binding while anything below is considered to not be unenriched, and therefore no binding by CTCF. Wild-type CTCF binds to the wild-type target site with no detectable binding to the variant binding sites as predicted by the bacterial B2H reporter assay

FIGS. 23A-B: Exogenous wild-type and variant CTCF binding activity in human cells. Two endogenous variant binding site sequences, matching one of the five variant binding sites that CTCF variants were selected on, were identified in the human genome (Variant site 1 and Variant site 2). Both wild-type CTCF with a 3xHA tag and one of the 3xHA tagged engineered CTCF variants, selected to bind to the variant binding site sequence of Variant site 1 and Variant site 2, were assayed for binding in human cells through ChIP-qPCR. **FIG. 23A:** 3xHA tagged wild-type CTCF binds to wild-type CTCF binding site and does not bind to either variant binding site. Human K562 cells were transfected with plasmid expressing 3xHA tagged CTCF and processed with HA antibody to enrich specifically for the exogenous CTCF (3xHA tagged) and not endogenous CTCF (no tag) binding. A negative control is provided to show ChIP-qPCR levels with no enrichment for a region that is not occupied by CTCF. These results demonstrate exogenous wild-type CTCF has the same binding activity as endogenous CTCF. **FIG. 23B:** 3xHA tagged variant CTCF binds to variant binding sites and does not bind to wild-type CTCF binding site. K562 cells expressing variant CTCF tagged with 3xHA were analyzed by ChIP-qPCR and treated with HA specific antibody. The same sites as in FIGs. 22 and 23A were investigated for variant CTCF binding. The variant CTCF could bind to the variant sites as

indicated by enrichment with variant specific HA antibody and no detectable binding was seen at the wild-type binding site as indicated by lack of HA antibody-based enrichment.

FIGS. 24A-B: Changes in gene expression relative to wild-type control of genes located around variant binding sites. A variant CTCF selected to the G3 binding site sequence and variant CTCF selected to the Other binding site sequence were expressed in wild-type K562s. The variant CTCFs were fused to GFP and RNA was isolated from GFP+ cells 72 hours post nucleofection. cDNA was generated from the RNA and quantified by RT-qPCR. Gene expression levels across samples were normalized to a house keeping gene (HPRT). Changes in gene expression are relative to gene expression levels in wild-type K562s expressing wild-type CTCF tagged with GFP. **FIG. 24A.** Changes in gene expression of genes around G3 variant binding site in the presence of variant CTCF relative to the wild-type CTCF control. **FIG. 24B.** Changes in gene expression of genes around Other variant binding site relative to the wild-type control.

FIG 25: Introduction of variant binding sites upstream of MYC leads to reduction of Endogenous MYC expression. The CTCF binding site ~2kb upstream of the MYC TSS was replaced with one of six different sequences used for CTCF variant selections (listed in table). The introduction of these sequences with 4-6 nucleotide changes from the wild-type CTCF binding site sequence result in a reduction of endogenous MYC expression to the same levels as when the CTCF binding site is deleted and loop formation is disrupted. WT_6 sequence has 4 point mutations from the native CTCF binding site, but these changes should have no impact on wild-type CTCF binding as indicated by results from the B2H reporter assay. This appears to be the case as MYC expression levels in the WT_6 cell line are comparable to wild-type K562 MYC expression levels. Because K562 vitality is linked to MYC expression, all variant cell lines were generated in a K562 cell line with exogenous MYC expressed off of a separate PGK promoter (exoMYC.K562).

FIGs. 26A-B: Variant CTCFs are able to bind the engineered G3 variant binding site and recover MYC expression. CTCF variants selected to bind to the G3 variant binding site sequence were expressed in the G3_3.K562 cell line. Cells were analyzed for MYC expression and CTCF occupancy on the DNA 72 hours post nucleofection. Residues of ZF helix of the variant and wild-type (indicated by (wt)) are listed in the legend. G3 binding site sequence and interacting recognition helix of the CTCF zinc finger array is also diagramed. **FIG. 26A.** Endogenous MYC levels are recovered to wild-type levels in the G3_3 cell line when CTCF variants are expressed.

Endogenous MYC expression levels were quantified by RT-qPCR and are relative to reduced endogenous MYC levels of G3_3 cell line. Endogenous MYC levels from the exoMYC.K562 cell line without any alterations to the CTCF binding site is shown as a positive control (separated by dashed lines). **FIG. 26B.** CTCF variants are able to bind to the introduced variant binding site in G3_3 cell line while the wild-type CTCF does not. CTCF Ab specific enrichment captures both wild-type and variant CTCF while HA Ab will only detect HA-tagged CTCF (transiently expressed). exoMYC.K562 is included as a control for ChIP-qPCR and is separated by dashed line. exoMYC.K52 has the native sequence at the CTCF binding site upstream of MYC and should demonstrate wild-type CTCF binding. The exogenously expressed CTCFs (variant and wild-type) are HA tagged and expressed in the G3_3 cell line. ChIP-qPCR was performed to investigate CTCF binding to the variant CTCF site replacing the wild-type site ~2kb upstream of MYC (MYC site). An endogenous G3 site elsewhere in the genome and a region with no known CTCF binding served as a positive and negative control respectively. The variant CTCFs are able to bind to the variant site as indicated by enrichment with both CTCF and HA antibody, while the wild-type CTCF does not.

FIGs. 27A-B: Variant CTCFs are able to bind the engineered A3 variant binding site and recover MYC expression. CTCF variants selected to bind to the A3 variant binding site sequence were expressed in the A3_4.K562 cell line. Cells were analyzed for MYC expression and CTCF occupancy on the DNA 72 hours post nucleofection. Residues of ZF helix of the variant and wild-type (indicated by (wt)) are listed in the legend. A3 binding site sequence and interacting recognition helix of the CTCF zinc finger array is also diagramed. **FIG. 27A.** Endogenous MYC levels are recovered to wild-type levels in the A3_4 cell line when CTCF variants are expressed. Endogenous MYC expression levels were quantified by RT-qPCR and are relative to reduced endogenous MYC levels of A3_4 cell line. Endogenous MYC levels from the exoMYC.K562 cell line without any alterations to the CTCF binding site is shown as a positive control (separated by dashed lines). **FIG. 27B.** CTCF variants are able to bind to the introduced variant binding site in A3_4 cell line while the wild-type CTCF does not. CTCF Ab specific enrichment captures both wild-type and variant CTCF while HA Ab will only detect HA-tagged CTCF (transiently expressed). exoMYC.K562 is included as a control for ChIP-qPCR and is separated by dashed line. exoMYC.K52 has the native sequence at the CTCF binding site upstream of MYC and should demonstrate wild-type CTCF binding. The exogenously expressed CTCFs

(variant and wild-type) are HA tagged and expressed in the A3__4 cell line. ChIP-qPCR was performed to investigate CTCF binding to the variant CTCF site replacing the wild-type site ~2kb upstream of MYC (MYC site). An endogenous A3 site elsewhere in the genome and a region with no known CTCF binding served as a positive and negative control respectively. The variant CTCFs are able to bind to the variant site as indicated by enrichment with both CTCF and HA antibody above the negative control, while the wild-type CTCF does not bind.

FIG. 28: Variant CTCFs recover MYC expression of the Other_10 variant binding site cell line. CTCF variants selected to bind to the Other variant binding site sequence were expressed in the Other_10.K562 cell line. Cells were analyzed for MYC expression 72 hours post nucleofection. Residues of ZF helix of the variant and wild-type CTCFs (indicated by (wt) are listed in the legend. Other binding site sequence and interacting recognition helix of the CTCF zinc finger array is also diagramed. A. Endogenous MYC levels are recovered to wild-type levels in the Other_10 cell line when CTCF variants are expressed. Endogenous MYC expression levels were quantified by RT-qPCR and are relative to reduced endogenous MYC levels of Other_10 cell line. Endogenous MYC levels from the exoMYC.K562 cell line without any alterations to the CTCF binding site is shown as a positive control (separated by dashed lines).

FIG. 29: Variant CTCFs recover MYC expression of the Aother_2 variant binding site cell line. CTCF variants selected to bind to the Aother variant binding site sequence were expressed in the Aother_2.K562 cell line. Cells were analyzed for MYC expression 72 hours post nucleofection. Residues of ZF helix of the variant and wild-type CTCFs (indicated by (wt) are listed in the legend. Aother binding site sequence and interacting recognition helix of the CTCF zinc finger array is also diagramed. A. Endogenous MYC levels are recovered to wild-type levels in the Aother_2 cell line when CTCF variants are expressed. Endogenous MYC expression levels were quantified by RT-qPCR and are relative to reduced endogenous MYC levels of Aother_2 cell line. Endogenous MYC levels from the exoMYC.K562 cell line without any alterations to the CTCF binding site is shown as a positive control (separated by dashed lines).

DETAILED DESCRIPTION

To date, there are no engineered CTCF variants available that are designed to bind to mutant CBSs with higher affinity than wild-type CTCF. Therefore, there is a need for

engineered CTCF variants that can bind to mutant CBSs with higher affinity than wild-type CTCF.

The present disclosure is based, at least in part, on the discovery that CTCF variants with alterations in the zinc finger array can be engineered to recognize CBSs that harbor one or more point mutations, i.e., mutant CBSs.

CTCF

CCCTC-binding factor (CTCF) is a multi-domain protein that acts as an essential genome organizer by maintaining higher-order chromatin structure while also having a role in cell differentiation and the promotion or repression of gene expression. CTCF maintains topologically associated domains (TADs) spanning megabases of the genome as well as smaller scale Sub-TADs leading to fine-tuned gene insulation or gene activation within gene clusters. In addition, CTCF has been found to regulate mRNA splicing by influencing the rate of transcription and more recently been implicated in promoting homologous recombination repair at double-strand breaks. Wild type CTCF binds throughout the genome via an 11 finger zinc finger array that recognizes canonical CTCF binding sites (CBSs).

Wild-type CTCF ZF arrays comprise the following sequences at ZFs 3-6 positions -1 to +6:

ZF3 positions -1 to +6:	TSGELVR (SEQ ID NO:1)
ZF4 positions -1 to +6:	EVSKLKR (SEQ ID NO:2)
ZF5 positions -1 to +6:	DTYKLKR (SEQ ID NO:3)
ZF6 positions -1 to +6:	QSGTMKM (SEQ ID NO:4)
ZF7 positions -1 to +6:	RKSDLGV (SEQ ID NO:5)

A wild-type CTCF has an amino acid sequence that has greater than 80%, greater than 85%, greater than 90%, greater than 95%, greater than 96%, greater than 97%, greater than 98% or greater than 99% sequence identity as compared to the amino acid sequence shown below:

MEGDAVEAIVEESETFIKGKERKTYQRRREGGQEEDACHLPQNQTDGGEVV
QDVNSSVQMVMMEQLDPTLLQMKTEVMEGTVAPEAEAAVDDTQIITLQVV
NMEEQPINIGELQLVQVPVPVTVPVATTSSVEELQGAYENEVSKEGLAESEPMI
CHTLPLPEGFQVVKVGANGEVETLEQGELPPQEDPSWQKDPDYQPPAKKTK
KTKKSKLRYTEEGKDVDVSVYDFEEEQQEGLLSEVNAAEKVVGNMKPPKPTK

IKKKGVKKTFQCELCSYTCPRRSNLDRHMKSHTDERPHKCHLCGRAFRVTL
 LRNHLNTHTGTRPHKCPDCDMAFVTSGELVRHRRYKHTHEKPFKCSMCDYA
 SVEVSKLKRHIRSHTGERPFQCSLCSYASRDYKLRHMRTHSGEKPYECYIC
 HARFTQSGTMKMHILQKHTENVAKFHCPHCDTVIARKSDLGVHLRKQHSYIE
 QGKKCRYCDAVFHERYALIQHQKSHKNEKRFKCDQCDYACRQERHMIMHK
 RTHTGEKPYACSHCDKTFRQKQLLDMHFKRYHDPNFVPAAFVCSKCGKTFT
 RRNTMARHADNCAGPDGVEGENGGGETKKS KRGRKRKMRSKKEDSSDSENA
 EPDLDDNEDEEEPAVEIEPEPEPQPVTPAPPPAKKRRGRPPGRTNQPKQNQPT
 AIIQVEDQNTGAIENIIVEVKKEPDAEPAEGEEEAQPAATDAPNGDLTPEMIL
 SMMDR

For the purpose of comparing two different nucleic acid or polypeptide sequences, one sequence (test sequence) may be described to be a specific percentage identical to another sequence (comparison sequence). The percentage identity can be determined by the algorithm of Karlin and Altschul, *Proc. Natl. Acad. Sci. USA*, 90:5873-5877 (1993), which is incorporated into various BLAST programs. The percentage identity can be determined by the "BLAST 2 Sequences" tool, which is available at the National Center for Biotechnology Information (NCBI) website. See Tatusova and Madden, *FEMS Microbiol. Lett.*, 174(2):247-250 (1999). For pairwise DNA-DNA comparison, the BLASTN program is used with default parameters (e.g., Match: 1; Mismatch: -2; Open gap: 5 penalties; extension gap: 2 penalties; gap x_dropoff: 50; expect: 10; and word size: 11, with filter). For pairwise protein-protein sequence comparison, the BLASTP program can be employed using default parameters (e.g., Matrix: BLOSUM62; gap open: 11; gap extension: 1; x_dropoff: 15; expect: 10.0; and wordsize: 3, with filter). Percent identity of two sequences is calculated by aligning a test sequence with a comparison sequence using BLAST, determining the number of amino acids or nucleotides in the aligned test sequence that are identical to amino acids or nucleotides in the same position of the comparison sequence, and dividing the number of identical amino acids or nucleotides by the number of amino acids or nucleotides in the comparison sequence. When BLAST is used to compare two sequences, it aligns the sequences and yields the percent identity over defined, aligned regions. If the two sequences are aligned across their entire length, the percent identity yielded by the BLAST is the percent identity of the two sequences. If BLAST does not align the two sequences over their entire length, then the number of identical amino acids or nucleotides in the unaligned regions of

the test sequence and comparison sequence is considered to be zero and the percent identity is calculated by adding the number of identical amino acids or nucleotides in the aligned regions and dividing that number by the length of the comparison sequence. Various versions of the BLAST programs can be used to compare sequences, e.g., BLAST 2.1.2 or BLAST+ 2.2.22.

CTCF Binding Sites (CBSs)

The CBS is typically 40bp in length with a highly conserved 15bp core sequence (or core motif). Sequence flanking the core sequence is significantly less well conserved, but still important for CTCF binding at sites throughout the genome (FIG. 1).

Wild type CTCF binds to a “consensus CBS motif” contains the following core sequence: 5'-NCDNHNGRNGDNNNN-3'.

In one embodiment, the consensus CBS motif contains the following core sequence: 5'-CCAGCAGGGGGCGCT-3' (SEQ ID NO:6). Other core sequences that are known in the art.

It is not known if the nucleotides flanking the core sequence are bound by the 11 finger ZF array present within CTCF. Co-crystal structures of the 11-finger Zinc Finger (ZF) array bound to a consensus CTCF Binding Sequence (CBS) suggests that only ZFs 3-7 of the 11-finger ZF array appear to bind directly to the highly conserved core sequence while ZFs 8-11 and 1-2 do not appear to mediate sequence-specific contacts. Progressive truncations of the ZF array suggest that ZFs 8-11 and ZFs 1-2 may improve DNA-binding of CTCF to CBSs and DNaseI foot printing, as well as ChIP-Seq and ChIP-Exo data, suggests that ZFs 9-11 may make important protein-DNA contacts (Rhee and Pugh, Cell (2011); Nakahashi et al., Cell Reports (2013)). Interestingly, the co-crystal structure of the CTCF Z array bound to a CBS only contains zinc finger 2-9 with the other fingers not visible in the structure, consistent with the idea that zinc fingers interacting with flanking regions of the motif may not make stable contacts with the DNA (Hashimoto, et al., Molecular Cell (2017)). Thus, it remains unclear what impact all 11 fingers of the array have on DNA binding activity of CTCF and if all zinc fingers, or a subset, contact the DNA.

CTCF binding is sensitive to changes in the conserved 15bp core motif of the CBS, where, in mice, single nucleotide changes at certain positions can lead to loss of CTCF binding (Nakahashi et al., Cell Reports (2013)). CTCF binding sites have been reported to be mutational hotspots in cancer with cancer-associated mutations localized to the core sequence of the CTCF binding site in primary samples from gastrointestinal cancer patients and with accompanying

atypical gene expression profiles of oncogenic and tumor suppressor genes (Guo et al., *Nature Communications* (2018)). Small deletions of CTCF binding sites have also been shown to lead to loss of expression of genes such as MYC and PTGS2, which both play a role in cancer development (Schuijers et al., *Cell Reports* (2018); Kang et al., *Oncogene* (2015)).

Methods described herein can be used to select and generate engineered CTCF variants comprising a plurality of zinc fingers, where the selected polypeptide has at least one amino acid residue in at least one zinc finger that differs in sequence from a wild-type CTCF, and where the engineered CTCF variant binds to a DNA sequence of interest (e.g., CBS harboring at least one mutation in the consensus CBS sequence) but does not bind to a consensus CBS. Using methods of the present invention, a scaffold polypeptide is re-engineered into a new scaffold-based zinc-finger polypeptide that has different structural and functional features, such that the new polypeptide binds to a sequence of interest but does not bind to a naturally occurring DNA binding site of the scaffold protein.

The term “zinc finger” or “Zf” refers to a polypeptide having DNA binding domains that are stabilized by zinc. The individual DNA binding domains are typically referred to as “fingers.” A Zf protein has at least one finger, preferably 2 fingers, 3 fingers, or 6 fingers. A Zf protein having two or more Zfs is referred to as a “multi-finger” or “multi-Zf” protein. Each finger typically comprises an approximately 30 amino acid, zinc-chelating, DNA-binding domain. An exemplary motif characterizing one class of these proteins is -Cys-(X) (2-4)-Cys-(X) (12)-His-(X) (3-5)-His (SEQ ID NO:7), where X is any amino acid, which is known as the “C(2)H(2)class.” A single Zf of this class typically consists of an alpha helix containing the two invariant histidine residues co-ordinated with zinc along with the two cysteine residues.

The term “bind to” or “binding” with respect to a nucleic acid binding factor and its target nucleic acid, e.g., CTCF (variant or wild-type) and CBS, refers to sequence-dependent binding of the nucleic acid binding factor to the target nucleic acid sequence of a nucleic acid through intermolecular interactions, e.g., ionic, covalent, London dispersion, dipole-dipole, or hydrogen bonding, in such a way that the binding allows the nucleic acid binding factor to mediate a biologically significant function, e.g., transcriptional activation, recruitment of other proteins to the binding site, and/or alteration of chromatic structure. Such binding can result in modulation of expression of genes, such as activation, overexpression, suppression, or inactivation of gene expression.

The term “does not bind to” with respect to a nucleic acid binding factor and its target nucleic acid, e.g., CTCF (variant or wild-type) and CBS, refers to the lack of sequence-specific binding of the nucleic acid binding factor to a nucleic acid through intermolecular interactions, e.g., ionic, covalent, London dispersion, dipole-dipole, or hydrogen bonding, as a result of the lack of presence of a target sequence in the nucleic acid (e.g., due to one or more point-mutations in the CBS). Such non-binding does not allow the nucleic acid binding factor to mediate a biologically significant function, e.g., transcriptional activation, DNA modification, DNA cleavage, recruitment of other proteins to the binding site, and/or alteration of chromatic structure.

Each finger within a Zf protein binds to from about two to about five base pairs within a DNA sequence. Typically a single Zf within a Zf protein binds to a three or four base pair “subsite” within a DNA sequence. Accordingly, a “subsite” is a DNA sequence that is bound by a single zinc finger. A “multi-subsite” is a DNA sequence that is bound by more than one zinc finger, and comprises at least 4 bp, preferably 6 bp or more. A multi-Zf protein binds at least two, and typically three, four, five, six or more subsites, i.e., one for each finger of the protein.

Compositions and Methods

Described herein are engineered CTCF variants that can bind to mutant CBSs with higher affinity than wild-type CTCF. The engineered CTCF variants can be used in regulating genes that are under the control of mutant CBSs (CBSs having at least one nucleic acid that is different in sequence from the nucleic acid sequence of a consensus CBS). The CTCF variants have at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF.

Exemplary engineered CTCF variants include those that contain:

(1) the amino acid sequence DHLQT (SEQ ID NO:8), EHLNV (SEQ ID NO:9), AHLQV (SEQ ID NO:10), EHLRE (SEQ ID NO:11), DHLQV (SEQ ID NO:12), EHLKV (SEQ ID NO:13), DHLQV (SEQ ID NO:14), EHLVV (SEQ ID NO:15), DHLRT (SEQ ID NO:16), DHLAT (SEQ ID NO:17), or DHLQT (SEQ ID NO:18) at ZF7 positions +2 to +6;

(2) the amino acid sequence DHLQT (SEQ ID NO:19), EHLNV (SEQ ID NO:20), AHLQV (SEQ ID NO:21), EHLRE (SEQ ID NO:22), DHLQV (SEQ ID NO:23), EHLKV (SEQ ID NO:24), DHLQV (SEQ ID NO:25), EHLVV (SEQ ID NO:26), DHLRT (SEQ ID NO:27), DHLAT (SEQ ID NO:28), or DHLQT (SEQ ID NO:29) at ZF7 positions +2 to +6;

(3) the amino acid sequence NAMKR (SEQ ID NO:30), EHMGR (SEQ ID NO:31), DHMNR (SEQ ID NO:32), THMKR (SEQ ID NO:33), EHMRR (SEQ ID NO:34), or THMNR (SEQ ID NO:35) at ZF6 positions +2 to +6;

(4) the amino acid sequence MNES (SEQ ID NO:36), HRES (SEQ ID NO:37), RPDY (SEQ ID NO:38), RTDI (SEQ ID NO:39), or RHDT (SEQ ID NO:40) at ZF6 positions -1 to +3;

(5) the amino acid sequence HGLKV (SEQ ID NO:41), HRLKE (SEQ ID NO:42), HALKV (SEQ ID NO:43), SRLKE (SEQ ID NO:44), DGLRV (SEQ ID NO:45), HTLKV (SEQ ID NO:46), or NRLKE (SEQ ID NO:47) at ZF5 positions +2 to +6;

(6) the amino acid sequence ATLKR (SEQ ID NO:48), QALRR (SEQ ID NO:49), GGLVR (SEQ ID NO:50), HGLIR (SEQ ID NO:51), ANLSR (SEQ ID NO:52), TGLTR (SEQ ID NO:53), HGLVR (SEQ ID NO:54), GGLTR (SEQ ID NO:55), HTLRR (SEQ ID NO:56), TVLKR (SEQ ID NO:57), ADLKR (SEQ ID NO:58), or HGLRR (SEQ ID NO:59) at ZF5 positions +2 to +6;

(7) the amino acid sequence AHLRK (SEQ ID NO:60), AKLRV (SEQ ID NO:61), GGLGL (SEQ ID NO:62), AKLRI (SEQ ID NO:63), TKLKV (SEQ ID NO:64), or SKLRV (SEQ ID NO:65) at ZF4 positions +2 to +6;

(8) the amino acid sequence ATLRR (SEQ ID NO:66), RRLDR (SEQ ID NO:67), TNLRR (SEQ ID NO:68), ANLRR (SEQ ID NO:69), GNLTR (SEQ ID NO:70), AMLKR (SEQ ID NO:71), HMLTR (SEQ ID NO:72), AMLRR (SEQ ID NO:73), or TMLRR (SEQ ID NO:74) at ZF4 positions +2 to +6;

(9) the amino acid sequence QQLIV (SEQ ID NO:75), SQLIV (SEQ ID NO:76), QQLLV (SEQ ID NO:77), GELVV (SEQ ID NO:78), QQLLI (SEQ ID NO:79), GQLIV (SEQ ID NO:80), GQLTV (SEQ ID NO:81), TELII (SEQ ID NO:82), QGLLV (SEQ ID NO:83), QQLLT (SEQ ID NO:84), GQLLT (SEQ ID NO:85), GELLT (SEQ ID NO:86), or QQLLI (SEQ ID NO:87) at ZF3 positions +2 to +6;

(10) the amino acid sequence AKLKK (SEQ ID NO:88), AKLRK (SEQ ID NO:89), AHLRV (SEQ ID NO:90), AKLRV (SEQ ID NO:91), or SKLRL (SEQ ID NO:92) at ZF4 positions +2 to +6; the amino acid sequence ERLRV (SEQ ID NO:93), NRLKV (SEQ ID NO:94), SRLKE (SEQ ID NO:95), or NRLKV (SEQ ID NO:96) at ZF5 positions +2 to +6; the amino acid sequence RPDY (SEQ ID NO:97), RTET (SEQ ID NO:98), or RADV (SEQ ID NO:99) at ZF6 positions -1 to +3; and the amino acid sequence DNLLA (SEQ ID NO:100),

SNLLV (SEQ ID NO:101), DNLMA (SEQ ID NO:102), or DNLRV (SEQ ID NO:103) at ZF7 positions +2 to +6;

(11) the amino acid sequence GHLKK (SEQ ID NO:104), AHLRK (SEQ ID NO:105), or GKLRI (SEQ ID NO:106) at ZF4 positions +2 to +6; the amino acid sequence SRLKE (SEQ ID NO:107), DALRR (SEQ ID NO:108), DGLKR (SEQ ID NO:109), or TRLRE (SEQ ID NO:110) at ZF5 positions +2 to +6; the amino acid sequence at RPDT (SEQ ID NO:111) or RTEN (SEQ ID NO:112) at ZF6 positions -1 to +3; and the amino acid sequence EHLKV (SEQ ID NO:113), DHLLA (SEQ ID NO:114), or HHLDV (SEQ ID NO:115) at ZF7 positions +2 to +6;

(12) the amino acid sequence SNLRR (SEQ ID NO:116), GNLVR (SEQ ID NO:117), GNLRR (SEQ ID NO:118), GNLKR (SEQ ID NO:119), ANLRR (SEQ ID NO:120), NNLRR (SEQ ID NO:121), or TNLRR (SEQ ID NO:122) at ZF4 positions +2 to +6; the amino acid sequence EHMKR (SEQ ID NO:123), EHMRR (SEQ ID NO:124), THMKR (SEQ ID NO:125), EHMNR (SEQ ID NO:126), or EHMAR (SEQ ID NO:127) at ZF6 positions +2 to +6; and the amino acid sequence DNLLT (SEQ ID NO:128), DNLLV (SEQ ID NO:129), DNLQT (SEQ ID NO:130), DNLLA (SEQ ID NO:131), DNLAT (SEQ ID NO:132), DNLQA (SEQ ID NO:133), DNLMA (SEQ ID NO:134), or DNLMT (SEQ ID NO:135) at ZF7 positions +2 to +6;

(13) the amino acid sequence GNLVR (SEQ ID NO:136), GNLRR (SEQ ID NO:137), GNLAR (SEQ ID NO:138), GNLMR (SEQ ID NO:139), ANLRR (SEQ ID NO:140), SNLRR (SEQ ID NO:141), or NNLRR (SEQ ID NO:142) at ZF4 positions +2 to +6; the amino acid sequence EHMNR (SEQ ID NO:143), EHMKR (SEQ ID NO:144), EHMRR (SEQ ID NO:145), SHMNR (SEQ ID NO:146), SHMRR (SEQ ID NO:147), THMKR (SEQ ID NO:148), or DHMNR (SEQ ID NO:149) at ZF6 positions +2 to +6; and the amino acid sequence EHLKV (SEQ ID NO:150), EHLAE (SEQ ID NO:151), STLNE (SEQ ID NO:152), DHLQV (SEQ ID NO:153), EHLNV (SEQ ID NO:154), DHLNT (SEQ ID NO:155), EHLQA (SEQ ID NO:156), or HHLMH (SEQ ID NO:157) at ZF7 positions +2 to +6; or

(14) the amino acid sequence GHLKK (SEQ ID NO:158), AHLKK (SEQ ID NO:159), TKLRL (SEQ ID NO:160), TKLKL (SEQ ID NO:161), GHLRK (SEQ ID NO:162), THLKK (SEQ ID NO:163), or AHLRK (SEQ ID NO:164) at ZF4 positions +2 to +6; the amino acid sequence TRLKE (SEQ ID NO:165) or SRLKE (SEQ ID NO:166) at ZF5 positions +2 to +6; and the amino acid sequence RADN (SEQ ID NO:167), RHDT (SEQ ID NO:168), RRDT (SEQ

ID NO:169), RPDY (SEQ ID NO:170), RTSS (SEQ ID NO:171), or RNDT (SEQ ID NO:172) at ZF6 positions -1 to +3.

In some embodiments, the engineered CTCF variants contain two or more combinations of the above-listed amino acid sequences.

In one embodiment of the present disclosure, mutations at certain positions within the consensus CBS substantially reduced binding by the wild-type CTCF zinc finger array in a bacterial two-hybrid system that was used to select for variants from randomized libraries that are capable of recognizing the mutated CBS sequence. Combining fingers together can be used to generate variant CTCF zinc finger arrays capable of recognizing CBSs harboring multiple point mutations. In some embodiments of the present disclosure, CTCF proteins harboring these zinc finger array variants are used to restore CTCF binding activity at sites bearing one or more mutations within a CBS (i.e., non-canonical CBSs). In some embodiments of the present disclosure, CTCF variants capable of recognizing alternative non-CBS sites in the genome. In some embodiments, such CTCF variants can be used to create artificial TADs and/or enhancer-promoter loops that can purposefully insulate genes and/or perturb the higher order structure of the genome and thereby alter expression of certain target genes of interest.

Diagnosis and Treatment of Diseases

The engineered CTCF variants described herein can be used for treating diseases where aberrant gene regulation due to mutant CBS is an underlying factor. The engineered CTCF variants described herein can, for example, bind to mutant CBSs that do not bind wild-type CTCFs, thereby altering or restoring gene regulation that can reverse or slow down progression of diseases. CTCF binding has been shown to regulate expression of oncogenes, such as *MYC*. Mutations accumulated in CTCF binding sites and loss of wild-type CTCF binding are associated to dysregulation of oncogenes and increased risk of carcinogenesis. Transcriptional dysregulation of *MYC* is one of the most frequent events in aggressive tumor cells and the dysregulation is a result of mutations in CTCF binding site disrupting enhancer-promoter loop. Engineered CTCF variants can bind to the mutated sites and restore normal gene expression levels, reducing risk of cancer development. In another case, Fragile X Syndrome is the result of a duplication in a repetitive region and the loss of *FMR1* expression. Duplication of a repeat region in the X chromosome disrupts a CTCF binding site, leading to the loss of an enhancer-

promoter loop driving the expression of FMR1. The engineered CTCF variants could restore the enhancer-promoter loop, leading to restoration of FMR1 expression. Human Papilloma Virus (HPV) and other integrating viruses (such as HIV) are often silenced by CTCF-mediated insulation of the viral genome from nearby enhancers. In the case of HPV18, there is a CTCF binding site in the promoter region of the viral genome. HPV18 that have mutations in the CTCF binding site are not silenced because these sequence mutations in the binding site can no longer be recognized by CTCF. Engineered CTCF variants would be able to bind to the mutated HPV integrated genomes and restore the insulating loop.

Kits

Also provided herein are kits comprising the engineered CTCF variant, and/or nucleic acids encoding an engineered CTCF variant as described herein and instructions for use.

Other Applications for the Engineered CTCF Variants

The engineered CTCF variants described herein can be used in a number of other applications, some of which are disclosed herein.

In some embodiments, the engineered CTCF variant, or nucleic acids encoding such engineered CTCF variant can be used to further elucidate the complex interactions of CTCF and other chromatin organization proteins. The structural maintenance of chromosomes is tightly regulated within cells and CTCF plays a major role. It still remains unclear how higher order structures are inherited across cell division and maintained through cell differentiation, the use of CTCF variants can help clarify that role. CTCF variants might be used to investigate how loops are formed across the genome and to modify or restore normal genomic architecture in a manner that impacts endogenous gene expression for research and therapeutic applications. They might also be used to re-establish ancestral CTCF binding sites so that we may better understand the evolutionary implications of TAD-based genome organization and epigenetic regulation of gene expression or to create alternative genomic architectures that impact endogenous gene expression for research and therapeutic applications.

EXAMPLES

Materials and Methods

The following materials and methods were used in the examples set forth below.

Construction of B2H reporter assay components

The zinc-finger bacterial expression plasmid contained the CTCF zinc finger array (or variants) fused to gal11P. The amino-terminal end of all or part of the CTCF 11-finger zinc finger array was fused to the carboxy-terminal end of gal11P with a Flag tag linker between them. The zinc finger expression plasmid contains a Kanamycin resistance gene. The second plasmid, known as the bacterial reporter plasmid, contained CTCF binding site sequence that was introduced via BsaI restriction digest followed by T4 mediated ligation of annealed oligos containing the CTCF binding site. The reporter plasmid contained bacterial lac promoter that promoted the expression of lacZ when the CTCF binding site was bound. The reporter plasmid also has a Chloramphenicol resistance gene.

Bacterial-two-hybrid (B2H) Randomized Library Construction

Complimentary oligos were synthesized by IDT with 'VNS' or 'NNS' variation introduced in the sequence by design. Oligos were annealed and ligated into the zinc finger expression plasmid (previously digested with XbaI and BamHI) using T4 ligase. Ligation reaction was purified using Qiagen Minelute column and the purified substrate was electro-transformed into electro-competent XL1blue E.coli strain. After 1 hour recover in SOC at 37°C, the transformation was inoculated into 150mL Luria broth (LB) with 50ug/mL of Kanamycin. After the culture reached a OD600 of 0.400-0.600 (about 10 hours growth at 37 °C) the culture was spun down and the library was harvested using Qiagen Maxiprep kit.

Bacterial-two-hybrid (B2H) Reporter Assay

600ng of gal11P-zinc finger expression plasmid and 600ng of reporter plasmid with CTCF binding site of interest were chemically transformed into 150uL of $\Delta\lambda$ E.coli strain with an alpha N-terminal domain of RNA polymerase (α -NTD)-Gal4 fusion. Plasmid and cell mixture was incubated on ice for 30 minutes, heat shocked at 42°C for 1 minute, recovered on ice for 2

minutes, followed by recovery in 500uL of Luria Broth for 1 hour. Post-recovery, transformation was plated on Kanamycin (50ug/mL), Chloramphenicol (12.5ug/uL) selective LB agar plates. After 14-16 hours of growth at 37°C, colonies were picked and grown overnight in 1mL of induction media (Luria broth with 50 ug/uL of Kanamycin, 12.5 ug/mL of Chloramphenicol, 10 ug/mL of ZnCl₂, and 500 ug/mL of IPTG). After 15-17 hours of growth, 25uL of the overnight culture was sub-cultured into 1mL of fresh induction media and grown for 2 hours at 37°C or until cultures were between OD₅₉₅ 0.157-0.268 as measured by spectrophotometer. 100uL of the subculture is then lysed for minimum of 15 minutes using 11ul of a 1:10 mixture of lysozyme and PopCulture soap. 15uL of the lysis mixture was then analyzed for fold activation of LacZ by previously described colorimetric ONPG assay. Binding was quantified by fold activation of LacZ. Fold activation was determined by calculating the fold increase of β -gal levels of a sample above the β -gal levels of the negative control (no zinc finger protein fused to gal11P).

Bacterial-two-hybrid (B2H) Selection Assay

Plasmids involved in the selection assay are the same as before with only one difference: The reporter plasmid is made to be a selective plasmid by swapping LacZ with BlaC, an antibiotic resistance gene for β -lactam ring class of antibiotics, such as Carbenicillin. Selections are carried out by constructing libraries of variants from a pool of oligos ligated into the zinc finger-gal11P expression plasmid. These are electro-transformed into electro-competent $\Delta\lambda$ E.coli strain containing the selective plasmid with the CTCF binding site of interest. Cells are recovered in 1mL of SOC for 1 hour at 37°C followed by induction of selective plasmid for 3 additional hours at 37°C in 4mLs of induction media (previously described). After four total hours, transformations are plated on low stringency plates (LB agar with 50 ug/mL of Kanamycin, 12.5 ug/mL of chloramphenicol, 100 ug/mL of Carbenicillin, 10 ug/mL of zinc chloride, and 200 ug/mL, IPTG and 0.45 ug/mL of Clavulanic acid). Plates are grown overnight at 37°C for 20-24 hours and then colonies are harvested off the surface with 2mL of LB. 50uL of the scrapped colonies are sub-cultured into 1mL of terrific broth (TB) with 50 ug/mL of Kanamycin, and 12.5 ug/mL of Chloramphenicol and grown 14-16 hours at 37°C. The next day, plasmid is harvested from the overnight cultures and chemically transformed into chemically competent $\Delta\lambda$ E.coli strain containing the same selective plasmid with the CTCF binding site of interest as before. The chemical transformation is performed as previously described with the

addition of 2 hour growth in induction media following a 1 hour recovery at 37°C. After a total of 3 hours of growth, cells are plated on high stringency selective gradient plates. The high stringency gradient plates contains 50 ug/mL of Kanamycin, 12.5 ug/mL of Chloramphenicol, 100 ug/mL of Carbenicillin, 10 ug/mL of ZnCl₂, 200 ug/mL of IPTG with a gradient of Clavulanic acid starting from ~1 up to 40 ug/mL in concentration. Plates were incubated 20-24 hours at 37°C. Colonies that grew on the gradient with the highest levels of Clavulanic acid were picked and grown in 1mL of TB with 50 ug/mL of Kanamycin and grown overnight in order to harvest the plasmid. The variant plasmid was then Sanger sequenced as well as analyzed for binding activity in the B2H β -gal reporter assay.

High Stringency Gradient Plates

The high stringency gradient plates contains 50 ug/mL of Kanamycin, 12.5 ug/mL of Chloramphenicol, 100 ug/mL of Carbenicillin, 10 ug/mL of ZnCl₂, 200 ug/mL of IPTG with a gradient of Clavulanic acid starting from ~1 to 40 ug/mL in concentration. To obtain a gradient of Clavulanic acid, rectangle plates are elevated using a pipette tip so as to have a ~25° slope (enough of an angle so that the thin end of the wedge is only barely covered with LB agar). 20-25 mL of LB agar with 50 ug/mL of Kanamycin, 12.5 ug/mL of Chloramphenicol, 100 ug/mL of Carbenicillin, 10 ug/mL of ZnCl₂, 200 ug/mL of IPTG and 4 ug/mL of Clavulanic acid is added to the inclined plate to form the bottom wedge. Once solidified, the plates are laid flat and 20-25 mLs of LB agar with 50 ug/mL of Kanamycin, 12.5 ug/mL of Chloramphenicol, 100 ug/mL of Carbenicillin, 10 ug/mL of ZnCl₂, 200 ug/mL of IPTG (with no Clavulanic acid) is poured on top. This creates plates with a gradient of Clavulanic acid ranging from ~1 ug/mL up to 4.0 ug/mL.

CTCF Binding Assay Using ChIP-qPCR

K562 cells were seeded 18-24 hours in advance of transfection at a density of 3×10^5 cells/mL. 3 million K562s per variant were transfected using Lonza Kit V using the provided optimized protocol and pooled in a 10cm dish. 5ug of plasmid expressing HA epitope tagged CTCF (wild-type or variant) expressed by a pCAG promoter was used for each 1 million cell reaction. 72 hours post transfection, approximately 10 million cells were crosslinked with 1% Formaldehyde at 37°C for 10 mins. Reaction was quenched with 1.2mL of 2.5M Glycine for 5 mins at 37°C. Cells were pelleted at 430g for 10 mins and sonicated on SFX250 Branson sonifier for 5.5 mins, 32% Amplitude, 1.3s off, 0.7s on. The samples were then split in half, one

precipitated overnight, rotating at 4°C with antibody specific to CTCF and the other precipitated overnight with HA specific antibody. The next day, antibody bound chromatin complexes were incubated with G-dynabeads for 2 hours at 4°C, rotating. Beads were washed three times in 1mL of ice-cold RIPA 150 Wash Buffer (0.1% SDS, 0.1% DOC, 1% Triton X-100, 1mM EDTA, 10mM Tris-HCl pH 8, 150mM NaCl), three time in 1mL of ice-cold RIPA 500 wash buffer (0.1% SDS, 0.1% DOC, 1% Triton X-100, 1mM EDTA, 10mM Tris-HCl pH 8, 500mM NaCl), three times in 1mL of ice-cold LiCl wash buffer (10mM Tris-HCl pH8, 250mM LiCl, 0.5% Triton X-100, 0.5% DOC), and once in 1mL of ice-cold 10mM Tris-HCl pH 8.5. The antibody chromatin complex was eluted from the beads in 100uL of Elution Buffer (10mM Tris-HCl pH 8, 0.1% SDS, 150mM NaCl) with 5mM DTT added fresh. Beads were incubated with elution buffer at 65 °C for 1 hour, shaking at 900rpm. Beads were pelleted by magnet and supernatant was moved to a clean tube where, after cooling to room temp, 1uL of RNase (Roche 11119915001) was added to the sample and incubated at 37 °C for 30 mins at 600rpm. 3 uL of Proteinase K [20 mg/mL] was added to samples and incubated overnight at 65°C (Lifetech #100005393). The next day, 100uL of SPRI beads with 160uL of PEG/NaCl (20% PEG, 2.5M NaCl) were added to samples, vortexed and incubated at room temp for 5 minutes before pelleting beads on a magnet. Pellet was washed twice with 80% ethanol and air dried for 5 minutes before final elution in 150uL of 10mM Tris-HCl pH 8. 3uL of recovered supernatant was mixed with 5uL of SYBR qPCR master mix and 2uL of primer mix for quantification of fragment enrichment over 1% input untreated by antibody by Real Time-qPCR.

Generation of variant binding site cell lines

Cell lines with the variant binding site introduced at the CTCF binding site ~2kb upstream of MYC TSS were generated by nucleofecting exoMYC.K562 with SpCas9-P2A-GFP, gRNA targeting the CTCF binding site, and one of 6 distinct ssODNs as HDR templates to introduce the 6 different variant binding sites. exoMYC.K562 is K562 cell line transduced with exogenous MYC construct expressed off of PGK promoter. This was necessary as any reduction of endogenous MYC expression can impact the survival of K562 cells. GFP+ cells were sorted at a high dilution into a 96 well plate for single-cell clonal expansion. Once expanded, gDNA and RNA was extracted to genotype and phenotype the clonal cell population. Clonal lines that had a reduction of endogenous MYC and also appeared homozygous at the target site for the desired HDR event were used in the study.

Quantifying MYC expression by RT-qPCR

Three million K562 cells genome edited to harbor the variant binding site upstream of MYC were nucleofected with 5ug of plasmid expressing a variant CTCF following the Lonza Kit V protocol. 72 hours post nucleofection, 1 million cells were isolated for RNA extraction following the NucleoSpin RNA Plus RNA isolation protocol. The RNA was converted to cDNA via Thermo High-Capacity RNA-to-cDNA Kit. 3uL of 1:20 dilution of cDNA was mixed with 5uL of Thermo Fast SYBRgreen Master Mix and run on RT-qPCR machine following standard PCR amplification protocol.

Results

Single Nucleotide Substitution at CBS Affecting CTCF Binding Efficiency

We reasoned we could use a bacterial two-hybrid (B2H) system to evolve the zinc finger array of CTCF to bind to mutated CBSs bearing single or multiple sequence changes that disrupt wild-type CTCF binding (Wright et al. Nature Protocols (2006); Sander et al., Nature Methods (2010); Maeder et al. Molecular Cell. (2008)). We used a previously described bacterial-two-hybrid (B2H) system to systematically define the impact of single nucleotide substitutions within a previously defined consensus CBS site (Joung et al., PNAS (2000)). In the B2H system, the binding of a DNA-binding zinc finger array to a target site of interest can be configured to result in increased transcription of a reporter gene (e.g., beta-galactosidase or an antibiotic resistance gene) (FIG. 2). To do this, two fusions are expressed in an E. coli cell bearing a reporter construct. The first fusion consists of a zinc finger array fused to a fragment of the yeast Gal11P protein, which interacts with a fragment of the yeast Gal4 fusion. The second fusion consists of a fusion of the N-terminal domain of the E. coli RNA polymerase alpha subunit to the yeast Gal4 fragment (the α -Gal4 fusion). The reporter construct consists of a weak E. coli promoter that drives expression of the reporter gene of interest with a binding site for the zinc finger array positioned upstream of the promoter. Binding of the zinc finger-Gal11P fusion to the zinc finger binding site results in recruitment of RNA polymerase complexes harboring the alpha-Gal4 fusion, resulting in increased transcription of the reporter gene. If the reporter gene is lacZ, which encodes for β -galactosidase (β -gal), the level of beta-gal expression can be easily quantified using a well-established colorimetric ONPG-based assay (FIG. 2).

In this B2H reporter assay, we determined the entire zinc finger array (ZF1-11) and the full CTCF binding site (CBS), not just the 15bp consensus CBS sequence, was required for optimal expression of the lacZ gene (FIG. 3), which mimics observed CTCF binding requirements in human cells^{10,11}. After optimizing positioning of the CBS site relative to the transcription start site, we then systematically introduce point mutations into the CBS and tested their impact on lacZ expression. These results demonstrated that mutation of nucleotides outside the 15bp core sequence had little impact on lacZ expression. By contrast, binding, however certain sequences at certain positions within the core sequence resulted in no or reduced binding (FIG. 4). Our results closely match ChIP-Seq data for CTCF binding sites in human cells and reflect other studies in the literature in which point mutations in the CTCF core lead to loss of CTCF binding. Taken together, these results strongly suggest that binding activity of the CTCF zinc finger array in the B2H system mimics the binding activity of intact CTCF protein in human cells.

Although most sequence changes in the flanking regions of the binding site had little impact on binding efficiency, certain alterations appeared to slightly improve the fold-activation of lacZ expression. Therefore, we tested whether a more “optimized” CBS bearing the “best” nucleotides as defined in the B2H assay might lead to higher-fold activation of lacZ expression but we did not observe any higher activity compared with the original consensus sequence (derived from Nakahashi et al. ChIP-seq data) (FIG. 5).

Generation of Engineered CTCF Variants That Bind to Mutated CBSs with Single Altered Nucleotide

Next, we sought to determine if we could use the B2H system to select for CTCF zinc finger array variants capable of recognizing mutated CBSs not recognized by the wild-type CTCF zinc finger array. To do this, we modified the B2H reporter construct, replacing the lacZ gene with the blaC gene (FIG. 6), which encodes beta-lactamase and therefore confers resistance to beta-lactam antibiotics (e.g., carbenicillin). This modification enables us to select for cells that express a CTCF zinc finger array variant that can efficiently bind a mutant CBS positioned upstream of the weak promoter driving blaC expression. Increasingly higher levels of blaC expression can be selected for by using media containing carbenicillin and increasingly higher concentrations of the beta-lactamase inhibitor clavulanic acid. Gradients of clavulanic acid can

be created within a single agar plate (FIG. 6; see Materials and Methods), thereby enabling sampling of cells at various concentrations of the inhibitor.

With this modified B2H selection system, we first sought to identify CTCF zinc finger array variants that can bind to CBSs bearing single point mutations that abolish binding by the wild-type CTCF zinc finger array in this system. In an initial set of selection experiments, we sought to identify CTCF zinc finger array variants that could bind to mutant CBSs bearing mutations of the C that is contacted by an aspartic acid (D) present at the third position (+3) of the alpha-helical recognition helix of ZF7 (shown by previously published co-crystal structures cited above). We created a randomized library of CTCF zinc finger array variants in which the codon encoding the ZF7 +3 position was randomized using a degenerate NNS codon (where N = G,A,C, or T and S = G or C). We then used the B2H selection system to interrogate this library to identify variants capable of recognizing CBSs bearing C to T, C to G, and C to A substitutions at the position contacted by ZF +3. Selections were initially performed on low stringency plates with clavulanic acid gradients ranging from 0 to 0.45 ug/ml) and surviving colonies harvested and plasmids encoding the variant zinc finger arrays were purified. This selected subset of variants was then subjected to high stringency selection in the B2H system on plates with carbenicillin and gradients of clavulanic acid ranging from 0 to 4 ug/ml). Plasmids encoding variant zinc finger arrays were purified from colonies that grew on the end of the gradient plate with highest concentration of clavulanic acid, sequenced, and then tested in the B2H reporter assay by beta-galactosidase assay.

As can be seen in FIGS. 7A-C, we obtained CTCF zinc finger array variants that showed preferential binding activity (as judged by the B2H reporter assay) for the mutated CBS compared with the original consensus CBS. These clones also showed selection for a particular amino acid at the ZF7 +3 position: for the C to T site, a threonine (T) was selected, for the C to A site, an asparagine (N) was selected, and for the C to G site a histidine (H) was selected. The identities of these amino acids is consistent with what might be expected to recognize the mutant nucleotide based on previous zinc finger selections using the Zif268 zinc finger array. However, although we successfully selected for mutants that had altered binding activity, in most cases, the binding activity of the variant for the mutated CBS was not as strong (as judged by the B2H reporter assay) as that of the wild-type CTCF zinc finger array for the consensus CBS (FIGS. 7A-C).

Based on our previous experience with re-engineering the DNA-binding specificities of the Zif268 zinc finger array, we hypothesized that obtaining stronger binding variants might require alteration of amino acids flanking the +3 position in ZF7. To test this idea, we created a larger library of variants in which we randomized positions +2, +3, +5 and +6 of ZF7 using degenerate VNS codons (where V = G, A, or C). Position +4 of ZF7 was not altered because it faces the internal core of the ZF domain and is not expected to make contacts to the DNA. We then performed B2H selections as described above using this library to identify variants that could identify a mutant CBS with a C to G mutation at the position contacted by ZF7 +3 in the wild-type CTCF zinc finger array. These selections identified variants that showed stronger binding activity for the mutant CBS and showed some degree of consensus in the identities of amino acids selected (FIG. 8).

Based on this success, we generated additional randomized libraries in which randomized positions -1, +1, +2, and +3 or +2, +3, +5 and +6 for ZF7, ZF6, ZF5, ZF4, and ZF3. We then performed selections as described above using these libraries against various matched mutant CBSs harboring nucleotide substitutions at positions expected to be contacted by residues randomized in the libraries (FIGS. 9-16). Analysis of variants from individual surviving colonies at the most selective end of the high stringency selection plates showed that many of these selections yielded variants with high activity for the mutant CBS of interest and sequencing of these clones showed that there was generally a degree of consensus in the amino acid sequences suggesting that selection was successfully occurring (FIGS. 9-16).

Generation of Engineered CTCF Variants That Bind to Mutated CBSs with Multiple Altered Nucleotides

Having successfully identified CTCF zinc finger variants that could recognize CBSs with a single altered nucleotide position, we next sought to identify variants that could recognize CBSs bearing multiple mutated nucleotides. To do this, we sought to recombine ZF variants each selected to bind to different “subsites” within the CBS that bear individual mutations. However, because of well-known context-dependent effects that exist between ZFs in a multi-finger array, we undertook a strategy in which we recombined together pools of selected ZF variants (rather than a single variant) for any given altered subsite to identify the combinations of mutated ZFs that best work together to recognize a CBS bearing multiple mutations. To isolate pools of ZF variants for various mutated CBS subsites, we harvested all remaining clones from the high

stringency selection plates we performed with the CBS sites bearing single mutations (depicted in FIGS. 9 – 16). Deep sequencing of the various selected clones in these pools yielded a variety of sequences with some degree of consensus within each selection as expected (Table 1).

We then recombined pools of variants for ZFs 4, 5, 6, and 7 to create CTCF zinc finger arrays that harbored various altered recognition helices for these positions and then performed B2H selections (see Materials and Methods) against five different mutated CBSs bearing combinations of various nucleotide substitutions in subsites for ZFs 4, 5, 6, and 7 (FIGS. 17-21). Sequencing of clones from these selections showed that certain recognition helix sequences for each finger were selected multiple times, suggesting that the selections were identifying combinations that work well together. Importantly, for all five of the multiply mutated CBSs, several of the CTCF zinc finger array variants identified showed good binding activity on the site for which they were selected as judged by B2H assay (FIGS. 17-21). In addition, for four of the five mutant CBS sites, we were able to identify variants that not only bind to the mutant CBS but also fail to bind to the original unmutated (consensus) CBS. Thus, we conclude that using our approach described here we are able to identify CTCF ZF array variants capable of recognizing multiply mutated CBSs that are not efficiently bound by the original wild-type CTCF zinc finger array.

Binding Specificity of Engineered CTCF Variants to Mutant and Wild-Type CBSs in Human Cells

Having successfully engineered variants that can recognize CBSs with multiple sequence changes across the motif, we next wanted to investigate whether the variants can bind to these same mutant binding sites in a human cell context while not binding to wild-type CBSs. First, we found a collection of sites in the human genome that matched the 15bp core sequence for each of the five mutated binding sites that we had selected CTCF variants to bind (described in FIG. 17-21). We then looked at two variant binding sites with sequence that matched one of the five mutated binding sites (sequence depicted in FIG. 20) as well as known CBSs to determine if endogenous CTCF could bind to the wild-type CBS and not bind to the variant binding sites as the B2H reporter assay would suggest (FIG. 20). Human K562s, an erythroleukemia cell line, were harvested and analyzed by ChIP-qPCR using CTCF specific antibody to detect CTCF-DNA binding. Wild-type CTCF showed no detectable binding to two different target sites that matched the mutated CBS but showed great enrichment for wild-type CTCF binding site, supporting the

results of the B2H reporter assay (FIG. 22). Next, we wanted to see if overexpressed, exogenous, 3xHA tagged wild-type CTCF delivered by plasmid transfection in K562s had the same binding profile observed with endogenous CTCF. Wild-type K562s were transfected with 3xHA-CTCF and 72 hours later were harvested and processed for ChIP-qPCR analysis with HA specific antibodies. Exogenous wild-type 3xHA-CTCF could bind to the wild-type CBSs and could not bind to the variant binding sites, same as endogenous wild-type CTCF, suggesting overexpression of CTCF by plasmid delivery reflects biologically relevant behavior (FIG. 23A). Based on these results, we next examined the ability of a variant CTCF to bind to the variant binding sites native to the human genome. The variant chosen was one pulled out from selection in the B2H selection assay and shown to bind to the variant site with the same sequence as variant site 1 and 2, used in FIGs. 22-23B, by the B2H reporter assay. K562s were transfected with the 3xHA-tagged CTCF variant and the same sites as before were examined for binding activity by ChIP-qPCR. Variant specific HA enrichment was present at the variant binding sites and lacking at the wild-type sites suggesting we successfully evolved a variant that can specifically bind to mutant CBS with as few as three nucleotide changes without binding native CBSs (FIG. 23B).

Gene Expression Regulation by Engineered CTCF Variants Via Looping

CTCF has the capacity to alter gene expression through CTCF-Cohesin mediated looping of the genome. We were curious to see if the variant CTCFs could reproduce the gene regulatory capacity of wild-type CTCF when bound to the endogenous variant binding sites. To investigate gene expression changes, we focused on genes within a 1Mb region of the variant binding sites. Eleven genes were identified within 1Mb region for Variant site 1.1 and 1.2 and another 10 genes were identified for Variant site 2.1 and 2.2. K562s were nucleofected with variant CTCFs fused to GFP that had the capacity to bind to Variant site 1 and Variant site 2. 72 hours post nucleofection, RNA was isolated from GFP+ cells and gene expression levels were compared to RNA extracted from K562s nucleofected with a wild-type CTCF control. Of the 11 genes for Variant site 1.1 and 1.2, 6 genes showed a change in gene expression relative to cells nucleofected with the wild-type CTCF control (JJ388) (FIG. 24A). 2 of the 10 genes identified for Variant site 2.1 and 2.2 had altered gene expression levels relative to wild-type control (FIG. 24B). This data suggests that not only do the variant CTCF proteins bind to their target sequence

in human cells, but it also reproduces the biological role of native CTCF to regulate gene expression possibly through the formation of loops or sub-TADs.

Next we wanted to demonstrate that the CTCF variants could replicate the biological function of wild-type CTCF at a known CTCF binding site that creates an enhancer-promoter loop. MYC expression is maintained by a loop formed between a CTCF binding site ~2kb upstream of the transcriptional start site (TSS) of MYC and a CTCF binding site ~1kb downstream of the MYC TSS¹⁴. When CTCF is bound to both sites, cohesin links both CTCFs via the CTCF's cohesin-interaction domain, creating a loop that maintains the expression of MYC. If one or both of the CTCF binding sites is disrupted the CTCF-mediated loop is lost and there is a reduction in MYC expression¹⁴. Five cell lines were generated containing the 5 different variant binding site sequences (defined in FIG. 25) at the CTCF binding site ~2kb upstream of the MYC TSS. This was done in K562 background transduced with a lentiviral construct expressing exogenous MYC via phosphoglycerate kinase (PGK) promoter (exoMYC.K562) to compensate for any reduced cell fitness that reduction of endogenous MYC expression may cause. An additional sixth cell line was generated where point mutations to the CTCF binding site were made that should have no affect on wild-type CTCF binding as indicated by results from the B2H reporter assay. RNA was isolated from the clonal cell lines homozygous for the variant binding sites and endogenous MYC gene expression levels were assayed by reverse transcriptase Real Time qPCR (RT-qPCR). Each of the isolated cell lines with the variant CTCF binding site demonstrated a reduced level of MYC expression suggesting that the CTCF-mediated loop is disrupted (FIG. 25).

Based on this result, we wanted to see if expression of the variant CTCFs in these modified cell lines could bind to the engineered sites and restore MYC expression. HA tagged wild-type CTCF and HA tagged CTCF variants were expressed in the cell line that contained their matching variant binding site. Variants selected to bind to the G3 variant binding site were expressed in the G3_3 cell line, A3 variants in the A3_4 cell line, etc. HA-tagged wild-type CTCF was also tested in each of the variant cell lines for binding and for recovery of endogenous MYC expression. The level of endogenous MYC expression in exoMYC.K562 served as wild-type control as there is no alteration to the CTCF binding site upstream of the MYC TSS. CTCF variants expressed in the engineered cell lines recovered endogenous MYC expression while expression of wild-type CTCF in these cell lines failed to recover MYC expression (FIGs. 26A-

29). The same samples were analyzed for occupancy of the variant binding sites by wild-type CTCF or the variant CTCFs by ChIP-qPCR enriching for CTCF-bound DNA fragments with CTCF or HA antibody. Wild-type CTCF had a reduced occupancy of the variant binding sites, consistent with continued reduction of MYC expression, while variant CTCF proteins could bind to the variant site they were selected for as well as rescue MYC expression (Figure 26-29). Together, this data suggests that we have evolved CTCF variants that can bind to novel sequences and still interact with cohesin to form loops that maintain gene expression profiles.

TABLES

Amino acid sequence of variants selected for on different CTCF binding sites. All amino acids sequences are listed from N to C terminal. Colonies growing on the highest stringency of selection were scrapped off, pooled, and plasmid encoding for the zinc finger was isolated and deep sequenced. The number of reads reflects how prominent the variant was in the population pooled from selections performed in triplicate.

Table 1. ZF7
selection on C:G
change at nt 2 of
core motif in CBS.
Sequences reflect
position 2 through
6.

Sequence	# reads				
DHLQT	2981	HTLME	245	DHLKA	68
EHLVV	2413	DHLRA	237	KDLTQ	66
DHLNT	1517	DHLAV	221	DKLMN	66
DHLRT	1442	HHLAE	221	GTLRE	66
EHLKV	1434	GHLMD	207	GHLTV	66
KDLVV	1357	DHLST	199	RLLTA	65
DHLQA	1114	EHLMV	197	SSLRE	63
DHLLV	1076	AHLVV	196	HTLKE	62
DHLLT	881	EHLAV	192	GHLAV	60
EHLTV	803	HTLAE	187	RLLAQ	58
STLME	786	STLQE	181	KDLAV	57
DHLAT	777	DHLAE	167	EHLQE	57
EHLNV	736	AHLQE	163	SHLNV	57
DHLQV	574	SSLNE	158	AGLPI	57
DHLKT	541	GHLNV	155	TTLME	56
EHLKE	517	EHLVE	144	AHLRV	56
DHLLT	506	DHLME	143	AHLMV	55
EHLRV	503	DHLRE	134	EHLME	55
STLRE	498	AHLNA	120	EHLQT	55
DHLMV	431	HTLVE	120	EVLNR	55
DHLKV	427	STLKE	112	HHLVV	54
DHLRV	394	EHLQV	107	KDLSV	54
DHLNV	389	GTLME	106	RHLVM	53
DHLLA	380	HHLAV	102	THLNE	50
DHLKE	368	HSLME	101	RDLRT	49
DHLNE	330	HSLTE	97	LLLGS	49
EHLRE	330	EHLMA	97	MVLGN	48
STLLE	323	DHLHT	94	KTLIE	47
DHLMA	305	AHLQV	94	AHLGV	46
KDLTV	296	DHLTV	93	SGLLA	46
DHLVT	284	EHLIV	90	DHLHV	45
AHLNV	270	SGLNE	89	EHLNT	45
AHLTV	268	AHLLV	85	STLLQ	44
		EHLLV	84	AHLKV	44
		VKLKI	83	AHLAV	42
		DHLQE	80	TNLID	41
		HTLTE	77	GTLNE	41
		STLHE	76	QVLTQ	40
		DHLVV	76	SSLME	39
		AGLAL	70	GHLVE	38
		STLND	69	HSLLE	38

SGLLE	38	DTLLA	17	DHLTI	9
GGLLE	36	STLDE	16	THLIV	9
STLRV	36	KDLVA	15	DTLMA	9
HTLAD	35	AHLHA	15	MKLQE	9
SHLME	35	KDLQV	15	TALGT	9
DHLAI	35	HHLTV	15	GHLLV	9
EHLA	35	SGLLD	15	GQLAI	8
HNLLL	34	ANLME	14	ANLES	8
PHLVV	34	DNLLV	14	AHLNT	8
KALGT	33	EHLKT	13	EHLLE	8
PHLVI	31	GSLAI	13	SNLTV	8
VLLII	30	EHLVS	13	STLLV	8
HHLRE	29	EHLNE	13	STLMV	8
GALRM	29	EHLVI	13	GTLVS	7
RGLHE	29	KDLKV	13	DNLKT	7
AHLLE	28	EGLGT	13	GHLQT	7
EHLKA	28	DNLQT	12	DNLLT	7
DTLLV	27	STLMS	12	EHLVT	7
EHLRT	26	AHLM	12	GALRE	7
SSLRD	24	IKLDG	12	SSLAE	7
EHLQA	23	VLLGA	12	DTLRQ	7
EHLAT	23	PGLSA	12	KALLG	7
SGLGE	22	AELNR	12	AMLNP	6
ATLQE	22	HQLVI	12	DTLHQ	6
DHLA	22	GHVV	12	DNLLQ	6
SNLLV	22	PHLLV	11	EHLAH	6
SHLLV	21	PRLAL	11	AHLKE	6
KDLMV	21	DHLNA	11	ATLAE	6
DHLQQ	20	KDLDV	11	EHLMD	6
ATLME	20	AHLHV	11	STLHM	6
GHQA	20	RVLGG	11	DTLAV	6
RTLTE	20	AHLQA	11	DHLVE	6
RRLAH	20	RQLRT	10	PTLGE	6
DTLQA	20	AHLQT	10	KGLPL	6
GHLEV	19	DNLLA	10	DTLLQ	6
HQLKL	19	EHLAE	10	AHLNE	6
EHLT	19	EHLAM	10	AHLAE	6
DGLRT	18	DRLSI	10	GHKLV	6
THLRP	18	GGLGA	10	SGLQV	5
DNLAT	18	GHNT	10	HHLV	5
EHLNA	17	AHLRT	10	EPLLP	5
STLVV	17	DTLRV	10	DNLAV	5
DNLMT	17	MSLRG	9	AHLLT	5

AHLST	5
DNLQA	5
DNLRT	5
DTLAL	5
DTLQV	5
EHLRA	5
SNLQV	5
KDLRV	5
DTLAT	5
DTLRA	5
QHLRV	4
SSLLE	4
SNLMV	4
SDLGG	4
DNLHT	4
DNLTA	4
DTLMV	4
EHLST	4
DTLSV	4
DNLMA	4
EHLVM	4
STLAE	4
KDLAE	4
SSLNV	4
SSLLV	4
AHLKT	4
AHLRE	4
KDLLV	4

Table 2. ZF7
selection on C:T
change at nt 2 of
core motif in CBS.
Sequences reflect
position 2 through
6.

Sequence	Read #
DTLLV	3772
DTLLA	1720
DTLLQ	1681
DTLQA	1340
DTLRV	1048
DTLQV	715
DTLRA	643
DTLMA	620
DTLMV	538
DTLAV	451
DTLAT	406
DSLLV	373
DTLSV	359
DTLLM	339
DTLRQ	334
DTLLI	306
DTLTQ	300
STLAE	269
DTLAA	268
DTLHQ	246
DTLSA	227
DTLKA	216
STLLV	213
STLQQ	201
DTLQQ	200
DTLLL	194
DTLMQ	189
STLQE	189
DTLNA	180
STLLA	176
DTLKV	163
STLNA	162
DTLRE	161

DTLTA	152	DTLGV	71
DTLQD	146	DVLRE	67
DTLVA	137	STLSA	66
DTLLS	123	DSLVS	65
STLTQ	122	DTLLE	63
DSLLA	116	STLAA	63
DTLRT	116	DTLKI	62
DTLQI	115	DTLKM	62
DTLMN	114	DTLQN	60
STLSE	114	STLME	60
SSLQV	112	TTLMT	60
TNLAV	109	TTLAE	59
DTLVV	108	STLTE	58
DTLHA	107	VELVQ	57
DTLMT	107	TTLNQ	56
SSLLV	107	DTLMI	54
STLLE	107	TTLMD	54
DSL RV	106	STLMA	51
DTLAE	105	DVLLA	50
STLNV	105	DVLLT	49
DTLRN	101	STLKE	49
DTLNV	100	TTLNE	49
DTLRD	99	MTLPT	48
DSLAV	94	SSLME	48
DTLVQ	94	STLHE	48
DTLQE	93	HTLVV	47
STLLD	92	TTLME	46
DTLTH	89	ATLTQ	45
SSLND	88	STLAS	45
STLTV	88	STLVV	44
STLMV	87	SSLLE	43
DTLML	86	SSLVE	42
STLLQ	85	DALQA	41
STLRE	85	DVLDA	41
STLQA	84	GSLMQ	41
DTLLD	83	DTLTM	40
DTLKQ	82	STLAQ	39
DTLLT	81	STLMI	38
DTLAL	76	DTLAM	37
DTLII	75	DTLHT	37
DTLLN	75	DTLQL	37
DSLLQ	73	DSLKQ	36
STLEQ	73	DSLRA	36

STLHV	35	DTLIA	18
STLMQ	35	DTLTE	18
DGLMA	34	GTLNH	17
DTLRL	34	STLAI	17
SSLLT	34	AHLGV	16
DSLQA	33	DNLLV	16
DTLRI	33	DQLVQ	16
STLGE	33	MPLIL	16
DALKE	32	TTLHQ	16
STLRA	31	TTLQV	16
DTLHH	30	ATLLE	15
DTLRG	30	DVLHE	15
DTLRM	30	ETLRA	15
DVLMT	30	KVLRS	15
DTLEI	29	SNLLV	15
SSLNE	29	DNLMT	14
DTLHV	28	DSLQ	14
GTLDE	28	DTLAN	14
SSLAV	28	GTLNV	14
STLKQ	28	HNLNV	14
DTLMD	27	QTLQA	14
GTLQT	27	RQLTT	14
SSLVQ	27	DTLSI	13
STLRV	27	DRLVG	12
LMLMG	25	ETLRQ	12
STLRQ	25	SSLGE	12
STLTA	25	SSLVV	12
DHLQT	24	DHLQA	11
DSLVA	23	DNLLT	11
SSLRV	23	DTLME	11
DSLRE	22	DTLTV	11
GRLQD	22	DTLVG	11
MALQD	22	ETLKA	11
STLLH	21	GVLSQ	11
STLVQ	21	LALMR	11
VRLTA	21	RTLVE	11
AVLGD	20	TTLII	11
PILVT	20	TTLNV	11
STLDD	20	DTLSE	10
DSLMI	19	SSLAE	10
STLID	19	STLAV	10
TKLDT	19		
ATLVA	18		

Table 3. ZF7
selection on C:A
change at nt 2 of
core motif in CBS.
Sequences reflect
position 2 through
6.

Sequence	# read
DNLLA	2659
SNLLV	2616
DNLMT	2555
DNLQT	1983
DNLLV	1945
DNLLT	1922
DNLAT	1457
DNLRA	1117
DNLMA	1038
DNLMV	901
DNLQV	845
DNLQQ	841
DNLLQ	813
DNLKT	582
DNLQA	571
SNLQV	565
DNLRQ	494
SNLMV	459
SNLTV	458
DNLNT	412
DNLHT	389
SNLVV	349
SNLQQ	334
DNLTA	323
DNLLS	322
DNLAV	316
DNLRT	309
DNLTT	300
DNLAA	295
SNLLA	295
SNLLQ	278
SNLAV	257

DNLNA	240	DNLIV	28
DNLGT	240	SNLMT	27
DNLRV	239	DTLMN	25
DNLKA	167	SNLTQ	23
DNLMQ	156	EILRI	23
DNLKV	148	IGLEA	22
SNLNV	132	HRLGG	22
SNLMA	128	DHLQT	21
SVLQD	113	DNLST	20
DNLQS	110	MRLHV	19
DNLSA	105	SNLTT	18
DNLAQ	103	SNLGV	16
DNLMS	98	SNLAT	16
DNLSQ	95	EHLVV	16
DNLNV	87	ANLMV	14
DNLGV	87	HVLVG	14
SNLLT	87	SNLRA	13
DNLIA	83	HNLQL	12
DNLNQ	83	DNLVA	12
SNLQT	80	SNLTA	12
SNLRV	79	KGLRM	12
SNLIV	79	DTLLA	12
DNLSV	74	PMLGV	11
SNLQA	60	GVLVA	11
SNLLL	57	DNLQD	11
SNLDV	56	MKLGT	11
DNLVQ	54	DTLLQ	11
SNLLI	54		
TGLAL	52		
SNLMQ	51		
DQLKI	40		
GDLGT	40		
SNLKV	39		
VPLVD	38		
DNLRI	37		
DNLLI	37		
TNLDV	36		
HDLKI	35		
DNLVV	35		
DTLLV	32		
DNLTV	31		
DNLVT	31		
SNLAQ	30		

10

Table 4. ZF7
selection on A:T
change at nt 3 of
core motif in CBS.
Sequences reflect
position -1 to 3.

Sequence	# Reads
RKHD	4641
RKAD	1938
RRSD	1299
RRHD	868
RKTD	182
NVSM	146
RQSD	76
RKND	69
SENV	69
VDHR	60
AQIV	58
KTPH	56
PKIV	51
GAEP	42
MLVE	40
VVGN	40
KGPE	36
GKVM	33
TEPG	33
TPHN	32
MPGG	31
DLEK	28
GTDN	27
ISRL	25
ATGL	21
ASNP	19
GAPT	17
HSPN	17
RPVA	16
RKDD	6
MLVD	4
RHRK	3
RKHV	3

RKQD	3	RKYD	1
RKSD	3	RMSD	1
DHHT	2	RRLD	1
GKHD	2	RRND	1
MKAD	2	RRRD	1
RKAE	2	RRSG	1
RRAD	2	RWHD	1
APIG	1	SHRL	1
AQNR	1	SQHV	1
DMDA	1	SSHD	1
EAPM	1	TTHV	1
EEMM	1	VHHV	1
EPIR	1	WKAD	1
GALE	1	WKHD	1
GENV	1		
GKAD	1		
GKVD	1		
GPLA	1		
GRIE	1		
IEKL	1		
KAAS	1		
KEEH	1		
LKVD	1		
LLVE	1		
LMTQ	1		
MASL	1		
MGIG	1		
MPGD	1		
MSLG	1		
NDMT	1		
NMHT	1		
NRIV	1		
PENA	1		
QKHD	1		
QVPD	1		
RASD	1		
REHD	1		
RGHD	1		
RKHA	1		
RKHY	1		
RKLD	1		
RKPD	1		
RKVD	1		

49

Table 6. ZF7
selection on A:C
change at nt 3 of
core motif in CBS.
Sequences reflect
position -1 to 3.

Sequence	Read #
RKHD	9
DTEN	6
GEHG	5
STKN	5
NIEI	5
QLPL	4
GREV	4
DHHT	4
DRDM	4
MVIN	4
VPDT	4
NIVP	4
MVPS	4
PNHP	4
KTDV	4
NEPN	3
SHRL	3
MGIG	3
HIKM	3
ILQI	3
NRIV	3
IVMQ	3
QTNS	3
ENMD	3
TVER	3
THDR	3
IRSP	3
MLPA	3
EEMM	2
ARIA	2
VPTV	2
EELI	2
KPLR	2
DDAT	2

NRLS	2	DRSM	1
PTLR	2	KNSI	1
MHIL	2	DDVS	1
GGGP	2	KPQP	1
MVEN	2	PHVP	1
DMDA	2	DTLQ	1
IVAT	2	KLGT	1
TLDR	2	IDPH	1
MEPL	2	HPNT	1
DTGV	2	KSRG	1
TSRS	2	RQMA	1
VLSI	2	KKEN	1
STVQ	2	QVLD	1
GPAQ	2	EPIR	1
VEQP	2	RRQM	1
MTKK	2	TPLP	1
PLIM	2	ILKN	1
QGTS	2	HQMK	1
AMTV	2	ELLN	1
SPMR	2	MDGG	1
EPNV	2	AAGS	1
MASL	2	STVV	1
MQIN	2	PARA	1
ALDE	2	ALQG	1
GRIE	2	SAPG	1
ALEH	2	PVLN	1
REKD	2	PENA	1
ELLA	2	TSLL	1
GVAR	2	KEEH	1
VDTL	2	HLDV	1
GHEN	2	IHIR	1
KAAS	2	SVTL	1
ELES	2	VKDR	1
DPDT	2	KMTI	1
SLEL	2	AGEM	1
TMNV	2	GDSE	1
VHHV	2	QPVK	1
IQPV	2	KVEA	1
MLQE	1	EQER	1
VMTV	1	IEKL	1
MVEE	1	GHHV	1
VARP...	1	GMHL	1
KAIG	1	RLRR	1

ATIR	1	ILKQ	1
RMDI	1	PVIP	1
SVIH	1	VESL	1
MDIG	1	IKQN	1
LART	1	EDNI	1
RLMA	1	THRD	1
RQPP	1	IPAG	1
MTMT	1	GLNH	1
EDTR	1	VDGR	1
NDMT	1	PHRM	1
MRGR	1	RTGA	1
ELHA	1	VSPD	1
TNGQ	1	KVGD	1
VNLT	1		
MHIR	1		
MLLQ	1		
GRGE	1		
NLRG	1		
HIML	1		
GRAH	1		
MNTL	1		
TTHV	1		
KTMG	1		
MTSV	1		
RLSM	1		
DTMV	1		
EAPM	1		
DMGM	1		
MLMM	1		
LMEM	1		
QAVS	1		
SRVL	1		
DEDP	1		
SGDR	1		
MMNC	1		
NIGM	1		
MVQR	1		
APHR	1		
LDAG	1		
RLAN	1		
MKGS	1		
KKLV	1		
VNQE	1		

Table 7. ZF6
selection on C:T
change at nt 5 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence	Read #
GHMRR	29
GHMNR	23
EHMRR	23
THMRR	19
THMKR	17
EHMNR	17
GHMKR	12
EHMAR	11
EHMQR	10
SHMRR	10
SAMRR	9
ENMGR	8
SHMKR	8
THMNR	7
NHMRR	7
EGMRR	7
GNMGR	7
SHMNR	6
NGMRI	6
EGMAR	6
ESMRR	6
GHMSR	5
EGMHR	5
TAMRR	5
TNMQR	5
VNMRR	5
AHMKR	4
NGMTA	4
DGMRR	4
GHMTR	4
EHMSR	4
EHMKR	4
GSMRR	4
TNMLR	4

NHMKR	4	NNMVR	2
ENMLR	4	TGMKR	2
SPMGV	3	TAMKR	2
TNMGR	3	AHMNR	2
SSMAR	3	QNMGR	2
GGMRR	3	TNMVR	2
GGMKL	3	NHMNR	2
SGMVR	3	EHMTR	2
EHMHR	3	GNMIR	2
THMSR	3	SGMRR	2
GSMKI	3	NHMSR	2
EKMKE	3	GGMRL	2
NGMAR	3	SPMKV	2
QNMVR	3	TNMRR	2
DNMRR	3	GNMRE	2
ENMER	3	ENMMR	2
NSMRR	3	THMER	1
SGMKR	3	QKMRT	1
ANMQR	3	GAMRR	1
GHMQR	3	TPMEV	1
ANMGR	3	GGMRE	1
DNMVR	3	GDMDR	1
QAMRE	2	GAMRA	1
GNMSR	2	PNMSR	1
ESMQR	2	EGMGR	1
TPMKV	2	EGTHR	1
SNMGR	2	QSMRE	1
GAMRI	2	THMKG	1
ANMNR	2	NNMGR	1
DNMMR	2	GHMNS	1
GSMKM	2	IDMKG	1
EHMGR	2	ESMTR	1
GNMAQ	2	SHMKI	1
EGMKG	2	HNMMR	1
SSMKI	2	SNMVR	1
TSMRR	2	TAMKV	1
DGMKR	2	DSMKR	1
DNMAR	2	SNMAR	1
SSMRR	2	ESMGR	1
GNMMR	2	EAMRR	1
NAMRG	2	GNMVR	1
THMKL	2	ANMRR	1
ENMAR	2	DGMKI	1

SHMHR	1
GAMKE	1
ESMRE	1
GSMLR	1
THMEV	1
TSMGR	1
EAMSK	1
NAMRQ	1
EGMRT	1
SHMQR	1
NGMKR	1
ESMKE	1
ANMHR	1
DHTKR	1
NGMRE	1
GSMRA	1
EGMNQ	1
GGMRM	1
PNMKR	1
NGMKI	1
SNMLR	1
SNMRR	1
SHMTR	1
TGMRR	1
SGMRI	1
DNMGR	1
EGMTR	1

Table 8. ZF6
selection on C:A
change at nt 5 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence	Read #
GNMGR	873
ESMRR	784
EGMRR	772
EGMAR	672
EGMHR	648
QNMVR	597
TNMLR	556
SGMKR	486
NGMTA	479
GSMRR	453
QAMRE	452
ENMGR	434
SAMRR	431
NGMAR	401
GSMRA	390
ENMER	389
SNMGR	378
EGMGR	376
DNMAR	368
DGMKI	347
ANMQR	342
GAMRR	322
VNMRR	297
NSMRR	295
ESMQR	293
SSMRR	289
TGMRR	289
ANMNR	279
GAMRA	275
EGMTR	273
EHMNR	265
GNMSR	263
TAMRR	260
TNMQR	257

DNMMR	253	GGMRL	124
DGMRR	241	QNMER	124
QNMGR	240	DSMKR	122
PNMKR	228	TNMRR	122
DNMGR	226	GAMRI	115
ETMRR	225	SNMAR	115
DNMKI	224	TAMKV	114
SSMKI	224	VSMKR	113
DNMRR	221	NGMRI	112
ESMTR	214	TNMMR	110
GGMRE	214	EGMKG	109
TNMGR	211	NAMRQ	108
SGMRR	204	EHMKR	107
GGMRR	203	NHMSR	106
NNMVR	201	GAMRM	102
TNMER	197	EAMSK	100
NGMKI	195	TAMNQ	99
ENMAR	194	ESMSR	96
ESMGR	194	GGMNQ	95
GNMMR	193	QSMRE	95
GNMRE	193	NAMRG	92
DNMVR	187	GGMKR	89
TNMAR	186	SNMVR	84
EHMRR	182	ESMRL	83
ESMRE	181	ESMKE	81
TNMVR	181	SAMRE	80
TSMRR	175	GGMQM	76
SGMVR	173	TGMKR	73
TAMKR	170	ENMMR	69
GNMIR	169	NSMKM	69
SSMAR	163	ESMKN	66
GSMKI	159	EGMRT	64
EHMKQ	149	GGMKL	64
ENMLR	149	TNMSR	63
DGMKR	144	DAMRV	61
GNMAQ	139	GNMER	60
SGMQR	138	GAMRE	59
SNMLR	133	GNMAR	54
GNMVR	130	EGMRK	53
ANMGR	129	GSMKM	50
HNMR	129	SGMAR	50
NNMGR	128		
GGMRM	127		

**Table 9. ZF6
selection on C:G
change at nt 5 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	# Read		
EHMRR	3207	DHMRR	105
GHMRR	2397	THMSR	97
THMRR	2025	AHMRR	92
GHMNR	1880	EKMRE	86
THMKR	1415	GHMAR	84
THMNR	1341	NGMKR	81
GHMKR	1208	VHMNR	77
EHMSR	1038	IDMKG	72
EHMAR	927	NHMTR	65
SHMKR	771	NHMSR	64
EHMQR	764	EGMRR	57
EHMNR	676	THMTR	57
SHMNR	646	GHMKI	56
SHMRR	579	SHMQR	52
EHMKR	511	EHMVR	43
EHMTR	460	SHMTR	43
NHMRR	436	TKMKE	42
EKMKE	381	EHMER	38
DHMNR	374	THMKT	37
NHMKR	342	GDMDR	36
SPMGV	322	NGMRR	35
GHMTR	318	EPMLM	34
NHMNR	285	GHMVR	31
DHMKR	264	THMRT	29
GHMSR	258	ESMRR	28
AHMNR	247	PHMKR	26
EHMHR	232	EHMRQ	24
AHMKR	227	EHMRT	23
EHMGR	210	DHMSR	22
GHMHR	129	QKMRT	22
THMKV	129	ETMMI	21
THMKL	117	SPMKV	21
TPMKV	110	SHMKL	21
GHMQR	105	TPMKL	21
		GHMKM	20
		GNMGR	19
		RQMLI	19
		GHMRM	18
		EGMKR	17
		EHMKA	17
		QIMPL	17
		SHMKV	16
		SGMNR	16
		THMAR	16
		QGMKR	15
		SAMRR	14
		TKMEG	14
		RPMGR	14
		VHMRR	13
		THMRV	13
		THMEV	12
		NHMKs	11
		THMKG	11
		AAMST	11
		TNMLR	11
		ENMER	10
		GKMRD	10
		THMEL	10
		SGMKR	10
		TPMRV	10
		SPMRV	10
		GGMKR	10
		EGMAR	10
		THMGV	9
		TAMRR	9
		DNMRR	9
		NGMRI	9
		ENMGR	9
		MGMGR	8
		VNMRR	8
		GKPSM	8
		NGMTA	8
		SHMRV	8
		SPMNR	8
		SAMNR	8
		SHMSR	8
		NGMPR	8
		TNMQR	8
		SPMRR	8
		QNMVR	8
		EGMHR	8
		DNMAR	7
		QNMGR	7
		GHMGV	7
		THMRL	7
		GSMRR	7

QHMKR	7	GNMSR	4	DHMQR	3
THMGR	7	GMMLT	4	TAMKV	3
DGMRR	7	QAMRE	4	AHMSR	3
THMQR	6	NSMRR	4	EHMRS	3
THMER	6	TGMRR	4	AHMTR	3
ENMAR	6	TPMKG	4	GHINR	3
RHMKR	6	TPMEV	4	QSMRE	3
SSMRR	6	THMHR	4	TNMAR	3
EHMRV	6	SHMGV	4	EYMRR	3
KHMKR	6	DGMKI	4	GQMNR	3
QHMINR	6	DGMKR	4	GHMKE	3
RAMKV	6	THMKS	4	GSMKM	3
NGMAR	6	THMRG	4	SHMHR	3
TNMGR	6	GHMKT	4	ESMGR	3
DNMVR	6	TSMRR	4	ESMQR	3
ESMRE	6	GNMMR	4	GHMNS	3
GHMRV	6	ANMQR	4	SNMAR	3
ENMLR	6	GSMRA	4	GNMAQ	3
NAMRG	5	GNMRE	4	TGMKR	3
SSMKI	5	NGMKI	4	SGMRR	3
TPMGV	5	GAMRI	4	ANMGR	3
GAMRR	5	NNMGR	4	GGMKL	3
GHMKV	5	THMRS	4	EHMMR	2
RHMNR	5	EGMKG	4	SHMRL	2
ANMNR	5	NHMQR	4	EGMRT	2
TPMEL	5	SNMGR	4	GGMQM	2
NNMVR	5	SHMAR	3	GGMKA	2
SGMVR	5	GNMVR	3	EAMSK	2
SPMKL	5	EAMKR	3	EHMPR	2
SPMKR	5	GSMRE	3	AHMKs	2
TNMRR	5	SPMEL	3	AHMQR	2
PNMKR	5	AHMAR	3	GHTRR	2
LAMEE	5	DSMKR	3	GHMKG	2
GAMRA	5	PPMMV	3	EPMKV	2
GAMRM	5	TNMVR	3	EHMAK	2
EGMGR	5	HNMR	3	GYMNR	2
GGMRL	5	KHMNR	3	THMSS	2
PGMMS	5	GNMIR	3	GDMNR	2
GGMRR	5	SNMLR	3	GHMRT	2
GSMKI	5	TPMKR	3	EHMKQ	2
DNMGR	5	QSMKR	3	QRMGV	2
EGMTR	4	RHMRR	3	GSMRQ	2
SHMEV	4	ESMKE	3	DHMTR	2

VEMER	2	TPMNV	1	PHMMS	1
SPMEV	2	EHVQR	1	SHVQR	1
GPMKV	2	ENMKE	1	ESMSR	1
TPMER	2	DHMKM	1	SHMGL	1
EHMDR	2	SHMNQ	1	TDMVA	1
EHVRR	2	NSMKM	1	PQMMS	1
DNMKI	2	GLMKR	1	KHMQR	1
GGMAR	2	APMNL	1	EHMQL	1
HHMKR	2	RHMSR	1	EHISR	1
GHMRS	2	EHMRG	1	SHMKK	1
EYMAR	2	DWMRR	1	EQMTR	1
KHMRR	2	GHRMH	1	TPMRG	1
EHMSS	2	QNMHR	1	GHISR	1
TPMRL	2	CHMRR	1	GPMGV	1
GHMSL	2	ERMRR	1	GYMRR	1
VHMKR	2	EHMKE	1	GHMTV	1
GHTNR	2	EPMKR	1	APMIM	1
GPMRT	2	AHINR	1	THINR	1
GGMRM	2	SHMRT	1	DHMMS	1
TNMER	2	PHMNR	1	GHMKL	1
ESMKN	2	AHMKV	1	EKMEE	1
EQMRR	2	THMGM	1	DPMRM	1
ESMTR	2	NGMKM	1	SHMKT	1
EHMKS	2	EKMKR	1	SPMGL	1
THMKM	2	EHMIR	1	SPMGE	1
GAMKE	2	NNMHR	1	DHISR	1
TAMKR	2	GNMNR	1	TPMKQ	1
DNMMR	2	KRMQR	1	GHMKW	1
SSMAR	2	EKMRR	1	EHMCR	1
ENMMR	2	TQMKQ	1	NNMKR	1
GTMKM	1	EHMKV	1	ESMKR	1
VHRIR	1	DHMKE	1	TEMLI	1
DHMNK	1	EHTTR	1	SHMKM	1
TPMNM	1	SPMRM	1	EHVNR	1
RQMII	1	GKMNR	1	GHMER	1
EHMRW	1	TNMKR	1	NHMDR	1
SPMRL	1	THKRR	1	GHMWR	1
GVMRA	1	SQTNR	1	THMKI	1
GHMQV	1	THLKR	1	QKMKE	1
GPMKL	1	SHMQS	1	THMNK	1
IDMKR	1	THMSV	1	AHMKQ	1
PGMMG	1	THMRH	1	DHMGR	1
KHMER	1	DPMKV	1	EGMKW	1

TQMKE	1	THRRR	1	GHMTM	1
TRMRR	1	DAMRV	1	EPMSG	1
AHMGR	1	ERMNR	1	NHMNM	1
TRMKR	1	NHMHR	1	GQMKR	1
KNLTR	1	DLMNR	1	TPMEG	1
PEMMS	1	GQMQR	1	KHMRV	1
EHLTL	1	RGMMI	1	SLMKR	1
RHMKV	1	TQMKR	1	DGMRN	1
PGMIR	1	EHMGV	1	RQMHI	1
THTKR	1	AHMTQ	1	EPMRV	1
EHIRR	1	TPMMV	1	GAMRE	1
THMPR	1	GHKRR	1	SHMRM	1
GKMKQ	1	GPMER	1	EQMAR	1
GPMRV	1	EPMQV	1	SHMRS	1
AHVNR	1	TAMNQ	1	EHMQV	1
EPMSR	1	GDMRR	1	EPMPM	1
PRMMV	1	EHLKR	1	IDMNR	1
ELMSR	1	DHMKK	1	TKMKQ	1
ETMRR	1	GDIDR	1	RQMLS	1
SNMNR	1	GHMKK	1	ATMML	1
TSMKT	1	TQMMI	1	PQMMI	1
GNMHR	1	SGMKA	1	NAMKI	1
TQMRR	1	TPMRM	1	GHMQS	1
SHMKG	1	SPMKG	1	EAMKK	1
DHMRT	1	KQLNR	1	THMRK	1
EHMRE	1	NHMKT	1	PHMRR	1
SQLNR	1	TKMRE	1	GHMKA	1
SHMGR	1	VSMKR	1	AHMNH	1
GHKNR	1	EHMAV	1	EYMSR	1
THMNL	1	EHMNS	1	EHMAW	1
GYMKR	1	DHMHR	1	NHMGR	1
SNMKV	1	AHMVR	1	GHMKS	1
GHMRC	1	GRMRR	1	EHMRL	1
NHMRV	1	GHMNV	1	ENMTR	1
SGMKT	1	GHMNL	1	TNMMR	1
EHLRR	1	GHVSR	1	QAMRV	1
VPMRR	1	GQMHR	1	EHMQP	1
DLMKR	1	EKMAR	1	THMSM	1
TSMKL	1	NHMGL	1	IDMKE	1
APMTV	1	EHMKG	1	EGTHR	1
ESMRL	1	EPMAL	1	HNMMR	1
EHMLM	1	AHLTR	1	PNMSR	1
EKMNR	1	KHMTR	1	SNMVR	1

ANMRR	1
GGMRE	1
EAMRR	1
GSMLR	1
SHMKI	1
ANMHR	1
TSMGR	1
DHTKR	1
NGMRE	1
NAMRQ	1
EGMNQ	1
SNMRR	1
SGMRI	1

Table 10. ZF6
selection on A:C
change at nt 6 of
core motif in CBS.
Sequences reflect
position -1 to 3.

Sequence	Read #		
HRES	6362	RMES	13
MNES	5959	LRES	12
VKES	3337	RTDN	12
LRDS	2986	HADH	12
HLES	1799	VNES	12
TRES	1285	ANES	12
MREA	648	RTEN	12
VRET	601	RNEH	11
MRET	284	MNET	11
LLES	222	RLDT	11
MRTS	192	RADV	10
ERKS	122	RLET	9
IKES	111	HRET	9
RPDT	95	HMR...	9
VRVT	61	NRES	8
RNES	51	TGEA	8
HVES	41	TGES	8
RTET	40	RHET	8
LSHT	33	MRES	7
RPES	33	RNDT	7
SRES	32	LVES	7
ENKA	25	VGSS	7
RADN	24	RHDT	7
TREN	23	RIDT	7
DSPQ	21	VREA	6
RRES	20	HMES	6
RGEN	17	ERKN	5
VRES	17	RPEA	5
HRDS	15	TPPI	5
HREA	15	RREA	5
LRDT	15	RQEN	5
RVES	15	VKDS	4
EKKS	14	RKES	4
GRES	13	MLGL...	4
		DRPN	4
		RKEA	4
		VMLGL...	4
		TRDS	4
		HLET	4
		HLDS	4
		PPAT	4
		ENAS	4
		VKET	4
		GREA	4
		TREA	4
		H...	4
		IRDS	3
		MNDS	3
		LLDS	3
		RTES	3
		RPET	3
		IDVH	3
		RTEH	3
		TRET	3
		HGES	3
		TMES	3
		LRVS	2
		PREA	2
		EGKN	2
		TSES	2
		VKFGHFCVLL*NV...	2
		YRES	2
		MKES	2
		RTDI	2
		MNEG	2
		MIES	2
		QRES	2
		MMEA	2
		MNER	2
		RGS	2
		RTSS	2
		RNAS	2
		RTDT	2
		TRVS	1
		TFNV	1
		VRVS	1
		FRDS	1
		IKER	1
		RLEN	1
		IKET	1
		HRVS	1
		DRKG	1
		VKEC	1
		MSEA	1
		LRDR	1
		INES	1

MSES	1	HWES	1
NLES	1	RHEA	1
LQDS	1	MTES	1
HAPT	1	GRDS	1
HRR...	1	VSET	1
HRKA	1	MRKA	1
LRGS	1	FKES	1
QSGT	1	ERKG	1
HQES	1	VKR...	1
ETGS	1	RNDH	1
SGT...	1	VPDA	1
MLGF...	1	TGR...	1
MNGS	1	RKDA	1
MRED	1	SPDT	1
TKES	1	TTTL	1
RPDH	1	RKDS	1
HRGS	1	RRLT	1
GNES	1	RTSN	1
LWDS	1	LRT...	1
MRDS	1	RQSA	1
IHES	1	ARFT	1
LRDG	1	DRKS	1
LRDC	1	RRDT	1
MYES	1	RMDS	1
RPNI	1	HRKS	1
EGRS	1	GTPP	1
TRR...	1	DKRN	1
RLES	1	RPERE...	1
LGLPTGR...	1	SGDS	1
ARES	1	TAG	1
HLGS	1	GR...	1
HSES	1	T...	1
PRTS	1	...QSGT...	0
MNKS	1		
RRDS	1		
RREN	1		
QGES	1		
LREA	1		
LLET	1		
MREV	1		
VEES	1		
MNEA	1		
RNEN	1		

Table 11. ZF6
selection on A:G
change at nt 6 of
core motif in CBS.
Sequences reflect
position -1 to 3.

Sequence	# Reads
RPDT	6216
RNES	2750
RTET	1736
RPES	1565
RADN	1412
RTEN	973
RMES	860
RLDT	734
RRES	690
RTDN	588
RVES	584
RNEH	575
RIDT	557
RPEA	516
RGEN	467
RADV	455
RNDT	452
RHET	413
RKEA	340
RLET	297
RPET	263
RREA	252
RHDT	247
HRES	239
RKES	231
RQEN	199
RGSA	186
RTSS	154
RTDI	152
ERKS	123
MNES	104
RNAS	90
RREN	88
RLDP	82

RRDT	80	RGET	6
RTEH	80	RKGS	6
RKDA	63	RLES	5
VKES	61	MRET	5
RRET	53	RNDH	5
LRDS	44	RNDS	5
RTDT	36	STET	5
RLEN	35	GREA	5
TFNV	29	SNES	5
RADT	28	RPDA	4
HLES	28	RNER	4
TRES	27	RPEN	4
RKET	24	RVET	4
ATNM	23	RAET	4
RREH	22	SHET	4
RTDH	21	RSDT	4
RQSA	21	Q...	4
RTES	20	ENAS	3
RNET	20	LPDT	3
RPDN	19	MMES	3
THVP	19	SPES	3
ARFT	18	RMEN	3
ENKA	18	RVEI	3
GTTP	17	RRDS	3
EASN	16	RMET	3
RMEG	14	SADN	3
RTAA	14	RAES	3
RPDH	14	RPDV	3
SPDT	14	RTEA	3
DSPQ	14	RHES	3
EKKS	13	RQEA	3
MREA	13	MRTS	3
RNEP	12	ERKN	3
VHDN	12	RNRS	2
RKEN	12	RAEA	2
RPYT	12	RVDN	2
RQES	11	RNEG	2
RSGS	11	RVEG	2
RPDS	10	RAEN	2
VRET	10	RVDT	2
MTGN	7	RDDN	2
VMLGL...	7	MLEA	2
RNEN	7	RPNT	2

RGES	2	NRES	1	EGRS	1
SPEA	2	RIPA	1	RGR...	1
RTAG	2	RMEA	1	MADV	1
MKEA	2	RHNT	1	HTDN	1
SRES	2	RNSS	1	RKEV	1
WNES	2	LPES	1	RADA	1
GNES	2	SLDP	1	RDAS	1
RKDS	2	STEN	1	MLDT	1
TTTL	2	RPKS	1	RPGS	1
RVEN	2	ATS...	1	RTEY	1
RMDS	2	MIDT	1	SLDT	1
RMEH	2	PPDT	1	RWES	1
RRLT	2	GLDA	1	ERKA	1
RKEH	1	RPEGE...	1	RIYT	1
ENRS	1	RHYT	1	TPVP	1
RNKS	1	RTEI	1	RQDA	1
RPGE...	1	SPEN	1	RMER	1
RKDT	1	APR...	1	RTSN	1
VPDA	1	LSL...	1	LRT...	1
RGEA	1	RHEN	1	MMEA	1
WIDT	1	REDV	1	VRVT	1
RNEY	1	RLKT	1	DRKS	1
RADI	1	RIET	1	TREN	1
RADY	1	RIES	1	HRKS	1
RTDD	1	LLES	1	LRES	1
RVDS	1	RPDI	1	RPERE...	1
HTET	1	MNDT	1	DKRN	1
HTEN	1	RLYT	1	VGTV	1
SGEN	1	ANES	1	...QSGT...	0
RTST	1	RAYN	1		
RAGR...	1	RADS	1		
SNAS	1	KNES	1		
RPGT	1	RVSA	1		
RAEH	1	RPED	1		
MHDT	1	RGEH	1		
REDN	1	RRER...	1		
REEV	1	LTET	1		
RRR...	1	LADN	1		
RMEW	1	GTR...	1		
RRER	1	RPER...	1		
RLDN	1	MLGLPGTR...	1		
RPT...	1	RPDP	1		
MVES	1	QADV	1		

Table 12. ZF6
selection on A:C
change at nt 6 of
core motif in CBS.
Sequences reflect
position -1 to 3.

Sequence	# Reads
HRES	7487
ERKS	7125
DSPQ	876
ENKA	801
EKKS	508
TRES	141
RPDT	126
ERKN	120
GREA	112
ENAS	103
LRDS	95
MNES	89
ANES	84
DRKG	73
DRKS	72
EGRS	69
ETGS	67
RNES	60
VKES	57
SRES	50
RTET	42
VPDA	39
RRLT	37
RADN	30
RPES	30
ERGG	27
HLES	25
DKRN	25
RTEN	21
TTTL	19
HRKS	19
RRES	19
RMES	18
EGKN	17

ERKG	16	ERKR	3
RGEN	16	RNAS	3
ENRS	15	ERKI	3
GRES	15	ERRS	2
RTDN	15	RLDP	2
RLDT	13	HSES	2
RVES	13	YSPQ	2
RIDT	13	EGKS	2
NRES	13	HRER	2
RNEH	12	QVK...	2
EKGT	11	DRKA	2
RHET	11	ESGN	2
MREA	10	QG...	2
RPET	9	ERES	2
Q GK	9	HKES	2
HMES	9	ESKS	2
VRET	9	QRES	2
RADV	9	EMKS	2
HMR...	9	SPDT	2
ERNS	8	RRDT	2
RQEN	8	MLGL...	2
RNDT	8	ARFT	2
RHDT	8	RTDT	2
HRDS	7	KRKS	1
RTSS	7	EASN	1
RKEA	7	TGDA	1
ENNS	6	NRKS	1
ERKA	6	RGK	1
RREA	5	EKNS	1
RGS	5	HRE...	1
QEK...	5	QGKS	1
MRTS	5	MTGN	1
LRES	4	DSTQ	1
RKES	4	TGE...	1
HREN	4	VRKS	1
RPEA	4	HRET	1
HGES	4	ENKV	1
RTDI	4	RLEN	1
RLET	4	RMEA	1
LLES	3	HREA	1
RKDA	3	RAEN	1
MRET	3	SLDT	1
HRGS	3	R...	1

TGES	1	RKDS	1
RADT	1	LRT...	1
QAK...	1	VRVT	1
DIPQ	1	TREN	1
QGT...	1	RPERE...	1
ERKC	1	GTTP	1
HSPQ	1	SGDS	1
RTES	1	...QSGT...	0
TREA	1		
RTAT	1		
QGR...	1		
TRKS	1		
GRKS	1		
ESKA	1		
ERK...	1		
YRES	1		
EKRN	1		
MGK...	1		
DSPH	1		
ERNG	1		
VSPQ	1		
QWK...	1		
EKKC	1		
LGLPTGR...	1		
ERNN	1		
RRET	1		
TNES	1		
HRKN	1		
RLF...	1		
DKSN	1		
DRNS	1		
KRN	1		
ERMS	1		
RREN	1		
EIAS	1		
HREC	1		
ERKT	1		
ETGN	1		
RQSA	1		
RTSN	1		
RMDS	1		
RTEH	1		
MMEA	1		

Table 13. ZF5
selection on G:T
change at nt 7 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence # Read

TRLKE 2129
HRLKE 1938
SRLKE 1530
TRLRE 1078
HRLRE 1073
NRLKE 1015
QRLRE 769
DALKR 700
DGLKR 681
SRLRE 534
HALKV 389
NRLKV 381
ERLRV 375
DGLKK 374
HGLKV 335
HRLKV 315
ERLRM 295
QRLKE 243
DGLVR 235
HTLKV 233
NRLRE 195
ARLRE 168
DALRR 168
ERLRQ 141
ARLKE 135
TRLRD 125
DGLRR 118
SRLNE 118
TGLKV 92
HRLSE 91
HRLNE 78
SHLKV 75
TTLKV 75
HRLGE 68

STLKV	66	HKLKV	11
DGLKV	65	GTLKV	10
DGLRK	61	ERLRR	9
HRLTE	60	HALKT	9
DRLKV	59	HGLKE	9
HSLKV	56	HHLVQ	9
DGLRV	47	NGLKV	9
SRLKV	45	DALKE	8
QRLKV	44	DALKV	8
HGLTV	43	HALKE	8
HRLME	43	HHLKQ	8
RLLPN	42	HHLKV	8
ERLKV	41	TRLKK	8
NRLRV	35	DRLRT	7
TRLKV	34	DRLRV	7
DGLKE	29	DTLRV	7
DTLKV	29	HRLKK	7
HGLRV	29	HTLKE	7
SALKT	28	NRLKK	7
HRLAE	25	STLKE	7
ERLIS	23	SRLIE	6
DGLTR	22	TRLME	6
DALVR	21	ATLKV	5
HRLKR	21	HGLVV	5
ERLRE	20	HRLRM	5
HQLKV	20	HRLRQ	5
TTLKQ	18	HTLKA	5
SRLKR	17	NRLRD	5
DRLKQ	16	TGLKE	5
HRLRV	16	TGLKT	5
TRLKR	16	TRLRQ	5
TRLNE	16	TTLKI	5
NRLKQ	15	TTLRV	5
TRLKD	14	DRLKE	4
TRLRV	14	HRLKA	4
EALKR	13	HRLKD	4
HTLKQ	13	HSLKE	4
NALKV	13	NRLKI	4
SALKV	13	NRLKR	4
SRLKD	13	STLKA	4
DGLRE	12	STLKQ	4
ERLKE	12	TRLKA	4
DTLKQ	11	TRLKQ	4

TRLRR	4	QRLRQ	2	DTHAG...	1
DTLKA	3	QSLIA	2	DTLKT	1
HALKR	3	QTLKV	2	DVLKK	1
HGLKA	3	RKLRS	2	EAAG...	1
HGLKR	3	RRLRE	2	EHLRQ	1
HPEG...	3	SALKE	2	ELLKV	1
HRLK...	3	SRLKK	2	EPLRV	1
HRLRK	3	SRLRK	2	ERLCV	1
HTLRV	3	SRLRV	2	ERLKK	1
NTLKQ	3	STLRV	2	ERLRR...	1
QRLRV	3	TMLKE	2	ERLVR	1
SRLME	3	TRLKG	2	ERLWE	1
SRPKE	3	TRLRM	2	ERPRM	1
TQLKV	3	TRLTE	2	ERPRV	1
TRLQE	3	TRRKE	2	ERQRM	1
TRLR...	3	AALKR	1	GGLKQ	1
ARLKR	2	AGLKR	1	GGLKR	1
ARLKV	2	AGLKV	1	GMLKV	1
ARLR...	2	AGLVR	1	GRLKE	1
ARLRV	2	ARLGE	1	GTLKQ	1
ARLVR	2	ARLME	1	HALKA	1
DALKK	2	ARLNE	1	HALKG	1
DALRV	2	ARLRD	1	HALPV	1
DAPKR	2	ARLRM	1	HAPEV	1
DRLRE	2	CRLKE	1	HGLKK	1
EGLKV	2	DALDR	1	HGLKQ	1
ERLLV	2	DALKT	1	HGLMV	1
ERLRA	2	DALKW	1	HGLPV	1
ERM RM	2	DALRK	1	HGLRD	1
GGLKV	2	DALTV	1	HGLVR	1
GGLVT	2	DELKR	1	HGQKE	1
HALRE	2	DELPG	1	HGRKV	1
HGLRE	2	DGLK...	1	HGRRG	1
HHLKE	2	DGLKG	1	HHLRV	1
HILKA	2	DGLKW	1	HILIA	1
HRLQE	2	DGLLR	1	HKLKE	1
HRLRR	2	DGLRQ	1	HKLRV	1
KRLKE	2	DGLTV	1	HMLKR	1
KTLKQ	2	DGLVW	1	HMLRE	1
NALKE	2	DGMKR	1	HNLKV	1
NRLNE	2	DKLKQ	1	HPLKV	1
NTLKV	2	DKLRQ	1	HQLKE	1
QRLKR	2	DRLRK	1	HQLRE	1

HQLRV	1	NRLKF	1	SKLKV	1
HR*A...	1	NRLKL	1	SNLKV	1
HRGCG...	1	NRLKT	1	SQLKV	1
HRLDE	1	NRLME	1	SRLIG	1
HRLIE	1	NRLND	1	SRLK...	1
HRLKF	1	NRLNV	1	SRLKA	1
HRLKG	1	NRLQE	1	SRLKG	1
HRLKL	1	NRLR...	1	SRLQE	1
HRLMV	1	NRLRM	1	SRLR...	1
HRLN...	1	NRLRQ	1	SRLRA	1
HRLR...	1	NRMKE	1	SRLRM	1
HRLRA	1	NRPKE	1	SRLRQ	1
HRLS...	1	NRPKV	1	SRLTE	1
HRLVR	1	NRQKE	1	SRQRE	1
HRMRE	1	NSLKE	1	SSLKE	1
HRPKE	1	NTLTV	1	SSLKV	1
HRPNE	1	PRLKE	1	SSQRE	1
HRQRE	1	PRLLP	1	STLKR	1
HRRKE	1	PRLRE	1	TAG...	1
HRRME	1	PRLTE	1	TGLKG	1
HRRRE	1	QAEG...	1	TGLKQ	1
HRVRE	1	QRLIS	1	TGLKS	1
HSACG...	1	QRLKK	1	TGLRV	1
HSLNV	1	QRLME	1	TGRRG	1
HSLRV	1	QRLRG	1	TLLRE	1
HTLAQ	1	QRLRM	1	TMQKE	1
HTLNV	1	QRLTE	1	TRL*L	1
HTMKV	1	QTA*R...	1	TRLAE	1
HVLKV	1	QTAW...	1	TRLE...	1
HWLRE	1	QTG*S...	1	TRLEE	1
KGLKQ	1	R...	1	TRLGE	1
MHLRS	1	RGLKV	1	TRLK...	1
MRLRE	1	RRLGD	1	TRLKY	1
MRLRM	1	RRLKE	1	TRLRG	1
NALKR	1	RRLNE	1	TRLRK	1
NGLKE	1	RRLTK	1	TRLSE	1
NLLRE	1	SALKK	1	TRPKE	1
NMLKE	1	SALKR	1	TRQRD	1
NMLNV	1	SCLKE	1	TRRRD	1
NPLRE	1	SGLAM	1	TRVRE	1
NRFKE	1	SGLAV	1	TSLRE	1
NRLIE	1	SGLKV	1	TTLKA	1
NRLKA	1	SHLKE	1	TTLKE	1

TTLKL	1
TTLKT	1
TTPRG	1
TTRKQ	1
TWLRE	1
VRRKV	1
YGLKR	1
YRLKE	1
YTLKV	1

**Table 14. ZF5
selection on G:C
change at nt 7 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	Read #
SRLKE	2533
TRLKE	2146
HRLKE	1984
NRLKE	1528
HRLRE	1001
SRLRE	799
TRLRE	625
HTLKV	499
HGLKV	320
QRLRE	299
STLKV	249
ARLKE	238
QRLKE	135
STLKE	126
TTLKV	102
DTLKA	95
HKLKV	87
DTLKV	84
HALKV	82
HILKA	80
TGLKV	80
ARLRE	78
HRLGE	75
NRLRE	75
HRLTE	61
RLLPN	55
DGLKV	50
HRLKV	46
HSLKV	43
HGLKA	37
NRLKV	30
HHLKV	27
QSLIA	26

DTLRV	25	SRLNE	7
TRLKV	25	TGLKE	7
TTLKE	25	DALRR	6
HTLKE	24	HGLKT	6
SRLKD	23	HGLTV	6
ARLVR	20	HKLKE	6
SKLKE	20	HRLME	6
DRLKV	19	SALKV	6
ERLRV	19	ATLKV	5
SRLKV	19	DTLKE	5
DALKR	18	GILND	5
DGLKR	18	SGLKV	5
GRLKE	18	SRLKA	5
GTLKV	18	DALVR	4
TRLRD	17	ETLKV	4
HTLKA	16	HRLSE	4
HRLAE	15	HSLKE	4
DTLKQ	14	HTLRV	4
HGLKE	14	NTLKV	4
HILKT	14	QRLKV	4
RKLRS	14	SRLK...	4
SGLAM	12	THLKE	4
TTLKQ	12	TRLKQ	4
STLKA	11	ARLKV	3
DGLKK	10	DGLRK	3
RRLGD	10	DGLRR	3
TRLKD	10	ERLRQ	3
TSLKV	10	GGLKV	3
DGLVR	9	HGLRE	3
ERLRM	9	HHLKQ	3
HHLKE	9	KRLKE	3
HHLVQ	9	NALRV	3
HRLNE	9	NGLKV	3
HTLKQ	9	NRLGE	3
TRLNE	9	NRLKK	3
HRLHE	8	RRLRE	3
SHLKV	8	SCLKE	3
SKLRM	8	SKLKV	3
DGLRV	7	SRLRD	3
ERLKV	7	SRLRV	3
GTLRV	7	STLKQ	3
HRLKA	7	TRLKA	3
HTLKS	7	TRLQE	3

YTLKV	3	DRLRE	1	HRLVR	1
APLLR	2	DRLRT	1	HRMRE	1
DKLKQ	2	DTLKT	1	HRRRE	1
DKLKV	2	DTPKA	1	HSG*G...	1
DRLKE	2	ERLIS	1	HSLKQ	1
ERLRE	2	ERLRR...	1	HSLRE	1
HALKE	2	ERPRM	1	HSVKA	1
HHLRV	2	GGLKQ	1	HTG*R...	1
HQLKE	2	GMLKV	1	HTLAQ	1
HRLEE	2	GRLKA	1	HTLEV	1
HRLRV	2	GRLKV	1	HTLME	1
HTLKG	2	GTLKQ	1	HTLMV	1
KTLKQ	2	GVLKE	1	HTLQE	1
MVLVV	2	GVLTG	1	HTLRQ	1
NRLKA	2	HALDV	1	HVLKV	1
NRLKD	2	HALKA	1	IRLKE	1
NRLKQ	2	HELV	1	IRQEE	1
NRLQE	2	HGLEA	1	KGLKQ	1
NRQKE	2	HGLKQ	1	KRLKV	1
NTLKA	2	HGLRG	1	LRLKK	1
QRLRV	2	HGMKA	1	NKLKE	1
QTLKV	2	HGPKV	1	NKLKG	1
RRLKE	2	HILIA	1	NRFKE	1
SRLKQ	2	HILKE	1	NRLAE	1
SSLKV	2	HILKV	1	NRLEE	1
STLRQ	2	HILNA	1	NRLKR	1
STLRV	2	HKLKG	1	NRLKS	1
TRLTE	2	HKLKQ	1	NRLKT	1
TSLRE	2	HKL RV	1	NRLRD	1
AGLKV	1	HMLRE	1	NRLRG	1
AQMKE	1	HPEG...	1	NRLRV	1
ARLNE	1	HPLKE	1	NRLTE	1
ARLRD	1	HQLKV	1	NSLKE	1
ARRRE	1	HRLGV	1	NTLKQ	1
CRLM...	1	HRLKD	1	PAEG...	1
CRLMV	1	HRLKG	1	PPPPE	1
DALKE	1	HRLKL	1	PRLRE	1
DALKV	1	HRLKR	1	QAEG...	1
DELPG	1	HRLLE	1	QGRRE	1
DGLKE	1	HRLQG	1	QRLEE	1
DKLRQ	1	HRLRA	1	QRLRG	1
DRLKA	1	HRLRS	1	QSLGR	1
DRLKT	1	HRLTV	1	SGLAV	1

SKLK...	1
SMLRE	1
SRLAE	1
SRLCE	1
SRLKG	1
SRLLE	1
SRLQE	1
SRLRA	1
SRLRK	1
SRPKE	1
SRQRE	1
SRRKE	1
SSLKE	1
SSQRE	1
STLRA	1
STLRE	1
TGLKQ	1
TGLRE	1
THLKV	1
TILYE	1
TLLKE	1
TRLKG	1
TRLME	1
TRLRV	1
TRLTV	1
TRMGE	1
TRMKQ	1
TTLKA	1
TTLKI	1
TTLKL	1
YTLKE	1

**Table 15. ZF5
selection on G:A
change at nt 7 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	Read #				
HTLKV	3934	STLRV	78	HHLRV	15
HGLKV	2682	DTLKA	75	NRLKV	15
STLKV	2167	GGLKV	75	GTLRV	14
RLLPN	1887	TRLKQ	75	HTLKE	14
TTLKV	1471	KTLKQ	69	SVLKV	14
HALKV	923	HTL*A	66	TGLKQ	14
DTLKV	888	HTLKA	64	ERLRM	13
TTLKQ	754	SRLRE	61	NRLRE	13
HKLKV	571	TRLRE	58	SKLKV	13
HTLKQ	513	NTLKQ	56	SNLKV	13
TGLKV	482	SKLKQ	55	TVLKV	13
GTLKV	473	QRLRE	53	ARLKE	12
DTLKQ	462	DKLKV	51	ARLRE	12
DGLKV	443	TTLRV	48	DGLKK	12
HSLKV	352	SSLKV	43	HHLRQ	12
HHLVQ	298	SHLKV	42	HTLKG	12
HHLKV	259	GTLKQ	40	HVLKV	12
HRLKV	210	TRLKV	40	TRLRQ	12
HRLKE	190	THLKV	38	NTLRQ	11
DTLRV	189	NTLKV	37	SGLAV	11
SRLKE	186	STLRQ	35	DALRR	10
TRLKE	178	HALRV	34	QTLKQ	10
SALKV	177	HGLKA	33	RTLKQ	10
ATLKV	155	DKLKQ	32	STLKE	10
HHLKQ	149	DGLKR	29	AGLKV	9
STLKA	140	EGLKV	29	DTLKT	9
NGLKV	136	ETLKV	29	HHLVV	9
NRLKE	124	GILKV	28	HNLKV	9
STLKQ	118	DRLKV	26	QRLKE	9
QTLKV	112	HGLRV	24	STLKG	9
HQLKV	94	SVLKQ	23	TVLKQ	9
SGLKV	91	DALKR	22	DGLVR	8
HRLRE	89	ERLRV	22	GGLVV	8
HTLRV	88	GQLHV	21	HGLQV	8
		TTLRQ	21	HRLGE	8
		DGLRV	20	HRLTE	8
		DTLKN	20	HTLRQ	8
		TTLKT	20	SRLKV	8
		GVLKV	17	DTLKG	7
		DKLRQ	16	GGLTV	7
		GTLKA	16	GVLKA	7
		GGLKQ	15	HALKA	7
		HGLKQ	15	HGLKT	7

TTLKA	7	DALVR	2	ATLKG	1
GTLRQ	6	DELPG	2	DALKE	1
HALKQ	6	DGLRA	2	DALKV	1
SRLNE	6	DVLKV	2	DELRO	1
STLKT	6	GALRV	2	DGLKA	1
TRLRD	6	GGLVQ	2	DGLKE	1
ATLKA	5	GNLKV	2	DGLKL	1
ATLKQ	5	GPLKV	2	DKLKG	1
DGLKQ	5	GTLKG	2	DRLKQ	1
DGLRR	5	GVLKQ	2	DRLRE	1
ERLKV	5	GVLRV	2	DRLRV	1
GTLNA	5	GVLVA	2	DSLKV	1
GVLKN	5	HALKG	2	DTLKE	1
HGLKE	5	HDLKV	2	DTLNQ	1
HGLVV	5	HGLEV	2	DTLQA	1
TTLKG	5	HGPKV	2	DTLRA	1
DGLRK	4	HHMVQ	2	DTLRL	1
ERLRQ	4	HILKA	2	DTLWQ	1
ETLRV	4	HKLKA	2	DTMKV	1
HGLKG	4	HKLKE	2	EGLKQ	1
HGLNV	4	HKLKQ	2	ERLRA	1
HGLTV	4	HRLKA	2	ERLRE	1
HHLKE	4	HRLKQ	2	ERPRM	1
HRLSE	4	HRLNE	2	ETLKE	1
NALKV	4	KGLKQ	2	ETRRV	1
STLTV	4	KTLKV	2	GGLAV	1
VGLGE	4	PTLKV	2	GGLRG	1
YTLKV	4	QSLIA	2	GGLRV	1
AGLAT	3	RLLPY	2	GHLKA	1
ARLVR	3	RLRPN	2	GILND	1
D*LPG	3	RTLAQ	2	GMLKV	1
DGLKG	3	RTLKV	2	GPLRA	1
DKLRV	3	SALTV	2	GQQHV	1
HALKE	3	STLKL	2	GTLQA	1
HRLME	3	TGLKT	2	GTPKV	1
HSLKQ	3	TKLKQ	2	HALES	1
HTLNV	3	TTLKI	2	HALKF	1
HTMKV	3	TTPKV	2	HALMV	1
HTQKV	3	AGLAS	1	HALPV	1
QSLGR	3	AGLKM	1	HAMKV	1
RKLRS	3	APLKV	1	HARKV	1
STLNV	3	ARLKV	1	HELVK	1
TTLKE	3	ARLRD	1	HGLKD	1

HGLKL	1	HTLTV	1	RTLKG	1
HGLKM	1	HTPKV	1	SALKQ	1
HGLKW	1	HTRKQ	1	SALRQ	1
HGRKI	1	HVLKF	1	SELKV	1
HGRKV	1	HVMKV	1	SFLKV	1
HHLAQ	1	HWLKV	1	SGLAM	1
HHLGQ	1	KADTV	1	SGLKQ	1
HHLMQ	1	KGLKG	1	SHLKQ	1
HHMKV	1	KRLKQ	1	SKLKA	1
HILIA	1	KTLAQ	1	SKLKE	1
HILKV	1	KTLRV	1	SRLKD	1
HKLKG	1	KTLTQ	1	SRLRA	1
HKLKM	1	LHLKV	1	SSLRV	1
HKLVN	1	LTLKQ	1	SSQRE	1
HKLQE	1	LTLKV	1	STLKK	1
HKLRV	1	MGLKV	1	STLKM	1
HMLNV	1	MPPK	1	STLMV	1
HPLDV	1	MRLKQ	1	STLQQ	1
HPLKV	1	NAVTE	1	STLTA	1
HPLQV	1	NGLKG	1	STMKA	1
HQLKA	1	NGLKL	1	STMKV	1
HQLKG	1	NRLKG	1	TALKV	1
HQLKT	1	NRLRD	1	TGLKA	1
HRLAE	1	NRLRV	1	TGLKD	1
HRLKG	1	NTLRV	1	TGLKE	1
HRLKL	1	PGLKV	1	TGLKG	1
HRLKR	1	QGLKV	1	TGLMV	1
HRLRQ	1	QRLKV	1	THLKE	1
HRLVR	1	QRLRV	1	THLKG	1
HRPKE	1	QRQRV	1	THLKL	1
HSLKA	1	QTLKA	1	THLKQ	1
HSLKE	1	QTLKG	1	THLMV	1
HSLKG	1	QTLKK	1	TKLKV	1
HSLKL	1	QTLKM	1	TPLQV	1
HSVKA	1	QTLMV	1	TRLKD	1
HTLAQ	1	RGLKV	1	TRLKG	1
HTLAV	1	RHLVQ	1	TRLPQ	1
HTLEV	1	RLLPT	1	TRLQE	1
HTLKN	1	RLLSN	1	TTLEV	1
HTLMV	1	RLMPD	1	TTLHV	1
HTLNA	1	RMLPN	1	TTLNQ	1
HTLQV	1	RRLGD	1	TTLQV	1
HTLTE	1	RSLKV	1	TTLRG	1

TTLVY	1
TTMKV	1
TVLRQ	1
VGLGG	1
VTLKV	1

Table 16. ZF5
selection on G:A
change position 8
of the CBS core
motif. Sequences
reflect position 2
to 6.

Sequence	# Read
GGLRR	341
GGLVR	336
TGLRR	274
EGLRR	267
DGLRR	232
SGLRR	206
AGLAR	179
SGLAR	178
GGLAR	177
GGLTR	168
DGLAR	152
AGLKR	148
TGLAR	135
DGLVR	129
GGLQR	127
GNLTR	124
GNLVR	123
HGLAR	123
GGLKR	111
TGLVR	108
AGLTR	105
SGLSR	102
AGLRR	100
GGLSR	94
HGLRR	91
HGLVR	87
SGLTR	84
NGLVR	80
AGLQR	79
GNLRR	79
AGLHR	76
GNLER	76
HNLLR	76

GNLAR	73	TGLSR	37
DGLTR	72	GNLLR	36
HALRR	69	NNLVR	36
HGLQR	69	TGLIR	36
NGLRR	69	DMLRR	35
SGLVR	68	GALKR	35
SNLDR	67	GNLDR	35
TNLRR	66	SALRR	35
HGLTR	63	SNLAR	35
SSLRR	63	SGLLR	34
DALRR	61	TNLNR	33
EGLTR	61	AGLLR	31
GGLER	61	GGLIR	31
DGLKR	60	DGLHR	30
TGLQR	60	DTLRR	30
HTLRR	59	HLLKR	30
AALKR	58	SALAR	30
AGLVR	55	SMLAR	30
AGLIR	54	VGLKR	30
HGLKR	54	DNLLR	28
ANLVR	53	GGLMR	28
EGLKR	53	SGLMR	28
SNLLR	51	AALRR	27
EGLAR	50	ETLRR	27
AGLSR	49	NGLAR	27
DGLIR	48	TGLRV	27
TGLKR	48	TGLTR	27
SGLQR	46	TNLQR	27
ETLKR	45	ANLAR	26
HGLLR	45	NNLAR	26
NGLQR	45	SNLSR	26
TGLMR	45	STLSR	26
ANLRR	43	AALAR	25
DNLVR	42	HALVR	25
TGLLR	42	HGLSR	25
DGLMR	41	SGLNR	25
ASLKR	39	STLAR	25
QGLRR	38	ANLIR	24
TNLVR	38	DGLDR	24
NGLTR	37	DGLSR	24
SGLDR	37	GTLKR	24
SGLHR	37	EALKR	23
TGLNR	37	NGLSR	23

SMLRR	23	SGLIR	15	ESLRR	11
HNLHR	22	TNLKR	15	NALRR	11
HNLRR	22	DGLGR	14	NGLKR	11
SGLKR	22	DSLQR	14	NNLLR	11
TGLGR	22	EGLNR	14	QGLKV	11
TNLMR	22	ENLRR	14	SNLRR	11
DALVR	21	GSLRR	14	STLRR	11
GTLTR	21	NGLNR	14	VNLSR	11
DGLNR	20	QALKR	14	DMLKR	10
SSLVR	20	SALSR	14	GALRR	10
TGLER	20	SSLGR	14	GGLDR	10
DTLKR	19	VNLKR	14	HGLMR	10
GNLSR	19	ATLRR	13	HNLVR	10
HGLIR	19	DGLLR	13	HQLIR	10
HSLVR	19	EMLKR	13	NALKR	10
AGLNR	18	GALVR	13	NTLKV	10
DALAR	18	GNLGR	13	QNLQR	10
GGLHR	18	GNLQR	13	SALKV	10
NGLIR	18	HALAR	13	SMLIR	10
QGLTR	18	HSLIR	13	TALRV	10
QMLKR	18	HTLER	13	TNLAR	10
QNLRR	18	HTLQR	13	TQLKR	10
TGLKV	18	NGLER	13	TTLKV	10
AILKR	17	NGLMR	13	TTLTR	10
GNLKR	17	QGLVR	13	VGLQR	10
GNLMR	17	TALKR	13	AALSR	9
HNLTR	17	TTLMR	13	ATLAR	9
HTLAR	17	VGLRR	13	DALGR	9
QGLKR	17	ANLKR	12	DTLNR	9
SGLER	17	ANLNR	12	EILKR	9
SGLGR	17	ATLTR	12	ESLKR	9
SNLVR	17	DNLRR	12	GGLNR	9
EALRR	16	ENLKR	12	GSLTR	9
GTLRR	16	GGLLR	12	HNLAR	9
HGLGR	16	GTLVR	12	MGLKR	9
HTLMR	16	HNLSR	12	NGLHR	9
NTLRR	16	NTLKR	12	NMLKR	9
TGLHR	16	SALER	12	PNLKR	9
TSLRR	16	SSLTR	12	SALTR	9
TTLQR	16	TALVR	12	SDLKR	9
DNLKR	15	ANLSR	11	STLGR	9
GALTR	15	DNLAR	11	AGLER	8
QTLRR	15	ENLSR	11	DILRR	8

DMLNR	8	QMLRR	6	HILTR	4
DTLAR	8	TNLHR	6	HLLLR	4
HALLR	8	TSLHR	6	HNLQR	4
HALSR	8	VGLAR	6	HTLGR	4
HNLGR	8	AALQR	5	IGLTG	4
NALVR	8	AGLDR	5	NGLLR	4
SMLTR	8	ATLKR	5	NSLRR	4
TALAR	8	DGLKK	5	PNLIR	4
TNLER	8	DTLQR	5	PNLRR	4
TNLGR	8	DVLKR	5	SALIR	4
TTLNR	8	GALSR	5	SILGR	4
DALLR	7	GMLKR	5	SPLVR	4
DSLAR	7	GTLSR	5	STLTR	4
GTLAR	7	HNLER	5	TALKT	4
GTLIV	7	NGLLV	5	TALTR	4
HALIR	7	NNLTR	5	TGLDR	4
HGLDR	7	QALAV	5	TSLKR	4
HGLER	7	QGLAR	5	TTLVR	4
HTLLR	7	QNLHR	5	VGLQN	4
NNLIR	7	SALMR	5	VNLRR	4
NNLMR	7	SLLLR	5	AALVR	3
QSLKR	7	SVLAR	5	ADLKR	3
SALGR	7	SVLTR	5	ANLGR	3
SALVR	7	TALRR	5	ATLSR	3
SNLMR	7	TMLRR	5	DNLQR	3
SQLRR	7	TQLRV	5	DNLTR	3
STLQR	7	TTLLR	5	DRLRR	3
STLVR	7	TTLRR	5	DTLVR	3
SVLKR	7	AALNR	4	EGLVR	3
TSLKV	7	ATLVR	4	GALNR	3
AALTR	6	DALHR	4	GDLKR	3
DSLKR	6	DALMR	4	GDLTR	3
DSLRR	6	DGLER	4	GGLGL	3
DTLMR	6	DGLQR	4	GSLQR	3
EGLLR	6	DGLRV	4	HALKR	3
ENLAR	6	DLLRR	4	HGLHR	3
GNLNR	6	DRLKV	4	HGLRV	3
GTLQR	6	GGLGR	4	HTLKR	3
HALDR	6	GNLHR	4	HVLKR	3
HVLER	6	GTLKV	4	NGLDR	3
IGLRR	6	GTLNR	4	NMLAR	3
INLTR	6	HALHR	4	NSLAR	3
NMLRR	6	HALMR	4	NTLAR	3

QGLHR	3	HSLHR	2	AALLR	1
SGLAV	3	HSLRR	2	AALMR	1
SILTR	3	HTLKV	2	AAPER	1
SILVR	3	HTLNR	2	ADLRR	1
SQLKR	3	HTLTR	2	AGLAW	1
SSLQR	3	HTLVR	2	AGLRW	1
TALHR	3	IGLKR	2	AGLTS	1
TALNR	3	ITLKR	2	AILTR	1
TALSR	3	MTLKR	2	AMLKR	1
AGLGR	2	NALHR	2	ANLPR	1
AGLMR	2	NALSR	2	ARLKR	1
ASLQR	2	NGLGR	2	ARLQR	1
ASLVR	2	NTLHR	2	ARLTR	1
ATLMR	2	QDLKR	2	ASLRR	1
AVLKR	2	QGLLR	2	ASLTR	1
DALNR	2	QNLLR	2	ATLDR	1
DALQR	2	QNLRW	2	ATLER	1
DALSR	2	QSLRR	2	ATLIR	1
DGLRK	2	QTLKR	2	ATLLR	1
DHLHR	2	SALKR	2	ATLQR	1
DHLVR	2	SALRV	2	AVLRR	1
DNLSR	2	SSLAR	2	DALKR	1
DTLSR	2	SSLSR	2	DALRV	1
DTLTR	2	STLDR	2	DGLSV	1
DVLRR	2	STLER	2	DILHR	1
EGLIR	2	STLHR	2	DQLRR	1
EGLSR	2	STLKV	2	DSLSR	1
GAEE...	2	STLMR	2	DTLAK	1
GALQR	2	TALGR	2	DVLLR	1
GALRV	2	TGLAT	2	EALNR	1
GDLRR	2	TGLSV	2	EALTR	1
GDLVR	2	TGLVT	2	EGLKV	1
GGLKV	2	TNLKV	2	EGLMR	1
GGLRV	2	TNLSR	2	EGLQR	1
GSLAR	2	TTLAR	2	EGLRL	1
GSLKR	2	TTLGR	2	EGLRV	1
HDLRR	2	TTLIR	2	EGVRR	1
HGLNR	2	TTLKR	2	ELLRR	1
HHLIR	2	TTLKT	2	ENLER	1
HMLKR	2	TVLRM	2	ETLLR	1
HMLRR	2	VQLAM	2	GALHR	1
HQLVR	2	VTLTR	2	GGHRR	1
HSLAR	2	A*S...	1	GGLAG	1

GGLAV	1	NTLIR	1	TSLGG	1
GGLDV	1	NTLNR	1	TSLGR	1
GGLGS	1	NTLQR	1	TSLQR	1
GGLQE	1	NVLKR	1	TSLVR	1
GGLVL	1	PALKR	1	VALAR	1
GGLVT	1	PGLLR	1	VALKR	1
GGPSH	1	PWS...	1	VALSR	1
GGPSR	1	QAAWG...	1	VGLKC	1
GGQRR	1	QALAR	1	VGLSR	1
GGVRR	1	QALTR	1	VGLTM	1
GGWR...	1	QDLIR	1	VNLAR	1
GILER	1	QTLAR	1	VNLIR	1
GKLRR	1	QTLQR	1	VNLNR	1
GMLAR	1	QVLRR	1	VTLGR	1
GNLIR	1	RGLTR	1	VTLKR	1
GSLER	1	RGLVR	1	VTLMR	1
GSLVR	1	SALDR	1	VTLRR	1
GTLER	1	SALMC	1	WGLER	1
GTLGR	1	SALNR	1		
GTLHR	1	SDLAR	1		
GTQVR	1	SDLQR	1		
GVLRR	1	SDLRR	1		
GVLTR	1	SGPRR	1		
HALGR	1	SLLSD	1		
HALKV	1	SMLHR	1		
HDLAK	1	SNLQR	1		
HGAAR	1	SSLIR	1		
HGLKK	1	SSLKR	1		
HGLKM	1	STLLR	1		
HGLKV	1	STLNR	1		
HGLSV	1	STLRK	1		
HGLTW	1	SVLGR	1		
HGPAR	1	SVLRR	1		
HKLAR	1	TALER	1		
HNLLS	1	TALRT	1		
HRLSR	1	TDLAR	1		
HSLNR	1	TDLRR	1		
HSLSR	1	TGLQV	1		
HTLHR	1	TGLVRR	1		
HVLAR	1	TGPAR	1		
INLSR	1	TMLKR	1		
NALAR	1	TNLPR	1		
NHLVQ	1	TSLAR	1		

**Table 17. ZF5
selection on G:C
change at nt 8 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence # Read

DGLRR 498
DALRR 388
EGLRR 348
DALVR 288
DGLVR 262
DGLAR 261
AGLKR 257
DGLTR 255
TGLRR 237
DTLRR 196
HGLRR 192
ATLRR 176
ETLRR 149
SGLRR 142
GTLRR 136
GGLVR 132
STLSR 132
GTLQR 131
DGLSR 127
GGLKR 126
GTLAR 123
HTLAR 114
SGLVR 111
GDLKR 109
QTLRR 107
TGLVR 106
STLGR 103
DTLVR 102
DGLNR 100
DVLRR 100
STLVR 100
DVLKR 99
ATLKR 96
GGLRR 93

STLRR 93
AVLKR 91
GDLRR 90
GTLKR 89
HGLKR 89
HTLRR 89
TGLQR 87
SGLKR 86
DTLAR 84
ATLAR 83
STLER 81
TTLRR 80
EGLKR 79
GTLTR 79
TGLAR 75
DGLHR 74
HGLVR 74
HTLNR 73
STLNR 71
SSLRR 69
TGLSR 68
ATLSR 67
DTLNR 67
NTLAR 66
ADLKR 65
DALAR 65
GTLVR 64
DTLQR 62
ATLVR 61
NGLVR 61
DGLKR 59
GGLAR 59
HSLVR 59
DALQR 58
TGLLR 58
DMLKR 57
GGLTR 56
ADLRR 55
SGLSR 55
SSLVR 54
TGLKR 53
AGLRR 52
STLAR 52

HTLKR 51
NTLRR 51
TTLAR 51
ATLTR 49
GDLVR 49
SDLKR 49
TTLVR 49
SGLAR 48
STLDR 48
STLMR 48
TGLNR 48
ETLKR 47
SVLRR 47
STLHR 46
SDLRR 45
SVLKR 44
HGLTR 43
DSLRR 42
TDLRR 42
HGLAR 40
HTLMR 40
SGLTR 40
STLLR 40
TTLLR 40
NGLKR 39
DTLTR 38
EGLNR 37
GTLNR 37
TMLKR 37
TSLRR 37
ATLMR 36
DALTR 36
STLQR 36
STLTR 36
GDLTR 35
GGLSR 35
HGLIR 35
ASLKR 34
SSLTR 34
AALKR 33
DALIR 33
TTLIR 33
EGLTR 32

HTLGR	32	TTLKR	22	QDLRR	14
HTLQR	32	EVLRR	21	TGLHR	14
QSLKR	32	GDLSR	21	TSLTR	14
DSLKR	31	HVLRR	21	AMLKR	13
SALKR	31	NTLHR	21	ATLGR	13
SSLAR	31	TDLTR	21	GDLQR	13
TTLMR	31	TTLNR	21	GGLQR	13
ATLER	30	DGLKK	20	GSLKR	13
NALKR	30	DMLRR	20	NTLVR	13
SGLIR	30	GTLHR	20	SILRR	13
SMLRR	30	MTLKR	20	SNLVR	13
GVLKR	29	TSLAR	20	SSLSR	13
TGLTR	29	ATLHR	19	VTLKR	13
EGLAR	28	DSLVR	19	ATLLR	12
HTLER	28	GTLDR	19	DSLQR	12
TS�KR	28	HTLVR	19	GGLNR	12
GDLHR	27	NTLLR	19	HGLSR	12
HNLVR	27	QTLKR	19	SGLLR	12
HTLRV	27	ESLRR	18	SQLKR	12
AGLAR	26	HDLQR	18	TGLMR	12
GDLNR	26	HGLNR	18	TSLGR	12
HGLLR	26	HMLRR	18	AGLQR	11
SILKR	26	TVLKR	18	DILRR	11
TTLGR	26	ATLNR	17	DTLER	11
DGLIR	25	DGLGR	17	DTLLR	11
SVLAR	25	NGLRR	17	GTLER	11
TTLQR	25	SGLNR	17	HTLLR	11
AGLTR	24	SGLQR	17	NGLAR	11
AGLVR	24	STLIR	16	NGLHR	11
DALSR	24	TALKR	16	SDLTR	11
DTLIR	24	TGLIR	16	SMLKR	11
EGLVR	24	TTLSR	16	TTLKV	11
QGLKR	24	DILKR	15	ATLIR	10
HDLRR	23	ETLNR	15	GMLKR	10
HVLKR	23	GSLVR	15	GSLRR	10
NTLTR	23	MDLKR	15	GTLLR	10
SDLAR	23	NGLQR	15	QGLRR	10
SSLKR	23	QTLAR	15	AGLHR	9
TDLAR	23	SSLIR	15	AGLSR	9
DALNR	22	AALRR	14	DALHR	9
HTLIR	22	GSLAR	14	DGLLR	9
NTLKR	22	HGLQR	14	DTLGR	9
QDLKR	22	NGLTR	14	GDLAR	9

HALKR	9	GILKR	6	GELTR	4
HGLHR	9	GVLRR	6	GNLTR	4
HTLKV	9	HALAR	6	GSLTR	4
HVLVR	9	HALRR	6	GTLGR	4
NMLKR	9	HLLKR	6	HALHR	4
NTLIR	9	HNLKR	6	HDLNR	4
QDLAR	9	NALHR	6	HNLLR	4
TGLGR	9	QDLQR	6	MTLRR	4
ASLRR	8	QGLVR	6	NGLMR	4
ATLQR	8	SALAR	6	NTLER	4
ETLAR	8	SDLGR	6	SALNR	4
GALTR	8	SGLHR	6	SDLQR	4
GALVR	8	STLKT	6	SELKR	4
HALVR	8	TDLSR	6	SGLER	4
HGLMR	8	TGLER	6	SSLGR	4
HILKR	8	DGLER	5	STLSM	4
HMLVR	8	DILVR	5	TDLMR	4
HSLRR	8	DTLKR	5	TNLRR	4
NTLSR	8	EALKR	5	TQLKR	4
QDLIR	8	EILKR	5	TSLLR	4
SDLVR	8	GVLVG	5	TSLMR	4
SGLMR	8	HSLTR	5	TSLVR	4
SMLTR	8	HTLDR	5	TTLER	4
TSLKV	8	NALAR	5	TVLRR	4
TTLTR	8	NGLIR	5	AALNR	3
AALTR	7	NGLSR	5	ADLTR	3
ETLQR	7	QQLQR	5	AGLER	3
GALKR	7	SALRR	5	ANLAR	3
GALRR	7	SALVR	5	ANLRR	3
GGLIR	7	SGLDR	5	ATLDR	3
GNLVR	7	SGLGR	5	DALMR	3
HDLGR	7	SNLLR	5	DILTR	3
HDLTR	7	SVLLR	5	DMLQR	3
HSLHR	7	TDLGR	5	DNLAR	3
HTLHR	7	TDLQR	5	DNLKR	3
NVLKR	7	VGLKR	5	DTLSR	3
TALAR	7	AGLIR	4	DVLVR	3
TDLKR	7	AILKR	4	EALRR	3
TGLKV	7	ALLKR	4	EGLSR	3
TSLNR	7	DSLAR	4	ETLRK	3
TSLQR	7	DSLRS	4	GNLER	3
VDLKR	7	DTLRK	4	GTLMR	3
DVLKK	6	ETLTR	4	GTLNR	3

HAEG...	3	HGLGR	2	ADLRV	1
HDLMR	3	HGLKM	2	AGLAW	1
HMLQR	3	HILMR	2	AGLKK	1
HNLTR	3	HNLHR	2	AGLLR	1
HSLKR	3	HNLSR	2	AILRR	1
HTLTR	3	HSLIR	2	ANLNR	1
HVLAR	3	HSLQR	2	ASLKS	1
NALSR	3	HVLHR	2	ASLQR	1
NNLVR	3	IDLKR	2	ASLTR	1
QSLNR	3	NGLLR	2	ASMKR	1
SILAR	3	NILVR	2	ATPVP	1
SMLHR	3	PALKR	2	AVLRR	1
STLRV	3	PGLAR	2	AVLTR	1
SVLTR	3	PTLMR	2	CGLRR	1
SVLVR	3	QGLTR	2	DAEA...	1
TALRR	3	QMLKR	2	DALER	1
TSLGG	3	QSLRR	2	DALKR	1
VTLGR	3	QTLNR	2	DALLR	1
AALAR	2	SALIR	2	DALPR	1
AALLR	2	SALSR	2	DARRR	1
AALVR	2	SELRR	2	DDLNR	1
ADLVR	2	SNLDR	2	DGAEE...	1
AGLNR	2	SNLRR	2	DGLKV	1
ANLVR	2	SNLSR	2	DGLWR	1
ASLAR	2	SQLRR	2	DGPAR	1
ASLIR	2	STLKR	2	DGPKK	1
ASLVR	2	STLRK	2	DGRRR	1
DALGR	2	TDLHR	2	DGVRR	1
DGLDR	2	TDLVR	2	DMLTR	1
DGLQR	2	TGLKL	2	DNLLR	1
DGLRK	2	TGLRV	2	DNLSR	1
DNLVR	2	TMLNR	2	DSLNR	1
DVLMR	2	TNLAR	2	DTLDR	1
DVLQR	2	TNLKR	2	DTLRV	1
EGLGR	2	TNLVR	2	DVLRK	1
EGLHR	2	TSLIR	2	DVLRN	1
EGLQR	2	TTLKA	2	DVLSR	1
ESLKR	2	VDLRR	2	DVQKR	1
GGLMR	2	VTLAR	2	EALVR	1
GNLAR	2	AALHR	1	EGLIR	1
GNLMR	2	AALQR	1	EGLKM	1
HDLSR	2	AAPER	1	EGLLR	1
HGLDR	2	ADLNR	1	EGLMR	1

EGLQC	1	HDLKR	1	QDLVR	1
EGLRS	1	HELTR	1	QGLAR	1
EGLRV	1	HGLRW	1	QNLHR	1
EGRRR	1	HGMRR	1	QNLQR	1
EGVRR	1	HILIR	1	QNLRR	1
EGWS...	1	HLLNR	1	QNLRW	1
ENLAR	1	HNLAR	1	QPACV	1
ENLSR	1	HPAP...	1	QTLHR	1
ESLAR	1	HQLIR	1	QTLQR	1
ETGWG...	1	HSLAR	1	QTLTR	1
ETLER	1	HSLSR	1	RGLKR	1
ETLHR	1	HTLNK	1	RPAA...	1
ETLVR	1	HTLRA	1	RTLKV	1
ETRRR	1	HTLRG	1	SALHR	1
EVLKR	1	HTLSR	1	SALKV	1
GAEE...	1	HVLER	1	SALMC	1
GALAR	1	HWLLR	1	SALMR	1
GALNR	1	IGLRR	1	SDLKS	1
GDLYR	1	IGLTG	1	SILKV	1
GDPAP...	1	INLTR	1	SILNR	1
GGLDR	1	ITLTR	1	SILVR	1
GGLGR	1	KGLPG	1	SMLAR	1
GGLQE	1	MDVKG	1	SMLLR	1
GGQTR	1	MTLIR	1	SMLR	1
GGVVR	1	NALRR	1	SNLAR	1
GHLQR	1	NALVR	1	SNLHR	1
GILRR	1	NGLER	1	SNLQR	1
GMLRR	1	NHLVQ	1	SPLHR	1
GNLDR	1	NMLAR	1	SSLKW	1
GNLLL	1	NNLAR	1	STPER	1
GNLLR	1	NNLLR	1	STQVR	1
GNLQR	1	NSLAR	1	SVLQR	1
GNLVM	1	NTLNR	1	SVLSR	1
GTLLV	1	NTLQR	1	TALNR	1
GTLRV	1	P*MGS	1	TALVR	1
GTLRW	1	PALKP	1	TGLDR	1
GTPHR	1	PGWAG	1	TGLKW	1
GVLAR	1	PTLKR	1	TGLNV	1
GVLNR	1	PTLRR	1	TGLQC	1
GVLVR	1	PWS...	1	TGLRQ	1
GWLSR	1	QALKR	1	TGPAR	1
HAEA...	1	QALTR	1	TGPNR	1
HALKV	1	QDLAT	1	TGQRR	1

TMLRR	1
TNLMR	1
TNLNR	1
TRLVR	1
TSLIS	1
TTLDR	1
TTLKK	1
TTLRT	1
TTLRV	1
TVLRM	1
VALAR	1
VALRR	1
VGLHR	1
VGLNR	1
VGLQR	1
VGLRR	1
VGLTM	1
VNLKR	1
YGLAR	1
YGLVR	1
YILRR	1

**Table 18. ZF5
selection on G:T
change at nt 8 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	Read #				
GGLVR	178	TALRR	69	AGLVR	29
AALRR	174	GALAR	66	ASLKR	29
GALVR	170	TGLKR	66	GTLAR	29
TGLRR	162	GGLTR	63	QALVR	29
SGLRR	158	SALVR	61	SGLIR	29
GGLRR	152	ASLRR	57	STLSR	29
EGLRR	148	DALRR	57	TGLVR	29
AGLRR	143	DTLRR	52	DGLAR	28
GALRR	126	GSLKR	50	DSLKR	28
HTLRR	125	HALHR	46	STLTR	28
GGLKR	117	HSLVR	46	AALSR	27
AALTR	111	AALAR	45	HALSR	27
AGLTR	108	SALKR	45	SALSR	27
GALKR	104	EALRR	44	STLMR	26
AGLAR	102	AALVR	42	HTLNR	25
HGLVR	101	EALKR	42	DGLTR	24
HGLKR	99	HSLRR	42	HGLIR	24
HALAR	97	SALAR	42	NLTKR	24
AGLKR	96	SALTR	42	TSLRR	24
HGLRR	96	SSLRR	41	VTLRR	24
AALKR	94	ATLAR	40	AGLNR	23
SGLAR	93	HALKR	40	GALSR	23
ATLRR	90	NLRR	40	SSLTR	22
ETLRR	90	SALNR	40	AGLHR	21
HGLAR	90	AGLQR	39	GALNR	21
EGLKR	83	DGLVR	38	HTLGR	21
SGLTR	82	ETLKR	38	QTLRR	21
HGLTR	79	QALRR	38	TGLTR	21
DGLRR	77	TALAR	36	AALQR	20
GALTR	75	TTLAR	36	GTLTR	20
GGLAR	74	TTLRR	36	HGLMR	20
HALRR	74	HALVR	35	TALKR	20
SALRR	71	HTLAR	35	ASLQR	19
SGLVR	70	GGLSR	34	ATLVR	19
		AALNR	33	DALKR	19
		GGLQR	33	ESLRR	19
		SSLVR	33	GGLNR	19
		DTLNR	31	ATLTR	18
		GSLRR	31	GGLIR	18
		HTLMR	30	GSLTR	18
		SGLKR	30	GTLKR	18
		STLAR	30	GTLQR	18
		AGLSR	29	NGLKR	18

SDLKR	18	SGLMR	10	TTLKR	6
STLVR	18	STLHR	10	AALER	5
GTLRR	17	STLRR	10	AALGR	5
SGLLR	17	AALHR	9	ATLGR	5
AGLIR	16	ATLIR	9	EGLNR	5
ASLVR	16	ATLNR	9	GGLQE	5
GSLAR	16	DTLAR	9	GNLTR	5
HTLHR	16	DTLKR	9	NALKR	5
SGLNR	16	HGLQR	9	NTLAR	5
SGLSR	16	HSLSR	9	QTLKR	5
DALVR	15	PALKR	9	QTLQR	5
GSLVR	15	SSLIR	9	TGLMR	5
HGLHR	15	TALTR	9	TGLNR	5
ATLER	14	TGLHR	9	TSLAR	5
HSLQR	14	ATLQR	8	DGLQR	4
SGLQR	14	HTLIR	8	DTLMR	4
TGLIR	14	HTLTR	8	EGLVR	4
TGLSR	14	NTLQR	8	GGLGR	4
TTLGR	14	QGLKR	8	GGLMR	4
TTLVR	14	SVLKR	8	GSLIR	4
VTLKR	14	TSLNR	8	GSLNR	4
ATLSR	13	AALIR	7	HALLR	4
DSLRR	13	AALMR	7	HTLKR	4
SALIR	13	DALTR	7	HTLVK	4
TALVR	13	DTLQR	7	NHLVQ	4
TTLLR	13	GSLHR	7	QGLRR	4
AALLR	12	GSLQR	7	QVLKR	4
HALTR	12	HGLSR	7	SALHR	4
QGLVR	12	NGLVR	7	SSLAR	4
TGLAR	12	SSLKR	7	STLGR	4
ATLLR	11	STLER	7	STLIR	4
DGLNR	11	TALNR	7	STLLR	4
GTLVR	11	ATLKR	6	TALQR	4
QALKR	11	AVLKR	6	TSLMR	4
SALLR	11	ETLAR	6	AALDR	3
SSLHR	11	HGLLR	6	AGLLR	3
STLNR	11	HTLVR	6	ANLAR	3
TGLQR	11	HVLKR	6	ANLRR	3
TTLQR	11	NALQR	6	ASLSR	3
ASLTR	10	NGLRR	6	ATLHR	3
DGLKR	10	NSLVR	6	DALAR	3
NSLRR	10	SGLER	6	DALQR	3
QSLKR	10	SGLHR	6	DGLIR	3

DGLSR	3	DNLRR	2	TNLVR	2
DMLRR	3	DNLVR	2	TSLHR	2
EGLAR	3	DSLAR	2	TTLNR	2
EGLTR	3	DTLLR	2	TTLTR	2
ESLKR	3	DTLSR	2	AALTG	1
GGLER	3	DTLTR	2	AAPER	1
GGLLR	3	DVLKR	2	ADLKR	1
GNLAR	3	DVLRR	2	AGLAW	1
GNLVR	3	EALTR	2	AGVIR	1
GSLSR	3	ETLQR	2	AGVTR	1
HALQR	3	GALDR	2	AMLKR	1
HNLTR	3	GDLKR	2	ANLNR	1
HSLKR	3	GDLTR	2	ASLAR	1
HTLQR	3	GGPRR	2	ASLRG	1
ITLKR	3	GTLER	2	ATLMR	1
MTLKR	3	HALNR	2	ATLRM	1
NSLAR	3	HDLRR	2	ATPRR	1
NSLSR	3	HGLDR	2	AVLAR	1
NTLNR	3	HGLGR	2	AVLRR	1
QALTR	3	HGLNR	2	AVLVR	1
QGLTR	3	HNLLR	2	DALNR	1
QTLHR	3	HSLHR	2	DGLGR	1
QTLVR	3	MPLTR	2	DGLKK	1
SALMR	3	NALHR	2	DGLRK	1
SNLLR	3	NGLAR	2	DGLWR	1
SSLGR	3	NGLIR	2	DGPAA...	1
STLDR	3	NGLLR	2	DMLKR	1
SVLRR	3	NGLQR	2	DSLQR	1
TGLLR	3	NGLTR	2	DTLVR	1
TNLMR	3	NTLHR	2	DVLKK	1
TNLRR	3	NTLIR	2	EALHR	1
TSLER	3	NVLKR	2	EALSR	1
TTLMR	3	QDLIR	2	EGLHR	1
VGLKR	3	QDLKR	2	EGLMR	1
AGLDR	2	QSLRR	2	EGLRG	1
AGLER	2	SILKR	2	EGLRV	1
AGLGR	2	SMLRR	2	ENLAR	1
ASLMR	2	SSLQR	2	ENLSR	1
ASLNR	2	STLQR	2	GAEE...	1
DGLDR	2	TALER	2	GALER	1
DGLHR	2	TALGR	2	GALGK	1
DNLKR	2	TGLRV	2	GALIR	1
DNLLR	2	TMLKR	2	GALKV	1

GALMR	1	INLTR	1	STLRK	1
GALQR	1	KGLPG	1	STPSR	1
GAPRR	1	MPLQR	1	SVLTR	1
GDLNR	1	NALAR	1	TALLR	1
GDLVR	1	NGLHR	1	TAPTR	1
GGLDR	1	NGLMR	1	TDLAR	1
GGLHR	1	NGLSR	1	TDLRR	1
GGPAR	1	NMLKR	1	TGLIK	1
GGPVR	1	NNLAR	1	TGPAR	1
GGQVR	1	NNLLR	1	TGQRR	1
GGVAR	1	NTLTR	1	TNLKR	1
GGWP...	1	PWS...	1	TNLNR	1
GMLAR	1	QAPWP...	1	TSLKR	1
GNLER	1	QDLRR	1	TSLQR	1
GNLMR	1	QGLAR	1	TTLIR	1
GNLQR	1	QMLKR	1	TTLKV	1
GSLRV	1	QNLHR	1	TTLNR	1
GTLGR	1	QNLQR	1	TVLRM	1
GTLHR	1	QNLRW	1	TWLRR	1
GTLMR	1	QTLSR	1	VALAR	1
GTLNR	1	RGLQR	1	VALQR	1
GTLSR	1	SALER	1	VGLQR	1
HAAQ...	1	SALGR	1	VGLTM	1
HALAS	1	SALMC	1	VNLKR	1
HALER	1	SALQR	1	VSLKR	1
HALVH	1	SAQR...	1	VSLRR	1
HAMRR	1	SARVR	1	VTLAR	1
HAQHR	1	SDLAR	1	VTLGR	1
HGLTL	1	SDLNR	1		
HGLVM	1	SDLQR	1		
HLLKR	1	SDLRR	1		
HNLAR	1	SELRR	1		
HNLHR	1	SGADA...	1		
HNLKR	1	SGLR...	1		
HNLSR	1	SGLVC	1		
HNLVR	1	SGPDP...	1		
HPAP...	1	SMLAR	1		
HQLIR	1	SNLDR	1		
HSLGR	1	SNLQR	1		
HTLRV	1	SNLSR	1		
HVLAR	1	SQLRR	1		
IGLRR	1	SSLPR	1		
IGLTG	1	STLKT	1		

**Table 19. ZF4
selection on G:T
change at nt 10 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	Read #				
AHLRK	4967	TKLKV	108	AKLRT	14
GHLKK	1446	THLRT	107	RHLRV	14
THLRA	1429	GHLRR	102	AALRK	11
EHLRR	1293	*R...	92	AHLHK	11
GHLRK	1082	SHLRL	90	GHLTK	11
HHLTK	876	SKLRV	80	QHLRR	11
AKLRI	867	GALV...	79	RSHS...	11
AKLRV	641	GHLKM	75	SHLNK	11
AKLRL	625	SKLRI	74	AHLQK	10
AKLKI	599	GILS...	71	GHLMK	10
SHLRK	566	SK*VL	63	SKLRT	10
AHLKK	560	SKLVL	62	AHLKV	9
THLKK	496	TR...	61	AHLRA	9
TKLRL	486	IRLGV	59	AHLRT	9
SKLRL	475	MALGL	58	EHLRL	9
SKLKV	466	EHLRK	54	GHLKI	9
TKLKL	466	GHLRM	54	SHLKL	9
QHLRK	457	EHLKR	50	EHLKK	8
AKLKL	443	ITLM...	48	GHLRT	8
GHLVK	419	AHLVK	40	GKLKM	8
GKLKI	302	THLRL	40	HHLKK	8
THLRK	268	GKLKV	38	SKLTI	8
AKLKV	258	GHLKL	34	THEKP...	8
GKLRI	246	AHLRR	32	*G...	7
GKLRL	224	GHLIK	30	AKLIL	7
GHLRL	213	EHLVR	28	AKLTI	7
TKLKI	199	GKL RV	27	HALAA	7
RSLGL	178	TALSM	26	TKLQV	7
AHLRV	177	EHLQR	25	AKLRM	6
AHLRL	153	EKLKV	25	EHLRI	6
TKLRV	152	QHLVK	25	GHLAK	6
SKLKI	146	TKLNL	25	GHLKR	6
SHLVG	132	GHLRA	23	GKLTL	6
GKLKL	116	GRLPK	21	SHLKK	6
		NGR...	21	SHLRR	6
		SKLKL	21	AILKA	5
		THLTK	21	AKLRK	5
		RLLSG	20	AKLTL	5
		TKLRI	19	ASLTG	5
		AHLRI	18	EHLRV	5
		GHLKV	16	EVLTM	5
		GHLRV	16	GHLKT	5
		GLLPG	16	NGRS...	5

THLRR	5	AMLMQ	2	TKLMV	2
AHLKL	4	AQLRI	2	TPLNI	2
GALVH	4	DALR...	2	TRLQK	2
GKLV	4	EHLRA	2	TSLTR	2
NGRSPV...	4	EHLRT	2	VGLGQ	2
QALSI	4	EKLKL	2	VHLRK	2
SHLRT	4	GGLQK	2	AALES	1
TRS...	4	GGLTM	2	AALRI	1
AALRL	3	GH*R...	2	ADLRK	1
AHLMK	3	GHLLR	2	AELLG	1
AHLRE	3	GHLRI	2	AELRI	1
AHLRQ	3	GHLVG	2	AGLAA	1
AKLNL	3	GHLVR	2	AGLKR	1
AKLRA	3	GKLNL	2	AGLMD	1
APLRK	3	GKLRR	2	AHLGL	1
EKLRI	3	GKLVP	2	AHLK...	1
GALMG	3	GLLGL	2	AHLKA	1
GALTG	3	GNLGM	2	AHLKI	1
GHLRG	3	GVLQK	2	AHLKT	1
GHLTL	3	HGLLP	2	AHLNK	1
GKLRR	3	HHLRV	2	AHLR...	1
GKLTV	3	HLLN	2	AHLSK	1
GKLV	3	IGLQR	2	AHLSP	1
HHLRK	3	KTLGV	2	AHLTV	1
MGLVG	3	LSLLK	2	AHLWK	1
SHLKV	3	MRLGE	2	AKFKI	1
SHLRI	3	NSLTR	2	AKIKH	1
SKLIL	3	NVLNK	2	AKIRI	1
SKLMV	3	PHLRK	2	AKIRL	1
SLLAG	3	PLLMP	2	AKIRV	1
THLKI	3	PRLRH	2	AKLHT	1
THLQK	3	QKLHL	2	AKLKE	1
VPLAG	3	QKLNL	2	AKLKG	1
AGLLG	2	SHLRV	2	AKLKM	1
AHLKM	2	SKLHL	2	AKLMN	1
AHLRN	2	SKLKR	2	AKLNI	1
AHLTK	2	SKLNL	2	AKLQL	1
AKLIV	2	SPLAE	2	AKLRG	1
AKLKA	2	SVLML	2	AKLRR	1
AKLKK	2	TH*R...	2	AKLSM	1
AKLTV	2	THLKL	2	AKSRV	1
AKLV	2	THLRV	2	AKVKL	1
AKSRI	2	TKLIL	2	AKVRI	1

ALLMA	1	EHLKM	1	GLLMG	1
ALLRR	1	EHLRQ	1	GLLRG	1
AMLIM	1	EHLRS	1	GMLGG	1
AMLKI	1	EHLSE	1	GPLGV	1
AMLRG	1	EHLSR	1	GPLRV	1
AMLR	1	EHLTK	1	GRLKI	1
ANLSN	1	EHLVK	1	GRLKK	1
ANVAQ	1	EQLGP	1	GSLST	1
APLKK	1	ERLAA	1	GSLVK	1
AQFRK	1	ERLGR	1	GTLKR	1
AQLVD	1	ERLRR	1	GVLG	1
ARLAG	1	ESLMA	1	GVLLV	1
ARLGT	1	ETLSH	1	GVLS...	1
ARLRA	1	EVLGI	1	GYLRK	1
ARLRK	1	FFLRV	1	HALRT	1
ASLRM	1	GALGR	1	HALVN	1
ATLKL	1	GALIM	1	HGLTG	1
ATLRV	1	GDLSG	1	HHLAK	1
C*LKI	1	GGLDL	1	HHLRR	1
DELMR	1	GGLDQ	1	HIRS...	1
DELRV	1	GGLKV	1	HTHEK	1
DGLES	1	GGLNM	1	IELVQ	1
DGLLR	1	GGLPE	1	IGLGL	1
DGLMD	1	GGLVV	1	IKLRL	1
DGLVG	1	GHFKT	1	IMLRE	1
DHLKK	1	GHFQN	1	IMLVE	1
DHLRK	1	GHLK...	1	IPLGD	1
DHLRR	1	GHLMN	1	IQLRK	1
DKLRK	1	GHLMV	1	IRLG...	1
DLLGV	1	GHLQR	1	IRLGG	1
DLLL	1	GHLR...	1	IRLVV	1
DNLRE	1	GILAG	1	IVLAA	1
DPLAR	1	GKLHE	1	KHLRA	1
DSLGE	1	GKLKA	1	KHLRL	1
EALMA	1	GKLKF	1	KILPE	1
EDLVK	1	GKLKT	1	KKLLE	1
EELGL	1	GKLR...	1	KMLPP	1
EELMM	1	GKLRA	1	KNLIK	1
EGLRG	1	GKLRM	1	KSLMP	1
EGLVE	1	GKLVA	1	LALGG	1
EHLG...	1	GKLVV	1	LGLGA	1
EHLHK	1	GLLGE	1	LGLVG	1
EHLKL	1	GLLLD	1	LHLTK	1

LQ...	1	QELGR	1	SHLRE	1
LRLIG	1	QGLPV	1	SHLRG	1
LTE...	1	QHLKK	1	SHLTK	1
LTLQR	1	QHLQR	1	SHLTM	1
LVLRR	1	QHLR...	1	SHLV...	1
MA*SHMK	1	QHLRI	1	SHLVK	1
MALRL	1	QHLRL	1	SKIRL	1
MALTR	1	QHGTK	1	SKLEG	1
MGLDP	1	QILLH	1	SKLGA	1
MGLGE	1	QKLRI	1	SKLKG	1
MGLQN	1	QNLHK	1	SKLRM	1
MHLRM	1	QPLIK	1	SKLRN	1
MKLEQ	1	QQVTA...	1	SKLRR	1
MLLRN	1	QTLAE	1	SLLEE	1
MLLSH	1	QVTLA	1	SLLGT	1
MLLVN	1	RALSA	1	SLLNG	1
MPLRA	1	RGL...	1	SQLKV	1
MQLGG	1	RGLGA	1	SQLLE	1
MRLAR	1	RGLTA	1	SRLMA	1
MRLMG	1	RGLVR	1	STLLM	1
MRLVG	1	RGLVV	1	STLVG	1
MSLER	1	RHLRA	1	TALRG	1
MTLPL	1	RHLRE	1	TG...	1
MTLSD	1	RHLRM	1	TGLAR	1
MVLAG	1	RHLRR	1	TGLGL	1
NG...	1	RILPR	1	TGLLK	1
NGLMR	1	RKLIV	1	TGLRV	1
NGLQR	1	RKLKL	1	TGLVD	1
NKLRL	1	RLLGA	1	THEKP	1
NLAH	1	RLLMP	1	THFRT	1
NLLPT	1	RLLRR	1	THIR...	1
NRLES	1	RMLVP	1	THLAR	1
NRLGG	1	RRLEG	1	THLKQ	1
NTLPK	1	RRLVN	1	THLLK	1
PGLHG	1	RTLML	1	THLMK	1
PGLRA	1	RTLQ	1	THLRP	1
PHFTK	1	SDLHV	1	THLVK	1
PILLQ	1	SDLRK	1	THMK	1
PKLGL	1	SGLGR	1	THVKK	1
PLLS	1	SGLLV	1	TKLKM	1
PQLTG	1	SGLVR	1	TKLKR	1
PREAM	1	SHLKM	1	TKLNM	1
PTLQR	1	SHLRA	1	TKLRK	1

TKLRP	1
TKLS...	1
TKLTI	1
TMLGG	1
TMLKL	1
TMLPG	1
TPLKR	1
TPLRA	1
TQLKK	1
TQLKL	1
TQLKV	1
TR*RL	1
TRLKL	1
TRLRE	1
TS...	1
TTLGI	1
TYLKK	1
VELDP	1
VELVN	1
VKLQQ	1
VKLRL	1
VKL RN	1
VKL RV	1
VLLKS	1
VLLQM	1
VMLKD	1
VMLMG	1
VPLAL	1
VPLER	1
VPLNT	1
VPLSS	1
VPLVP	1
VQ*G...	1
VRLEE	1
VRLQA	1
VVTA...	1
WHLKK	1
YG...	1

Table 20. ZF4

selection on G:C

change at nt 10 of

core motif in CBS.

Sequences reflect

position 2 to 6.

Sequence	Read #	GKLRI	8	ATLRV	3
AKLRV	5924	HHLTK	8	DGLES	3
TKLRV	4888	THEKP...	8	GGLKG	3
TKLKV	3542	TK*RG	8	GHLRG	3
SKLKV	3056	TKLRT	8	GHLRK	3
AKLKV	2451	AKLKL	7	GKLIV	3
SKLRV	1583	AKLRE	7	GKLKG	3
SKLRT	474	AKLTV	7	GKLKI	3
GKL RV	320	EKLKV	7	GKLVP	3
AKLRI	254	SKLKE	7	KKLHW...	3
GKLKV	237	TKLRG	7	QALSI	3
AKLKI	164	AKLRG	6	RTLS...	3
AGLKR	132	EHLRR	6	SKLRA	3
TKLKI	129	EKL RV	6	SKVRV	3
GGLKV	78	RALW...	6	SLLAG	3
SKLKI	76	SKLKA	6	TK*SV	3
SKLRI	76	AKLKG	5	TKLAV	3
ARLGT	64	AKLRM	5	TKLRE	3
VALGS	48	EVLTM	5	TKSRV	3
GVLQK	46	QHLRK	5	TKVKV	3
TRS...	39	SKLVV	5	VMLMM	3
AHLRK	30	STLVV	5	VPLAG	3
AKLVV	26	TKLKE	5	AGLLG	2
TKLRA	24	TKLNV	5	AILQV	2
LGLRG	18	GHLVK	4	AKIRV	2
MRLAR	15	GKLRL	4	AKLIV	2
TKLKA	14	KDALQYESECG...	4	AKLKE	2
SKLKG	13	LSLVD	4	AKLMV	2
TKLRI	13	QKLKV	4	AKLSV	2
AKLRT	12	RELKE...	4	AKVKV	2
NGR...	12	RILGS	4	ANLSN	2
PNLAV	12	THLKK	4	DKLRV	2
GGLEV	10	THLRA	4	ETLMH	2
GHLKK	10	TKIRV	4	GALMG	2
PREAI	10	TKLRL	4	GGLQK	2
		TKLRM	4	GGLTM	2

GHLKM	2	AGLKH	1	DGLTG	1
GKSKV	2	AGLRR	1	DGLVG	1
GLLGE	2	AGLSR	1	DHLKR	1
GLLGL	2	AGLTR	1	DHLNV	1
GNLGM	2	AGLVR	1	DLLGV	1
GTLKR	2	AHLKA	1	DLLLN	1
HLEEN	2	AHLKV	1	DNLRE	1
IGLQR	2	AHLRV	1	DPLAR	1
IKLRV	2	AHLSP	1	DRITI	1
KALHT	2	AKIRE	1	DVLKG	1
KGLMM	2	AKLAV	1	DVLRG	1
MELAE	2	AKLGV	1	EALVH	1
MGLVG	2	AKLMI	1	EELMM	1
MRLGE	2	AKLNV	1	EGLRG	1
MTLPL	2	AKLRF	1	EGLVE	1
NGLMR	2	AKLRN	1	EHLRV	1
NGRS...	2	AKLTI	1	EHLVR	1
NKLKV	2	AKLWV	1	EQLGP	1
PRLLA	2	AKRRV	1	EQLMT	1
PRLRH	2	AKSKV	1	ERLGR	1
PRLSR	2	AKVRG	1	ESLMA	1
QGLEA	2	ALLKV	1	ETLSH	1
SELKV	2	AMLIM	1	EVLAA	1
SKLKR	2	AMLKV	1	EVLGI	1
SKLRE	2	AMLMQ	1	G...	1
SKLRG	2	AMLRG	1	GDLSG	1
TH*R...	2	AQLKV	1	GGLDQ	1
TKIKV	2	AQLRV	1	GGLKD	1
TKLKL	2	AQLVD	1	GGLKI	1
TKLMV	2	ARLKV	1	GGLNR	1
TKLQV	2	ARLRI	1	GGLPE	1
TKLRD	2	ARLRM	1	GH*R...	1
TKLSV	2	ARLRV	1	GHLKR	1
TPLNI	2	ATLQV	1	GHLLR	1
TRLQK	2	AVLKV	1	GHLMV	1
TSLTR	2	AYPRE	1	GHLRR	1
TTLRV	2	CGLHW...	1	GHLRV	1
V	2	CKLRV	1	GHLTL	1
AALES	1	DALDR	1	GHLVG	1
AELKA	1	DELRV	1	GILAG	1
AELKV	1	DGLKV	1	GILRM	1
AELLG	1	DGLLR	1	GK*RG	1
AGLAA	1	DGLMD	1	GKLKA	1

GKLKM	1	KNLIK	1	NTPK	1
GKLML	1	KSLMP	1	NV...	1
GKLQV	1	KTLGV	1	NVLGG	1
GKLRA	1	LALGG	1	NVLNK	1
GKLRQ	1	LGLGA	1	PGLAA	1
GKLRT	1	LGLVG	1	PGLHG	1
GKLTl	1	LQ...	1	PKLGL	1
GLLLD	1	LRlIG	1	PKLRA	1
GLLMG	1	LSLDG	1	PLLKS	1
GLLPg	1	LTLQR	1	PLLMP	1
GLLRG	1	LVLRR	1	PNLAG	1
GPLGQ	1	MA...	1	PNYW...	1
GPLGV	1	MALGL	1	PQLTG	1
GPLMG	1	MALTR	1	PREAM	1
GQLKA	1	MELDR	1	PTLQR	1
GRLAV	1	MGLDP	1	PVLDH	1
GRLNA	1	MGLGE	1	Q	1
GSLST	1	MGLQN	1	QALTN	1
GSLVK	1	MKLEQ	1	QELGR	1
GVLAg	1	MKLQA	1	QGLPV	1
GVLlV	1	MKLRV	1	QILLH	1
HALRT	1	MLLRN	1	QKLNl	1
HGLLP	1	MLLVN	1	QNLHK	1
HIRS...	1	MPLLA	1	QPLIK	1
HPLTV	1	MPLRA	1	QTLAE	1
HRLTR	1	MRLARHIRSHTGERP...	1	QVLRK	1
IELVQ	1	MRLMG	1	RALSA	1
IGLGL	1	MSLER	1	RELVR	1
IKLKV	1	MSLVN	1	RGL...	1
IMLKS	1	MTLSD	1	RGLDM	1
IMLVE	1	MVLAG	1	RGLDR	1
IQSGE	1	MVLQE	1	RGLTA	1
IQVTLA	1	MVLVG	1	RGLVA	1
IRLAL	1	N...	1	RGLVR	1
IRLG...	1	NDALEYESECgP...	1	RGLVV	1
IRLGV	1	NDALQYESVCVP...	1	RHLRE	1
ITLM...	1	NGLQR	1	RILPR	1
IVLAA	1	NGLVV	1	RKLIV	1
KALRG	1	NK*NV	1	RKLKA	1
KKLLE	1	NKLRV	1	RKLKV	1
KKLRE	1	NLAH	1	RKLRE	1
KKLVR	1	NLLPT	1	RKLRV	1
KMLPP	1	NRLGG	1	RKVRV	1

RLLGA	1	TGLGL	1	VKLQQ	1
RLLMP	1	TGLRV	1	VKLRN	1
RMLQE	1	THEKP	1	VKLRV	1
RMLVP	1	THIR...	1	VLLKS	1
RPLEV	1	THLAR	1	VLLMA	1
RRLVN	1	THLQK	1	VLLPS	1
RTLML	1	THLRK	1	VMLKD	1
RTLQ	1	THLRL	1	VMLMG	1
S*G...	1	THLRT	1	VNLLE	1
SDLHV	1	TKLHV	1	VPLAL	1
SGLGR	1	TKLKD	1	VPLER	1
SGLLV	1	TKLKM	1	VPLNT	1
SGLVR	1	TKLKR	1	VPLSS	1
SHLKV	1	TKLKT	1	VPLVP	1
SHLRL	1	TKLMA	1	VQ*G...	1
SKFKV	1	TKLNM	1	VQLPV	1
SKFRV	1	TKLQI	1	VRLEE	1
SKIRT	1	TKLR...	1	VRLQA	1
SKLHL	1	TKLTV	1	VTLGR	1
SKLKD	1	TKLWV	1	YTHMK	1
SKLKL	1	TKSRD	1		
SKLKM	1	TMLGG	1		
SKLQI	1	TMLKV	1		
SKLRL	1	TMLPG	1		
SKLSV	1	TMLRV	1		
SKLTV	1	TPLKR	1		
SKLVL	1	TRLKV	1		
SKSRT	1	TRLRE	1		
SKVKV	1	TRLRG	1		
SKVRT	1	TRLRV	1		
SLLEE	1	TRSHS...	1		
SLLGT	1	TTIRV	1		
SNLKG	1	TTLGI	1		
SNLTH	1	TTLKV	1		
SQLE	1	TTLA	1		
SRLKV	1	TTLVP	1		
STLLM	1	TVLAP	1		
TALIS	1	TVLPM	1		
TALRG	1	VALTK	1		
TELIG	1	VELVN	1		
TELKV	1	VGLGQ	1		
TG*S...	1	VGLLR	1		
TGLAR	1	VKLLV	1		

Table 21. ZF4
selection on G:A
change at nt 10 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence	Read #	GGGGL
AKLRV	408	GHLTL
GKLRV	294	GPLHI
TRS	180	MLLVN
TKLKV	170	VELNS
GKLRL	166	AKLIT
NGRS	155	GKLTLL
TKLRV	124	SNLTH
AKLRM	109	AT*RR
TKLRL	109	DLLLN
GKLKV	107	GMLGG
AKLKI	92	RELVR
AKLKV	88	SPLLS
EKLRI	84	DKLRR
MSLER	68	GALIM
NGRSPVC	67	GLLG
GALMG	66	GLMM
AELIR	63	IHLAD
SGLGR	63	TLTQ
NVLGG	61	TRSHSS
RGLT	61	ALMQ
TLLMG	58	ARLRV
GKLVP	57	GHLRL
VPLAG	57	HLLN
QILLH	55	HTLNM
TLPL	55	PMLVD
*MLTS	54	SKLHL
EMLTS	53	GK*KL
SKLKV	53	AMLMMQ
IGLGL	52	DPLAR
TKLKI	52	GKLKL
AHLSP	51	NKLRV
AKLMV	51	TALRG
DALRG	51	AILQV
LGLGA	51	AKLIV
		AKLRL

51	HIRS	38
50	GKMKM	37
49	GLLGE	37
49	GNLGM	37
48	GILAG	36
48	GPLAL	36
48	MVLQE	36
47	PKLGL	36
47	ARLGL	35
46	EELLK	35
46	MLLRN	35
46	QPLIK	35
45	AHLAV	34
45	EVLTM	34
45	MRLGE	34
44	EALMA	33
44	PLLGV	33
44	PQLTG	33
44	EVLAA	32
44	HPLQQ	32
43	PGLAA	32
43	QKLNL	32
43	SKLNN	32
42	TRLRN	32
42	AKLVL	31
42	DLLV	31
42	DSLLA	31
41	GELRT	31
41	RLLGV	31
41	AALTR	30
41	GGLQK	30
41	NGLMR	30
41	NRLQ	30
41	PALGN	30
40	PLLGM	30
39	PPLMQ	30
39	TQLEE	30
39	VGLEG	30
39	DLLGV	29
39	GGLDL	29
38	GHLRG	29
38	KTLRE	29
38	PRLR	29

PSLGV	29	DGLAM	25	GHLLR	22
RR*PS	29	GGLPE	25	GILAK	22
TGLVD	29	GLPV	25	GMLPD	22
THLQK	29	KSLMP	25	GPLGV	22
DGLMDHIRSHTGERPF	28	NGLNR	25	GSLPM	22
LSLLK	28	SHMK	25	GSLVK	22
MVLVP	28	TPLNI	25	GVLAR	22
SELTG	28	TVLAP	25	LGLVG	22
SGLKH	28	VLLME	25	LTLQR	22
TPLKR	28	AGLLG	24	NGRSPVET	22
VGLG	28	GALPR	24	PGLRA	22
AHLRK	27	GKLIL	24	PMLRV	22
AKLMN	27	GKLML	24	TLML	22
AKLRI	27	GVLLV	24	AHLRV	21
DRLGP	27	KQLTD	24	ALLMA	21
GLLGR	27	LKLIG	24	ASLGQ	21
IMLRE	27	LRLIG	24	DGLTG	21
KQLQP	27	LRLMS	24	EGLRG	21
MA*S	27	NRLGG	24	EHLAV	21
NGR	27	PNYWP	24	ELILE	21
RHLRE	27	RHLVP	24	GH*RS	21
RPLLR	27	SRLGA	24	GHLAM	21
RSLRL	27	AKLRN	23	GKLRM	21
SKLRV	27	DRLAS	23	GLLP	21
SLLAG	27	DSLGE	23	GTLAI	21
TTLGI	27	ERLAA	23	IRLKK	21
AELLG	26	GKLRI	23	KELRR	21
AGLTR	26	GSLS	23	KILPE	21
DGLVG	26	HRLGG	23	LHLPI	21
GALG	26	MDLLL	23	MGLVG	21
GDLSP	26	MTLGA	23	MPLLA	21
GGLDQ	26	PPLER	23	NELRG	21
GHLMV	26	PVLPD	23	NVLNK	21
GKLKI	26	QELGR	23	PHLNG	21
GKLSL	26	RTLS	23	PLLMP	21
GVLAG	26	SLLQG	23	RLLGS	21
LRLNL	26	TGLRV	23	RTLIS	21
MTLGN	26	TKLMV	23	SC*AS	21
PMLAA	26	VPLER	23	SDLHV	21
SKLRT	26	APLGM	22	SKLRL	21
TKLNM	26	EHLRR	22	VKLMN	21
ALIG	25	GALVR	22	VTLIG	21
AQLAN	25	GGLNR	22	AGLQE	20

ALHT	20	VPLSS	19	KGLEG	17
DPLVD	20	VQLPT	19	QGLDN	17
E	20	AALRR	18	RALVS	17
EALDA	20	AGLD	18	RGLAT	17
GALAT	20	AMLIM	18	SKLMV	17
GSLS	20	AMLRG	18	SKLVV	17
GTLLM	20	DVLPG	18	SRLMA	17
IKLRP	20	EQLGP	18	TLHE	17
LQ	20	GHLKR	18	TRLRG	17
NGP	20	GILRM	18	ARLRI	16
QNLHK	20	GLLV	18	EHLRV	16
RRLLD	20	GLMN	18	GHLKS	16
SLLGT	20	GMLVG	18	GLLKH	16
TELIG	20	GPLTI	18	GRLNA	16
TGLMG	20	GRLE	18	GVLSI	16
TKLLL	20	GSLQS	18	GVLST	16
TTLGA	20	GVLVS	18	HALRT	16
VE*DP	20	HKLLK	18	KKLVR	16
VGLLR	20	IELVQ	18	LVLRR	16
AGLGI	19	IPLGD	18	MPLVP	16
AGLLQ	19	LALGG	18	NLLPT	16
ARLAG	19	MLLSH	18	PKLQP	16
AVLSH	19	MRLKV	18	PVLMG	16
DELRV	19	MRLRS	18	QALIG	16
DRLAG	19	MSLSP	18	RGLIT	16
ERLSN	19	PALGG	18	RGLTA	16
ETLM	19	PGLHG	18	RRLVN	16
GELRG	19	PTLQR	18	RVQD	16
GKLVA	19	QPLAG	18	SLLEE	16
GRLNR	19	SK*VV	18	TELPM	16
GRLRL	19	SKLRG	18	TGL	16
IMLAG	19	TLIN	18	TMLGG	16
IVLDP	19	TLLTP	18	VPLVP	16
KVLAP	19	DALME	17	APLDL	15
LMLGM	19	EALNK	17	ARLGR	15
MALTR	19	EGLPT	17	DALSA	15
MPLRE	19	ELLKS	17	EGLAG	15
RLLGP	19	GELTD	17	GGLVR	15
SKLRA	19	GKLRQ	17	GGLVS	15
SMYRS	19	GMLRR	17	GHLRV	15
THLAK	19	GPLVS	17	GKLIV	15
VELDP	19	GQLMM	17	GLLRG	15
VGLTR	19	GQLVG	17	GMLGT	15

GPLLG	15	GRLLE	14	SHLRV	13
HIRSH	15	KKLLE	14	SRLGV	13
IGLQR	15	KTLGV	14	TGLAR	13
IMLV	15	MALPE	14	TLMG	13
IRLAL	15	MRLMG	14	TRLMM	13
KALGT	15	NDALQYES	14	TRLREHIRSHTGERPF	13
KALRG	15	NRLES	14	VELGP	13
LHLQG	15	NSLTR	14	VHLAR	13
MELMT	15	PKLRS	14	VKLVG	13
MPLGG	15	PRLPP	14	AGLAA	12
PGLAD	15	PVLKL	14	APLRV	12
PTLEV	15	QKLAN	14	EALV	12
RQLGM	15	QKLKL	14	EVLPE	12
RVLRG	15	RALPK	14	GALMN	12
SGLLR	15	RILPR	14	GLQA	12
SVLRV	15	THLGR	14	GLTG	12
TGLGL	15	TMLPG	14	GTLGD	12
TVLAG	15	VALGT	14	HLLGP	12
VGLA	15	VKLHE	14	LKLKL	12
VGLRG	15	VTLG	14	MALRK	12
VMLKD	15	ARLLG	13	MVLTG	12
VPLNT	15	ARLTG	13	NGLIE	12
VTLGR	15	ASLGA	13	NKLVV	12
WR	15	DLLSG	13	PALNV	12
A	14	DNLRE	13	PMLRL	12
AALHH	14	EALTI	13	PQLLG	12
AHLKA	14	EELMM	13	PVLRV	12
ALLGV	14	ETLS	13	QPLKR	12
AQLVD	14	GALGS	13	RGLDM	12
ARLHA	14	GHLRT	13	RGLEN	12
DGLG	14	GPLVL	13	RLLGA	12
DHLVG	14	GRLGA	13	RRLMV	12
DILRG	14	GRSYMA	13	SGLVR	12
DQLVE	14	GVLGS	13	SPLSG	12
DQLVG	14	HPLLV	13	SQLLE	12
EKLMM	14	ITLSP	13	SRLGR	12
ELLTP	14	MGLDP	13	TGLVG	12
ERLGR	14	MLLNG	13	THLRR	12
GALRS	14	MRLAE	13	TKLRM	12
GGLTM	14	NMLSR	13	TKLVM	12
GKLHE	14	PGLGG	13	TLLG	12
GKLNI	14	PGLVP	13	TMLPR	12
GKLVL	14	PLLKS	13	TNLRL	12

TPLGE	12	AGLTE	10	AGLDG	9
TPLVG	12	AGLVA	10	AGLVR	9
TRLLT	12	ALSA	10	DELMR	9
VGLGR	12	ATLMK	10	DSLVI	9
VKLQ	12	DGLAR	10	EKLKA	9
VLLKS	12	DHLNV	10	EKLRV	9
AGLML	11	EALAI	10	GKLIA	9
AKLTL	11	EELVE	10	GNLVT	9
ANLSN	11	EMLIP	10	GRLLI	9
ARLLT	11	EPLAA	10	GRLRS	9
ATLLR	11	ERLQE	10	GSLIR	9
EGLGG	11	GGLKD	10	GTLKR	9
EGLHL	11	GKLRA	10	HELMK	9
GHLKM	11	GKLVV	10	KMLGG	9
GRLAV	11	GMLRV	10	LGLIQ	9
GVLG	11	GPLME	10	LKLER	9
LGLEG	11	GVLSV	10	LPLNG	9
LNLQP	11	IKLMG	10	MGLGV	9
LRLRT	11	IPLNR	10	MVLAG	9
MELGD	11	MLLKG	10	MVLVG	9
MLLQR	11	MRLPR	10	NGLAR	9
MLPP	11	MSLRE	10	PILLQ	9
MSLGG	11	PNLAG	10	PMLTV	9
PKLII	11	PPLMV	10	PPLII	9
PNLQT	11	PTLGV	10	QRLVE	9
PPLLS	11	RGLRN	10	RKLIV	9
PTLGM	11	RGLVV	10	RKLKE	9
QKLMT	11	RSLIV	10	RRLHE	9
QTLAE	11	RTLGE	10	RVLGA	9
RLLMP	11	SSLGV	10	SALAR	9
RRLVG	11	TALIS	10	SC*RP	9
SNLIM	11	TGLGT	10	SGLDA	9
STLLM	11	THLRL	10	SQLDR	9
THLAR	11	TKLAV	10	TGLLR	9
TLTM	11	TKLLG	10	TKLKT	9
TRLGG	11	TLIG	10	TSLTE	9
TRLQK	11	TNLLR	10	AGLKM	8
VGLLA	11	TTLGG	10	AGLRS	8
VKLRM	11	VILGA	10	AHLGQ	8
VLLGG	11	VNILE	10	AHLR	8
VQ*GG	11	AALES	9	ALME	8
VRLEE	11	AALGL	9	ASLRR	8
AGLSG	10	AELMR	9	DALDR	8

DGLHG	8	AVLEN	7	ARLKV	6
DGLLQ	8	GDLSG	7	DGLLR	6
EELGL	8	GELGV	7	DKLHR	6
EKLRS	8	GKLV	7	DKLKV	6
EQLMT	8	GLLD	7	ERLPV	6
ERLAR	8	GLLMG	7	GDLVE	6
GALGR	8	GLMA	7	GELGE	6
GELKA	8	GLNR	7	GGLMQ	6
GGLVV	8	GLVV	7	GLLT	6
GHLRM	8	GPLPV	7	GLPG	6
GLML	8	GSLTQ	7	GSLRT	6
GLQN	8	GVLRG	7	GTLQV	6
GLTA	8	HPLAV	7	GVLKS	6
GMLGE	8	HTLGM	7	HGLVN	6
GPLRR	8	IQLGG	7	IELGR	6
GVLDT	8	KLLGD	7	KPLEL	6
GVLNT	8	KNLIK	7	MKLE	6
IQLAD	8	MALAR	7	NLTPK	6
KGLTM	8	MELEP	7	PALMR	6
MELGN	8	MGLAN	7	PHLVV	6
MPLMR	8	MGLGE	7	PPLVV	6
MTLSD	8	MPLDG	7	QALVP	6
NGLAM	8	NVLGR	7	QELGG	6
NGLQD	8	PGLPE	7	QHLRR	6
NTLDV	8	PHLQN	7	QTLGV	6
PHLSM	8	PRLGS	7	RILEP	6
PILLG	8	PSLLV	7	RLLMN	6
PVLQG	8	PTLAR	7	RPLVG	6
QGLGG	8	QMLER	7	RRLEP	6
QKLQI	8	RDLGS	7	SGLRA	6
QPLIA	8	RGLGN	7	SKLMA	6
RGLVA	8	RLLEK	7	SKLSV	6
SLLNG	8	RMLVP	7	TMLEP	6
SRLTD	8	SVLSG	7	TRSQ	6
TLLGD	8	TGLVN	7	VALRK	6
TRSHSSV	8	TLA*SH	7	VDLSG	6
TSLTR	8	TRLHT	7	VMILG	6
TTLGD	8	VALTK	7	VPLSE	6
VKLAP	8	VMLMG	7	AGLDR	5
VQLPV	8	VVLAG	7	ARLPV	5
AALRK	7	AGLVG	6	ARYGC	5
AHLKK	7	AKLKL	6	ATLKV	5
AKLHP	7	AR*PS	6	DGLRA	5

ERLLQ	5	GELV	4	VSLGV	4
ETLMG	5	GGLRP	4	VSLTA	4
GHLML	5	GHLKK	4	VSLVG	4
GHLOQ	5	GKLKA	4	AGLKR	3
GKLMV	5	GLIG	4	AGLQN	3
GPLG	5	GLIS	4	AGLRV	3
GPLTM	5	GLLGN	4	ALLRR	3
GQLV	5	GMLVN	4	ARLRT	3
GSLTL	5	GPLED	4	ASLQK	3
GTLRA	5	GPLQA	4	ASLR	3
GTLTG	5	GTLTV	4	ATLSR	3
HHLTK	5	GVLGI	4	DILGE	3
IVLVR	5	IDLGM	4	EELRM	3
MALVR	5	IELGG	4	EGLTG	3
MELGK	5	IGLAT	4	EMLKE	3
MGLEG	5	KKLMP	4	ESLLG	3
MGLMA	5	KLLGE	4	ESLMA	3
MPLNR	5	KLLLG	4	ETLAG	3
NMLGG	5	KMLPP	4	EVLVQ	3
NPLEL	5	MGLTL	4	GALKR	3
NSLGG	5	MNLGM	4	GGLGR	3
PRLQ	5	MPLMV	4	GHLRK	3
PRLVK	5	MPLRA	4	GKLRS	3
RGLVR	5	MQLGG	4	GLKT	3
RHLRS	5	MRLRM	4	GLLGV	3
RSLVV	5	PALTV	4	GMLLP	3
RSPV*ERMWILRA	5	PGLAL	4	GMLSG	3
RTLNA	5	PGLMG	4	GPLMG	3
TELN	5	PHLMS	4	GRLAP	3
VKLRA	5	PQLSA	4	GSLLR	3
VLLQD	5	PRLKA	4	GTLTM	3
VMLG	5	QKLIR	4	GVI	3
AGLNG	4	RELGV	4	ILLQQ	3
AHLRM	4	RGLHQ	4	KLLQM	3
AKLRA	4	RGLIG	4	LGLPG	3
AR*RA	4	RGLMG	4	MELVL	3
ARLPE	4	RTRSH	4	MGLAG	3
AVLNK	4	SQLDT	4	MGLPV	3
DALQYESECGGLNH	4	TEGG	4	MGLQN	3
DTLLR	4	THLKK	4	MQLAD	3
EGLRD	4	THLRA	4	MSLLR	3
ESLMG	4	TKLGV	4	MSLPE	3
G	4	TMLEG	4	NGLKQ	3

NGLQR	3	GDLVP	2	PPLLG	2
NGRSPV*E	3	GELRI	2	PQLTE	2
NPLSR	3	GGLEL	2	PVLDG	2
NQLVA	3	GHLSP	2	QGLTT	2
NTLGL	3	GKLEA	2	QRLAV	2
PRLRV	3	GKLKR	2	RELGG	2
PVLLM	3	GKLRR	2	RGLDG	2
PVLTG	3	GKLVI	2	RGLTE	2
QHLRK	3	GLHQ	2	RHLGA	2
QQLL	3	GLLR	2	RSLMI	2
RGLVN	3	GLMV	2	RSLRP	2
RHLVV	3	GLTL	2	SKLGA	2
RLLA	3	GNLVR	2	SKLGE	2
RLLPG	3	GPLVG	2	T*LT	2
RPLIT	3	GQLVD	2	TALKV	2
RVLNM	3	GRLSV	2	THLR	2
RVLQR	3	GVLAV	2	TRLKV	2
SGLER	3	HGLTG	2	TRLPP	2
TKLKL	3	HVLEL	2	VELGD	2
TLLPG	3	IELEM	2	VELVN	2
TRLRE	3	IGLQA	2	VGLGG	2
TSLER	3	KGLGN	2	VGLKD	2
VGLPA	3	KILPV	2	VKLHV	2
VPLRP	3	KPLPG	2	VKLLS	2
VRLMP	3	KSLRM	2	VQLTK	2
VSLGE	3	KTLGT	2	VRLK	2
AALTK	2	LGLAA	2	VRLPP	2
AALVK	2	LGLGG	2	AALEN	1
AHLTP	2	LVLQE	2	AALGP	1
AILRT	2	MGLAS	2	AALGT	1
AKLNS	2	MLLEE	2	AALKI	1
AKLNV	2	MLPA	2	AALMN	1
AKLRG	2	MRLAR	2	AALMQ	1
ALLGA	2	MSLRQ	2	AALMR	1
ARLLR	2	MTLGT	2	AALRV	1
ARLRA	2	NGLIV	2	AALSS	1
DVLG	2	NHLRM	2	AELGP	1
EELQS	2	NLA	2	AELRA	1
EGLVE	2	PALIM	2	AELRI	1
ELLGP	2	PGLAG	2	AGIAA	1
ERMC	2	PLLRA	2	AGILQ	1
EVLAG	2	PPLDG	2	AGLDS	1
GALGE	2	PPLIM	2	AGLG	1

AGLGG	1	AT*RS	1	EKLIV	1
AGLGN	1	ATLAK	1	EKLKV	1
AGLGP	1	ATLEV	1	ELLAR	1
AGLGQ	1	ATLKI	1	ELLPS	1
AHFRV	1	ATLMG	1	EMLVA	1
AHLRG	1	ATLNM	1	EQLGT	1
AHLRP	1	ATLNV	1	ERLAV	1
AKFRM	1	AVIG	1	ERLRV	1
AKLE	1	CGLGR	1	ETLNS	1
AKLGE	1	DALQP	1	ETSSH	1
AKLGL	1	DALTV	1	EVLAV	1
AKLHA	1	DELM	1	EVLGI	1
AKLKG	1	DELMN	1	EVLIQ	1
AKLLG	1	DELRA	1	EVLQE	1
AKLML	1	DGLE	1	GALGL	1
AKLQP	1	DGLEK	1	GALGV	1
AKLRF	1	DGLES	1	GALIS	1
AKLRQ	1	DGLML	1	GALMQ	1
AKLS	1	DGLTGHRSHTGERPF	1	GALRD	1
AKLTN	1	DGVAM	1	GALRG	1
AKLWL	1	DHLVD	1	GAVMN	1
ALDA	1	DILG	1	GE*GI	1
ALIM	1	DILRT	1	GELKV	1
ALKG	1	DKLKG	1	GELML	1
ALLGE	1	DKLMM	1	GELMR	1
ALLRS	1	DLLA	1	GELRV	1
ALTG	1	DLLAR	1	GELTG	1
ALTR	1	DNLRV	1	GFLAR	1
AMLPD	1	DRLAA	1	GGFRD	1
AMLR	1	DRLGG	1	GGLA	1
APLAG	1	DSLPE	1	GGLAE	1
APLGP	1	DSLIV	1	GGLGA	1
AQLAD	1	DVLRG	1	GGLGE	1
AQLLL	1	DYLVN	1	GGLGP	1
AR*RG	1	EALA	1	GGLHP	1
ARLAA	1	EALKV	1	GGLKV	1
ARLGT	1	EALMV	1	GGLMD	1
ARLMS	1	EALTN	1	GGLMT	1
ARLRS	1	EELAP	1	GGLNI	1
ARLTE	1	EELMMHIRSHTGERPF	1	GGLRG	1
ARYGR	1	EELVEHIRSHTGERPF	1	GGLRL	1
ASLGP	1	EHLRL	1	GGLSG	1
ASLRP	1	EHLVR	1	GGLVG	1

GGVGL	1	GPLVP	1	IGIQR	1
GHLAI	1	GQLGP	1	IGLGA	1
GHLQC	1	GQLLE	1	IGLRL	1
GHLQR	1	GR*ML	1	IHLAG	1
GHLRR	1	GRLGG	1	IHLRM	1
GHLSV	1	GRLLG	1	IKLTG	1
GHLVG	1	GRLMP	1	IMLPR	1
GHLVK	1	GRLVS	1	IQLMG	1
GILAR	1	GRYGC	1	IQLRL	1
GILSG	1	GSLRV	1	IRLAA	1
GKLAI	1	GSLSK	1	IRLGP	1
GKLGG	1	GSLSP	1	IRLGV	1
GKLIG	1	GTLKL	1	IRLRR	1
GKLII	1	GTLLL	1	ISLVG	1
GKLIT	1	GTLLV	1	ITLMV	1
GKMKHRSHTGERPF	1	GTLMT	1	ITLRG	1
GKLLK	1	GTLRV	1	ITLRP	1
GKLNA	1	GTLTE	1	ITLVG	1
GKLPT	1	GVIN	1	IVLPG	1
GKLQA	1	GVL	1	KG	1
GKLR	1	GVLDN	1	KGLAT	1
GKLRA	1	GVLE	1	KGLDL	1
GKLRE	1	GVLKD	1	KGLMR	1
GKLT	1	GVLQK	1	KGRSPVET	1
GKLTM	1	GVLRL	1	KIIV	1
GLAA	1	GVLSG	1	KILLA	1
GLIV	1	GVLTG	1	KKLAG	1
GLLEK	1	GVMN	1	KKLGV	1
GLLGG	1	GVPV	1	KKLRI	1
GLLMV	1	HELMR	1	KLLAG	1
GLLPG	1	HLLVP	1	KLLRV	1
GLLQD	1	HPLDR	1	KPLAA	1
GLLTG	1	HPLLS	1	KPLMV	1
GLSG	1	HPVKE	1	KRLEG	1
GLSGR	1	HTLKM	1	KSLVG	1
GLSV	1	HTLLK	1	KTLEG	1
GLVN	1	HTLNI	1	KTLRG	1
GLVQ	1	HTLNK	1	KTLRV	1
GMLAG	1	HTLRP	1	KTLVG	1
GNLSN	1	IALPG	1	KVLPV	1
GPLA	1	IELAL	1	LAHGT	1
GPLKP	1	IELG	1	LGLGP	1
GPLRP	1	IELHL	1	LGLGV	1

LKVKL	1	NGLIVHIRSHTGERPF	1	PVLKV	1
LNLHT	1	NGR	1	PVLMT	1
LRLIM	1	NGRPPG*E	1	QALKR	1
LRVIG	1	NGRSPVR	1	QALRG	1
LSLSG	1	NILMG	1	QALSP	1
LTQQ	1	NKLAR	1	QGLHL	1
LVLRG	1	NKLRA	1	QGLPV	1
MALRG	1	NKLRG	1	QILLQ	1
MELIG	1	NKLVA	1	QILLRHIRSHTGERPF	1
MGLRV	1	NKLVK	1	QILLY	1
MLAA	1	NMLGV	1	QILPE	1
MLLIS	1	NNLIN	1	QMLAR	1
MLLLP	1	NRLRE	1	QPLAV	1
MLLMV	1	NRLRI	1	QPLTM	1
MLLPP	1	NSLV	1	QRLGG	1
MLLPV	1	NSLVA	1	QTLAV	1
MLLV	1	NVHP*VVGLAA	1	QTLGG	1
MLLVG	1	NVLGE	1	QTLGP	1
MLVG	1	PALAG	1	REIVR	1
MMLDP	1	PALGP	1	REIRR	1
MPLGA	1	PALV	1	RGLAA	1
MPLGL	1	PASV	1	RGLDN	1
MPLLG	1	PDLRA	1	RGLNS	1
MRLEE	1	PGITE	1	RGLRS	1
MRLGA	1	PGLAP	1	RGLTG	1
MRLGG	1	PGLHE	1	RGLVE	1
MRLGR	1	PGVAA	1	RGYGT	1
MRLVG	1	PGVVP	1	RHE	1
MSLHG	1	PHLKR	1	RHLKM	1
MSLQQ	1	PKLIF	1	RLLGL	1
MTLER	1	PLRG	1	RP*SG	1
MVL	1	PMLAG	1	RPLAG	1
MVLMN	1	PMLTM	1	RQLGK	1
MVLNT	1	PNLAS	1	RQLLE	1
MVLRG	1	PNLAV	1	RRLEA	1
MVLVT	1	PNYW	1	RRLET	1
MVVAS	1	PNYWS	1	RRLGD	1
NDALQYD	1	PQLVV	1	RRLGS	1
NDALQYSECGP	1	PQSRG*RG	1	RRLSE	1
NELLR	1	PR*GA	1	RRLTP	1
NELMR	1	PRLRL	1	RRVVG	1
NELRV	1	PSFQ	1	RSH	1
NGLG	1	PTLAK	1	RTLKL	1

RTLVG	1	TELIR	1	TVLGG	1
RVLEP	1	TELPR	1	TVLT	1
RVLRE	1	TELRV	1	V*KS	1
SC**A	1	TGLAD	1	VALHT	1
SCLK	1	TGLGA	1	VDLLL	1
SGILV	1	THLAN	1	VELAP	1
SGLGG	1	THLAV	1	VELN	1
SGLGL	1	THLRK	1	VELNN	1
SGLGT	1	TKIRV	1	VELRV	1
SGLLG	1	TKLKA	1	VGLPV	1
SGLNL	1	TKLLR	1	VGLQA	1
SGLRL	1	TKLME	1	VGLQR	1
SGLVG	1	TKLNV	1	VGLRN	1
SHLRL	1	TKLR	1	VGLRV	1
SKLIL	1	TKLRA	1	VGLSP	1
SKLKA	1	TKLRI	1	VGLSQ	1
SKLKG	1	TKLRR	1	VHLAL	1
SKLLG	1	TKLVL	1	VKLMA	1
SKLRI	1	TKSGV	1	VKLQN	1
SKLRM	1	TLIS	1	VKLRL	1
SKLVL	1	TLLIR	1	VLLAA	1
SL*HG	1	TLLM	1	VLLIE	1
SLLRT	1	TLLMQ	1	VLLKI	1
SNLTY	1	TLNG	1	VLLTP	1
SNYWP	1	TLQP	1	VLMV	1
SPLIG	1	TMLDP	1	VLQR	1
SPLKI	1	TMLRE	1	VMLRG	1
SPLRN	1	TNLVG	1	VPLAL	1
SQLKV	1	TPLIV	1	VPLVG	1
SQMK	1	TPLMQ	1	VQLPM	1
SR*G	1	TPLSD	1	VQLRV	1
SRLKV	1	TPLSI	1	VRLEG	1
SRLMT	1	TQLED	1	VRLGG	1
SRLVT	1	TRLGA	1	VRLQA	1
SSLGA	1	TRLMI	1	VRLVR	1
SSLGL	1	TRLRL	1	VTG	1
STLQK	1	TRLRV	1	VTLER	1
SVLVG	1	TRLTG	1	VTLGS	1
SVLVS	1	TSLSE	1	WRN	1
T	1	TTLEP	1		
TALEA	1	TTLGV	1		
TALKG	1	TTLKV	1		
TELE	1	TTLRV	1		

**Table 22. ZF4
selection on G:A
change at nt 11 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	Read #				
GNLRR	3407	ANLAR	102	TALKR	12
ANLRR	1937	ATLRR	96	KNLER	11
GNLVR	1794	HNLRR	90	NNLVR	11
SNLRR	1771	AALRR	81	THLRR	11
SNLKR	1208	SNLLR	77	SMLQR	10
TNLRR	862	ANLER	73	TNLMR	10
GNLKR	850	SMLRR	62	TTLRR	10
GNLAR	805	ANLHR	59	AGLRR	9
SNLVR	764	ANLLR	58	ATLTR	9
GNLQR	562	SMLKR	51	HNLKR	9
GNLTR	531	SNLSR	47	KNLVR	9
NNLRR	486	TNLQR	47	SGLRR	9
GNLIR	475	GALKR	44	VNLRR	9
ANLVR	455	GALRR	44	AMLTR	8
GNLNR	373	AHLRR	42	AVLRR	8
GNLIR	360	HMLRR	40	GHLKR	8
ANLKR	326	HMLKR	36	TNLTR	8
SNLAR	269	RNLQR	35	ATLNR	7
SNLQR	261	AMLKR	31	ATLVR	7
GNLMR	251	GMLKR	31	HNLMR	7
SNLMR	228	GMLRR	29	TALRR	7
GNLHR	220	SALKR	28	TNLLR	7
SNLTR	209	SNLER	26	AMLQR	6
NNLKR	202	KNLQR	25	ANLSR	6
SNLIR	199	RNLRR	24	ASLKR	6
ANLMR	191	GTLRR	21	ASLTR	6
ANLNR	179	TMLKR	21	ATLMR	6
TMLRR	158	GNLER	20	GMLER	6
SNLNR	155	QNLKR	19	RLLIN	6
ANLTR	136	RRLDR	19	SGLLK	6
ANLQR	125	STLRR	19	TNLAR	6
TNLKR	118	TNLNR	17	AHLVR	5
AMLRR	111	QNLRR	16	ATLHR	5
GNLSR	107	SALRR	16	ATLQR	5
		TNLHR	16	ENLRR	5
		ANLIR	15	HMLQR	5
		AALKR	14	HMLTR	5
		ATLKR	14	ISLRV	5
		ASLRR	13	NNLAR	5
		NTLRR	13	SNLHR	5
		TNLVR	13	STLKR	5
		GHLRR	12	AHLKR	4
		NNLMR	12	ATLIR	4

DNLRR	4	ANLRL	2	ANIQR	1
GALTR	4	ANLRW	2	ANLDR	1
GNLRK	4	ATLAR	2	ANLGR	1
KKLLR	4	DALLV	2	ANLVG	1
MNLRR	4	GGLIR	2	ANSRR	1
MVLLR	4	GILAR	2	ANVRR	1
NNLQR	4	GILRR	2	APLRR	1
QNLVR	4	GN*S...	2	ASLQR	1
RNLAR	4	GNLDR	2	ATLLR	1
SHLRR	4	GNLNK	2	ATLRS	1
SMLHR	4	GNLRP	2	AYFRR	1
TNLER	4	GNLRS	2	CNLAR	1
TVLLV	4	GTLIR	2	CNLNR	1
AALNR	3	GTLMR	2	CNLVR	1
AALVR	3	GTLVR	2	DNLKR	1
AGLKR	3	HGLET	2	DNLVR	1
ETLRR	3	HMLNR	2	GALNR	1
GHLQR	3	HNLVR	2	GDLAR	1
GHLVR	3	KNLMR	2	GDLRR	1
GMLNR	3	NNLLR	2	GDLTR	1
GMLTR	3	NSLRR	2	GGLKR	1
GMLVR	3	PGLLG	2	GGLRR	1
GNLGR	3	RNLVR	2	GGVRR	1
GNLRG	3	SMLNR	2	GHLNR	1
GNLVK	3	SMLTR	2	GNFRR	1
GSLRR	3	SNLDR	2	GNFVG	1
GTLKR	3	STLMR	2	GNLAG	1
HTLRR	3	SVLRR	2	GNLAS	1
HVLRR	3	TGLRR	2	GNLHK	1
KNLRR	3	TMLQR	2	GNLLS	1
MNLKR	3	TSLKR	2	GNLMS	1
NGRSPV...	3	TTLKR	2	GNLNH	1
NMLRR	3	TVLRR	2	GNLQS	1
PNLIR	3	AALRS	1	GNLRH	1
QHLRR	3	ADLER	1	GNLS...	1
SALNR	3	ADLVR	1	GNLTK	1
STLEV	3	AGLMR	1	GNLTK	1
STLNR	3	AHLTR	1	GNLTV	1
TALRS	3	AILTR	1	GNLVW	1
THLKR	3	AMLAR	1	GNLWR	1
TNLIR	3	AMLHR	1	GNSKR	1
AALTR	2	AMLIR	1	GNSQR	1
AMLNR	2	ANFRR	1	GNSRR	1

GNVQR	1	SNLVW	1
GNVTR	1	SNSRR	1
GQLAL	1	SNVKR	1
GSLKR	1	SNVRG	1
GTLNR	1	STLVR	1
GY*LR	1	TMFRR	1
HNLAR	1	TMLNR	1
HNLQR	1	TNLGR	1
ITLQR	1	TNLLS	1
KILGN	1	TPTRS	1
KNLKR	1	TQLVL	1
KNLTR	1	TSLRR	1
KSLRR	1	VNLTR	1
LNLRR	1	VTLRR	1
LNLVR	1		
NMLKR	1		
NNLIR	1		
NNLNR	1		
NNLTR	1		
NNSRR	1		
NTLAR	1		
NTLIR	1		
NTLKR	1		
NTLNR	1		
PRLRG	1		
QHLKR	1		
QMLKR	1		
QTLRR	1		
RLIIN	1		
RNLKR	1		
SALQR	1		
SGLKR	1		
SHLKR	1		
SILNR	1		
SKLTR	1		
SMLIR	1		
SMLVR	1		
SNLFR	1		
SNLIH	1		
SNLRK	1		
SNLRQ	1		
SNLSG	1		
SNLTS	1		

**Table 23. ZF4
selection on G:C
change at nt 11 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence	Read #
AMLRR	3064
TMLRR	2212
SMLRR	1556
GMLRR	1320
GMLKR	1160
SMLKR	1049
AMLKR	797
TMLKR	515
HMLRR	478
HMLKR	429
ATLRR	261
SMLQR	248
AMLQR	212
AMLNR	125
AMLTR	124
TMLQR	123
NMLRR	119
NMLKR	102
SMLTR	98
HMLTR	93
GMLTR	88
AMLHR	72
GMLNR	68
SMLHR	63
SMLNR	59
ATLKR	58
GMLVR	50
HMLNR	47
HMLQR	47
AMLAR	46
AMLVR	45
TMLNR	38
GMLHR	34

SMLVR	34	NGRSPV...	7
SMLAR	29	SGLKR	7
STLRR	27	TMLMR	7
SVLRR	27	TTLKR	7
RRLDR	26	ATLQR	6
GNLRR	24	ATLTR	6
TTLRR	24	GMLMR	6
AVLRR	23	GNLTR	6
AMLIR	22	GTLKR	6
GMLAR	22	SNLKR	6
GMLQR	22	SSLKR	6
GTLRR	19	AMLER	5
ASLRR	18	AMVRR	5
HMLAR	17	ANLVR	5
NTLRR	17	GNLKR	5
ANLRR	16	SNLTR	5
QMLRR	16	TMLAR	5
ATLNR	15	VELNS	5
QMLKR	15	ATLAR	4
RRLKN	15	ATLIR	4
AMLMR	14	EMLKR	4
ATLMR	14	GNLAR	4
GMLIR	14	GNLMR	4
EMLRR	13	NNLKR	4
GNLVR	13	SMLMR	4
RTLAL	13	STLKR	4
SMLSR	13	TGLRR	4
SNLRR	13	ANLER	3
SMLIR	12	ANLKR	3
AGLKR	11	ANLNR	3
TRS	11	ANLQR	3
ATLVR	10	ASLKR	3
TRS...	10	DMLRR	3
AALRR	9	GALRR	3
ATLHR	9	GNLNR	3
SNLVR	9	HLLRR	3
TMLTR	9	HMLHR	3
TNLRR	9	HVLRR	3
TMLVR	8	CTLRR	3
TVLRR	8	LL...	3
GGLKR	7	NNLRR	3
GNLIR	7	SGLTR	3
GNLQR	7	SMLKN	3

SMLLR	3	AMITR	1	HLLNS...	1
SNLQR	3	AMKTR	1	HMLLR	1
STLNR	3	AMLCR	1	HMLMR	1
AALKR	2	AMLHS	1	HMVRR	1
AALNR	2	AMLPR	1	HNLMR	1
ALLRR	2	AMLR...	1	HVLKR	1
AMLLR	2	AMLRG	1	KILGN	1
AMLRH	2	AMLRK	1	KMLKR	1
AMLRS	2	AMLTM	1	LMLGK	1
AMLRW	2	AMLWR	1	MLRR	1
AMLSR	2	AMYT...	1	NLLKR	1
ANLHR	2	ANLAR	1	NMLGR	1
AVLKR	2	ARLRR	1	NTFRR	1
GMLER	2	ARLRS	1	NTLIR	1
GMLKN	2	ARLRV	1	NTLNR	1
GMLRW	2	ASLNR	1	PMLMR	1
GMVRR	2	ATLER	1	PVVKR	1
GSLRR	2	ATLGR	1	QSLKR	1
GVLRR	2	ATLRS	1	RMFRR	1
HILKR	2	ATLSR	1	RMLRR	1
HILRR	2	CMLRR	1	SALNR	1
HMLRS	2	DMLKR	1	SALRR	1
HMLVR	2	DMLQR	1	SGLRR	1
HNLRR	2	DMVKR	1	SILKR	1
HTLRR	2	EMLNS	1	SKLKR	1
MGLST	2	ETLRR	1	SKLTR	1
NMLIR	2	GLLKR	1	SLLNR	1
NTLKR	2	GLLQS	1	SMFRR	1
QTLRR	2	GLLSR	1	SMIKR	1
RMLKR	2	GMIKR	1	SMLGR	1
RMLQR	2	GMLKT	1	SMLKW	1
RN*P...	2	GMLRM	1	SMSRR	1
SMFKR	2	GMLTW	1	SMVKR	1
SNLAR	2	GNLHR	1	SNLLR	1
STLMR	2	GNLLR	1	SNLNR	1
TLLRR	2	GRLKR	1	SQLKR	1
TMIRR	2	GRLKS	1	SRLKR	1
TMVRR	2	GRLRV	1	SRLRR	1
VIKR...	2	GTLNR	1	SSLAR	1
AKLQR	1	GTLVR	1	STLVR	1
ALLKR	1	GVLKR	1	SVLKR	1
AMFRR	1	HALRR	1	TILRR	1
AMIRR	1	HLLKR	1	TMLER	1

TMLGR	1
TMLHR	1
TMLLR	1
TMLRH	1
TMLWR	1
TNLKR	1
TNLSR	1
TPTRS	1
VMLKR	1
VSLRK	1
VTLRR	1
WMLKR	1
WMLRR	1
YMLKR	1
YMLRR	1

Table 24. ZF4
selection on G:T
change at nt 11 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence	Read #				
ATLRR	6399	STLNR	58	ASLTR	13
RRLDR	1155	SALKR	47	GTLMR	13
GTLRR	1073	AALNR	46	HMLKR	13
TTLRR	1024	HALRR	44	STLLR	13
STLRR	970	GALRR	41	STLTR	13
ATLNR	770	QTLKR	41	TTLNR	13
ATLVR	727	HTLKR	39	GNLTR	12
ATLKR	708	SMLKR	39	KILGN	12
NTLRR	670	HALKR	36	TNLRR	11
AALRR	657	SVLRR	36	ARLRI	10
ATLMR	456	AVLKR	35	ETLKR	10
ATLAR	418	SMLRR	34	GSLRR	10
GTLKR	399	GMLRR	33	GTLAR	10
ASLRR	366	GMLKR	31	KTLEK	10
ATLTR	363	STLMR	31	QTLMR	10
QTLRR	298	STLVR	31	SILRR	10
ETLRR	292	GTLVR	28	SNLKR	10
ATLQR	291	GALKR	27	TALKR	10
ATLIR	261	GTLNR	27	ASLQR	9
STLKR	252	TTLQR	27	GVLKR	9
NTLKR	237	GVLRR	25	NNLRR	9
HTLRR	227	GNLRR	24	ATLDR	8
AVLRR	208	SNLRR	24	GNLAR	8
ATLLR	171	TSLRR	24	GNLIR	8
AALKR	141	ANLRR	23	KTLLR	8
ATLER	134	VTLRR	23	RTLRR	8
ATLHR	130	AALTR	22	SMLQR	8
TTLKR	125	AMLKR	22	SSLKR	8
ASLKR	95	STLQR	22	ARLRV	7
AMLRR	93	ATLRK	21	GALVR	7
HVLRR	81	GNLVR	21	GTLIR	7
SALRR	63	HMLRR	20	HVLKR	7
HILRR	60	ATLSR	17	MTLRR	7
TMLRR	59	RTLQR	17	SNLTR	7
		TALRR	17	SNLVR	7
		GSLKR	16	ANLKR	6
		STLIR	16	GNLKR	6
		AALQR	15	HILNR	6
		ASLVR	15	MTLMR	6
		STLHR	15	AALVR	5
		SSLRR	14	AMLQR	5
		TMLKR	14	GNLQR	5
		TVLRR	14	NTLNR	5

NTLVR	5	GPLVR	3	QRLSV	2
QTLTR	5	GSLSR	3	RPLVG	2
RTLKR	5	HMLTR	3	RRIDR	2
SALTR	5	HSLRR	3	RRIDS	2
SVLKR	5	HVLNR	3	RRVDR	2
AMLTR	4	NSLKR	3	RSLIR	2
ANLNR	4	NTLMR	3	RTLIR	2
ANLVR	4	NVLRR	3	SDLTV	2
APLRR	4	QTLQR	3	SMLHR	2
ATLGR	4	RRLNR	3	SRLKI	2
ATVRR	4	SALNR	3	SSLVR	2
DTLRR	4	SALQR	3	STVRR	2
GILRR	4	SVLTR	3	TTLTR	2
GMLNR	4	AGLKR	2	TTLVR	2
GTLIR	4	AGLRR	2	TVLKR	2
GTLQR	4	AGLVR	2	AALAR	1
GVLSR	4	AMLHR	2	AALLR	1
HRLKI	4	AMLIR	2	AALNS	1
HTLVR	4	ARLKI	2	AALRK	1
KTILR	4	ASLNR	2	AALRL	1
KTIRR	4	ASLSR	2	AALRS	1
NNLKR	4	ATFRR	2	AAVRR	1
NSLRR	4	ATLNW	2	AKLQR	1
NTLIR	4	ATLRW	2	AKLRR	1
SMLTR	4	ESLRR	2	ALLKR	1
SNLAR	4	ETLVR	2	AMLAR	1
SNLQR	4	GALNR	2	AMLNR	1
STLAR	4	GALQR	2	AMLVR	1
TILAR	4	GMLVR	2	ANLHR	1
TSLKR	4	GNLLR	2	ANLQR	1
TTLAR	4	HALTR	2	ANLTR	1
TTLMR	4	HMLNR	2	AP*C...	1
AILRR	3	HNLRR	2	APLHR	1
ANLMR	3	HTLMR	2	APLKR	1
ASLAR	3	HTLQR	2	APLMR	1
ASLHR	3	IALAG	2	APLVR	1
ATLNK	3	MSLRR	2	APYP...	1
ATLRG	3	MTLLR	2	ARLRR	1
EALRR	3	MTLVR	2	ARLTR	1
GILKR	3	NGRSPV...	2	ARLVG	1
GMLTR	3	NMLKR	2	ASFRR	1
GNLNR	3	NMLRR	2	ASLER	1
GPLPV	3	PTLRR	2	ASLMR	1

AT*G...	1	GTLMW	1	QTLIR	1
ATFKR	1	GTLRK	1	QTLRQ	1
ATFRT	1	GTLTR	1	QTLVR	1
ATFTR	1	GTLVS	1	RN*P...	1
ATIRR	1	GTSNR	1	RRLAG	1
ATLES	1	GTSRR	1	RRLAR	1
ATLFR	1	GVLRK	1	RRLDG	1
ATLHW	1	GVVRR	1	RRLHR	1
ATLIS	1	HALMR	1	RRLVR	1
ATLNH	1	HALQR	1	RRSDR	1
ATLNS	1	HILKR	1	RRVEK	1
ATLQG	1	HILQR	1	RTLER	1
ATLQW	1	HTLAR	1	RTLNR	1
ATLRI	1	HTLLR	1	RTLRG	1
ATLRP	1	HTLNR	1	SAVKR	1
ATLWR	1	HTLRG	1	SGLKR	1
ATSVR	1	HTLSR	1	SKLTR	1
ATVAR	1	HVLHR	1	SMLIR	1
AVLGR	1	KTLLR	1	SMLNR	1
AVLLR	1	KTLVR	1	SMSRR	1
AVLNR	1	MALRM	1	SNLIR	1
AVLTR	1	MPLAR	1	SPLRR	1
DKLRR	1	MPLNR	1	SRLRI	1
DMLKR	1	MPLVR	1	STLCR	1
DRLRA	1	MRS	1	STLER	1
DTLNR	1	MTLKR	1	STLKS	1
EPLVM	1	NRLRI	1	STLRI	1
ETLAR	1	NTLAR	1	STSRR	1
ETLQR	1	NTLHR	1	SVLRK	1
GALTR	1	NTLLR	1	TALIR	1
GDLRR	1	NTLQR	1	TALMR	1
GMLAR	1	NTLRW	1	TALTR	1
GNLMR	1	NTLTR	1	TMLQR	1
GPFKR	1	NTLVS	1	TMLRG	1
GPLGL	1	NTVRR	1	TNLIR	1
GPLKR	1	NVLKR	1	TNLKR	1
GSLGA	1	PPLKR	1	TPIMM	1
GSLQR	1	PSLKR	1	TPTRS	1
GSLTR	1	PTFHR	1	TRLRV	1
GTFRR	1	QKLA...	1	TRSP...	1
GTLDR	1	QMLKR	1	TTLGR	1
GTLER	1	QSLKR	1	TTLIR	1
GTLGR	1	QTLHR	1	TTLRS	1

TVLNR	1
VSLRR	1
VTLLR	1
VTLLQ	1
VVLGN	1
WRLDR	1
WTLRR	1

**Table 25. ZF3
selection on G:A
change at nt 13 of
core motif in CBS.
Sequences reflect
position 2 to 6.**

Sequence Read #

GQLTV 1094
GQLVV 906
GELVV 766
AELIV 643
TELIV 552
QELLV 528
GELIV 525
GELTV 505
GQLIV 476
QELLT 457
SELIV 416
GQLLV 372
SGLIV 372
GQLII 361
AELLV 311
VELLI 277
AELVV 271
AQLIV 267
SQLIV 265
TELII 251
QGLLV 247
SQLII 243
QQLLI 224
AGLIV 221
QELVV 209
GELLV 206
GELLT 202
SQLLV 199
GELVI 194
QQLIV 179
QELII 177
TQLIV 176
VELII 172
VELLV 160

GELLI 151
GQLLT 150
NELLI 149
GQLLI 148
SQLLI 140
AQLLV 136
GQLIT 132
GQLTI 129
TELIT 122
TELLI 118
TELLV 116
QELLI 112
AQLVV 106
GSLLV 104
AQLLI 102
HPPEE 100
SQLVV 100
QQLLV 98
QELIV 95
SELII 91
AQLII 90
QQLVV 90
TGLLV 88
NQLII 88
GQLVI 81
AGLLV 80
NQLLV 73
QELGV 69
GALVV 68
SQLTV 67
GELTT 67
GELII 65
SGLLV 63
AELII 60
TQLII 59
QQLII 59
AQLIT 58
SQLIT 58
SSLIV 57
SELTV 57
NELLV 57
TQLLV 56
QGLIV 55

QELVI 55
NELIV 55
TELLT 53

Table 26. ZF3
selection on G:T
change at nt 13 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence	Read #		
QQLLI	1145	LVLAD	95
GELTV	1108	EALRA	94
GQLTV	933	GELVI	87
GQLLI	748	STLKA	87
GQLVV	545	GQLVI	85
GQLII	518	AGLIV	82
GQLIV	479	GQLTL	82
GELVV	477	NVLGT	81
GELIT	438	KGLGP	79
SQLLV	432	MQLRR	79
SQLI	431	GDLQR	75
GQLLT	404	VLLPN	71
SQLSM	365	MRLGD	69
QQLLT	349	GQLAQ	67
QQLIV	312	NELRG	67
HPPEE	308	GELTT	66
QELLT	300	GELVT	64
SQLLI	282	STLVV	63
GKLNA	281	VDLAV	61
GELIV	263	AQLTI	59
GQLLV	225	DALPA	57
SQIV	219	SVLQL	57
RALLI	216	GPLGN	56
ENLLI	201	GHLLL	52
AQLLV	174	DVLDP	51
PDLKR	174	SSLSI	50
GELLT	172	KMLAD	50
AQLIT	164		
GQLVT	138		
GQLLS	116		
QELLV	112		
GELNP	112		
GQLIQ	98		
PTLVG	98		

Table 27. ZF3
selection on G:C
change at nt 13 of
core motif in CBS.
Sequences reflect
position 2 to 6.

Sequence	# Reads		
RKHD	4641	GAPT	17
RKAD	1938	HSPN	17
RRSD	1299	RPVA	16
RRHD	868	RKDD	6
RKTD	182	MLVD	4
NVSM	146	RHRK	3
RQSD	76	RKHV	3
RKND	69	RKQD	3
SENV	69	RKSD	3
VDHR	60	DHHT	2
AQIV	58	GKHD	2
KTPH	56	MKAD	2
PKIV	51	RKAE	2
GAEP	42	RRAD	2
MLVE	40	APIG	1
VVGN	40	AQNR	1
KGPE	36	DMDA	1
GKVM	33	EAPM	1
TEPG	33	EEMM	1
TPHN	32	EPIR	1
MPGG	31	GALE	1
DLEK	28	GENV	1
GTDN	27	GKAD	1
ISRL	25	GKVD	1
ATGL	21	GPLA	1
ASNP	19	GRIE	1
		IEKL	1
		KAAS	1
		KEEH	1
		LKVD	1
		LLVE	1
		LMTQ	1
		MASL	1
		MGIG	1
		MPGD	1
		MSLG	1
		NDMT	1
		NMHT	1
		NRIV	1
		PENA	1
		QKHD	1
		QVPD	1
		RASD	1
		REHD	1
		RGHD	1
		RKHA	1
		RKHY	1
		RKLD	1
		RKPD	1
		RKVD	1
		RKYD	1
		RMSD	1
		RRLD	1
		RRND	1
		RRRD	1
		RRSG	1
		RWHD	1
		SHRL	1
		SQHV	1
		SSHD	1
		TTHV	1
		VHHV	1
		WKAD	1
		WKHD	1

REFERENCES

1. Ong, Chin-Tong & Corces, V.P., *Nat Rev Genet*. 2014 Apr;15(4):234-46.
2. Phillips, J. & Corces, V.P., *Cell*. 2009 Jun 26; 137(7): 1194–1211.
3. Ali, T. et al., *Curr Opin Genet Dev*. 2016 Apr;37:17-26.
- 5 4. Nora, E. P. et al., *Nature*. 2012 Apr 11;485(7398):381-5.
5. Rao, S. S. et al., *Cell*. 2014 Dec 18; 159(7): 1665-1680.
6. Phillip, J., et al., *Cell*. 2013 Jun 6; 153(6): 1281–1295.
7. Shukla, S., et al., *Nature*. 2011 Nov 3;479(7371):74-9.
8. Hilmi, K., et al. *Sci Adv*. 2017 May 24;3(5):e1601898.
- 10 9. Han, D., et al. *Sci Rep*. 2017 Mar 6;7:43530.
10. Rhee, S., & Pugh, F.B., *Cell*. 2011 Dec 9;147(6):1408-19.
11. Nakahashi, H., et al., *Cell Rep*. 2013 May 30;3(5):1678-1689.
12. Hashimoto, et al., *Mol Cell*. 2017 Jun 1;66(5):711-720.e3.
13. Guo, A. et al., *Nat Commun*. 2018 Apr 18;9(1):1520.
- 15 14. Schuijers, J. et al., *Cell Reports* (2018). *Cell Rep*. 2018 Apr 10;23(2):349-360.
15. Kang, J.Y. et al., *Oncogene*. 2015 Nov 5;34(45):5677-84.
16. Wright, D., et al. *Nat Protoc*. 2006;1(3):1637-52.
17. Sander, J., et al. *Nat Methods*. 2011 Jan;8(1):67-9.
18. Maeder, M., et al. *Mol Cell*. 2008 Jul 25;31(2):294-301.
- 20 19. Joung J. K. et al., *Proc Natl Acad Sci U S A*. 2000 Jun 20;97(13):7382-7.

OTHER EMBODIMENTS

It is to be understood that while the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended
5 claims. Other aspects, advantages, and modifications are within the scope of the following claims.

CLAIMS

WHAT IS CLAIMED IS:

1. An engineered CCCTC-binding factor (CTCF) variant comprising at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF, wherein the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS) with a higher affinity than wild-type CTCF, the mutant
5 CBS comprising at least one nucleotide base that differs in sequence from the nucleotide sequence of a consensus CBS, wherein the at least one amino acid residue that differs in sequence from the amino acid sequence of the wild-type CTCF is selected from the group consisting of amino acid residues at position(s) -1, +1, +2, +3, +5, and +6 of any of ZF7, ZF6, ZF5, ZF4, and ZF3 of the engineered CTCF variant.
10
2. The engineered CTCF variant of claim 1, wherein the mutant CBS has a Thymine (T), Adenine (A), or Guanine (G) residue at position 2 of the consensus CBS motif, the engineered CTCF comprising an amino acid residue threonine, asparagine, or histidine at ZF7 position +3.
15
3. The engineered CTCF variant of claim 1, wherein the mutant CBS has a G residue at position 2 of the consensus CBS motif, the engineered CTCF comprising the amino acid sequence DHLQT, EHLNV, AHLQV, EHLRE, DHLQV, EHLKV, EHLVV, DHLRT, or DHLAT at ZF7 positions +2 to +6.
20
4. The engineered CTCF variant of claim 1, wherein the mutant CBS has a T, G, or C residue at position 3 of the consensus CBS motif, the engineered CTCF at ZF7 positions -1 to +3 comprising:
the amino acid sequence RKHD or RRSD, wherein the mutant CBS has a T
25 residue at position 3 of the consensus CBS motif;
the amino acid sequence RKAD, IPRI, RKHD, or RKDD, wherein the mutant CBS has a G residue at position 3 of the consensus CBS motif; or

the amino acid sequence GIVN, ELLN, QALL or PHRM, wherein the mutant CBS has a C residue at position 3 of the consensus CBS motif.

5. The engineered CTCF variant of claim 1, wherein the mutant CBS has a T, G, or A residue at position 5 of the consensus CBS motif, the engineered CTCF at ZF6 positions +2 to +6 comprising:

the amino acid sequence NAMKR, GNMAR, EGMTR, SNMVR, or NAMRG, wherein the mutant CBS has a T residue at position 5 of the consensus CBS motif; or

- the amino acid sequence EHMGR, DHMNR, THMKR, EHMRR, or THMNR, wherein the mutant CBS has a G residue at position 5 of the consensus CBS motif.

6. The engineered CTCF variant of claim 1, wherein the mutant CBS has a T, G, or C residue at position 6 of the consensus CBS motif, the engineered CTCF at ZF6 positions -1 to +3 comprising:

- the amino acid sequence MNES or HRES, wherein the mutant CBS has a T residue at position 6 of the consensus CBS motif; or

the amino acid sequence RPDT, RTDI, or RHDT, wherein the mutant CBS has a G residue at position 6 of the consensus CBS motif.

7. The engineered CTCF variant of claim 1, wherein the mutant CBS has a C, A, or T residue at position 7 of the consensus CBS motif, the engineered CTCF at ZF5 positions +2 to +6 comprising:

the amino acid sequence HGLKV, HRLKE, HALKV, SRLKE, or DGLRV, wherein the mutant CBS has a T residue at position 7 of the consensus CBS motif;

- the amino acid sequence HTLKV, or HGLKV, wherein the mutant CBS has an A residue at position 7 of the consensus CBS motif; or

the amino acid sequence SRLKE, HRLKE or NRLKE, wherein the mutant CBS has a C residue at position 7 of the consensus CBS motif.

8. The engineered CTCF variant of claim 1, wherein the mutant CBS has a C, A, or T residue at position 8 of the consensus CBS motif, the engineered CTCF at ZF5

positions +2 to +6 comprising:

the amino acid sequence ATLKR, QALRR, GGLVR, or HGLIR, wherein the mutant CBS has a T residue at position 8 of the consensus CBS motif;

the amino acid sequence ANLSR, TGLTR, HGLVR, or GGLTR, wherein the mutant CBS has an A residue at position 8 of the consensus CBS motif;

the amino acid sequence HTLRR, TVLKR, ADLKR, or HGLRR, wherein the mutant CBS has a C residue at position 8 of the consensus CBS motif.

9. The engineered CTCF variant of claim 1, wherein the mutant CBS has a T, A, or C residue at position 10 of the consensus CBS motif, the engineered CTCF at ZF4 positions +2 to +6 comprising:

the amino acid sequence AHLRK, wherein the mutant CBS has a T residue at position 10 of the consensus CBS motif;

the amino acid sequence AKLRV, EKLRI, or AKLRI, wherein the mutant CBS has an A residue at position 10 of the consensus CBS motif; or

the amino acid sequence TKLKV, wherein the mutant CBS has a C residue at position 10 of the consensus CBS motif.

10. The engineered CTCF variant of claim 1, wherein the mutant CBS has a T, A, or C residue at position 11 of the consensus CBS motif, the engineered CTCF at ZF4 positions +2 to +6 comprising:

the amino acid sequence ATLRR or RRLDR, wherein the mutant CBS has a T residue at position 11 of the consensus CBS motif;

the amino acid sequence TNLRR, ANLRR, or GNLTR, wherein the mutant CBS has an A residue at position 11 of the consensus CBS motif; or

the amino acid sequence AMLKR, HMLTR, AMLRR, or TMLRR, wherein the mutant CBS has a C residue at position 11 of the consensus CBS motif.

11. The engineered CTCF variant of claim 1, wherein the mutant CBS has a T, A, or C residue at position 13 of the consensus CBS motif, the engineered CTCF at ZF3 positions +2 to +6 comprising:

the amino acid sequence QQLIV, SQLIV, QQLLV, GELVV, or QQLLI, wherein the mutant CBS has a T residue at position 13 of the consensus CBS motif;

the amino acid sequence GQLIV, GQLTV, GKLVT, TELII or QGLLV, wherein the mutant CBS has an A residue at position 13 of the consensus CBS motif; or

5 the amino acid sequence QQLLT, GQLLT, GELLT, or QQLLI, wherein the mutant CBS has a C residue at position 13 of the consensus CBS motif.

12. The engineered CTCF variant of claim 1, wherein the mutant CBS has A, G, T, and T residues at positions 2, 6, 7, and 10 of the consensus CBS motif, respectively, the
10 engineered CTCF comprising:

(i) the amino acid sequence AKLKK, AKLRK, AHLRV, AKLRV, or SKLRL at ZF4 positions +2 to +6 of the engineered CTCF;

(ii) the amino acid sequence ERLRV, NRLKV, SRLKE, or NRLKV at ZF5 positions +2 to +6 of the engineered CTCF;

15 (iii) the amino acid sequence RPDT, RTET, or RADV at ZF6 positions -1 to +3 of the engineered CTCF; and

(iv) the amino acid sequence DNLLA, SNLLV, DNLMA, or DNLRV at ZF7 positions +2 to +6 of the engineered CTCF.

20 13. The engineered CTCF variant of claim 1, wherein the mutant CBS has G, G, T, and T residues at positions 2, 6, 7, and 10 of the consensus CBS motif, respectively, the engineered CTCF comprising:

(i) the amino acid sequence GHLKK, AHLRK, or GKLRI at ZF4 positions +2 to +6 of the engineered CTCF;

25 (ii) the amino acid sequence SRLKE, DALRR, DGLKR, or TRLRE at ZF5 positions +2 to +6 of the engineered CTCF;

(iii) the amino acid sequence at RPDTMKR or RTENMKM at ZF6 positions -1 to +6 of the engineered CTCF; and

30 (iv) the amino acid sequence EHLKV, DHLLA, or HHLDV at ZF7 positions +2 to +6 of the engineered CTCF.

14. The engineered CTCF variant of claim 1, wherein the mutant CBS has A, G, and A residues at positions 2, 5, and 11 of the consensus CBS motif, respectively, the engineered CTCF comprising:

- (i) the amino acid sequence SNLRR, GNLVR, GNLRR, GNLKR, ANLRR, NNLRR, or TNLRR at ZF4 positions +2 to +6 of the engineered CTCF;
- (ii) the amino acid sequence EHMKR, EHMRR, THMKR, EHMNR, or EHMAR at ZF6 positions +2 to +6 of the engineered CTCF; and
- (iii) the amino acid sequence DNLLT, DNLLV, DNLQT, DNLLA, DNLAT, DNLQA, DNLMA, or DNLMT at ZF7 positions +2 to +6 of the engineered CTCF.

10

15. The engineered CTCF variant of claim 1, wherein the mutant CBS has G, G, and A residues at positions 2, 5, and 11 of the consensus CBS motif, respectively, the engineered CTCF comprising:

- (i) the amino acid sequence GNLVR, GNLRR, GNLAR, GNLMR, ANLRR, SNLRR, or NNLRR at ZF4 positions +2 to +6 of the engineered CTCF;
- (ii) the amino acid sequence EHMNR, EHMKR, EHMRR, SHMNR, SHMRR, THMKR, or DHMNR at ZF6 positions +2 to +6 of the engineered CTCF; and
- (iii) the amino acid sequence EHLKV, EHLAE, STLNE, DHLQV, EHLNV, DHLNT, EHLQA, or HHLMH at ZF7 positions +2 to +6 of the engineered CTCF.

20

16. The engineered CTCF variant of claim 1, wherein the mutant CBS that has G, T, and T residues at positions 6, 7, and 10 of the consensus CBS motif, respectively, the engineered CTCF comprising:

- (i) the amino acid sequence GHLKK, AHLKK, TKLRL, TKLKL, GHLRK, THLKK, or AHLRK at ZF4 positions +2 to +6 of the engineered CTCF;
- (ii) the amino acid sequence TRLKE or SRLKE at ZF5 positions +2 to +6 of the engineered CTCF; and
- (iii) the amino acid sequence RADN, RHDT, RRDT, RPDT, RTSS, or RNDT at ZF6 positions -1 to +3 of the engineered CTCF.

30

17. The engineered CTCF variant of any one of claims 1-16, wherein the engineered CTCF variant interacts with cohesion to mediate the formation of an enhancer-promoter loop to modulate gene expression.

5 18. A method of treating a subject in need thereof, the method comprising administering to the subject a therapeutically effective amount of an engineered CTCF variant according to any one of claims 1-17.

19. The method of claim 18, wherein the subject has cancer.

10

20. A method of activating or repressing expression of a gene under the control of a mutant CBS of any one of claims 1-17, the gene being aberrantly expressed under the control of the mutant CBS, the method comprising contacting the mutant CBS with an engineered CTCF according to any one of claims 1-17, thereby regulating the expression
15 of the gene.

21. The method of claim 20, wherein the engineered CTCF activates or represses expression of the gene by interacting with cohesion to mediate the formation of an enhancer-promoter loop.

20

22. A pharmaceutical composition comprising an engineered CTCF variant according to any one of claims 1-17.

23. A gene expression system for regulation of a gene, the system comprising a
25 nucleic acid encoding an engineered CTCF variant according to any one of claims 1-17.

24. A method of altering the structure of chromatin comprising contacting an engineered CTCF variant according to any one of claims 1-17 with a mutant CBS to form a binding complex, such that the structure of the chromatin is altered.

30

25. A method of modulating expression of a gene that is under the control of a CBS bearing one or more mutations, the method comprising contacting the CBS bearing one or more mutations with an engineered CTCF variant according to any one of claims 1-17.
- 5 26. A kit comprising an engineered CTCF variant according to any one of claims 1-17 and instructions for use in a method described herein.

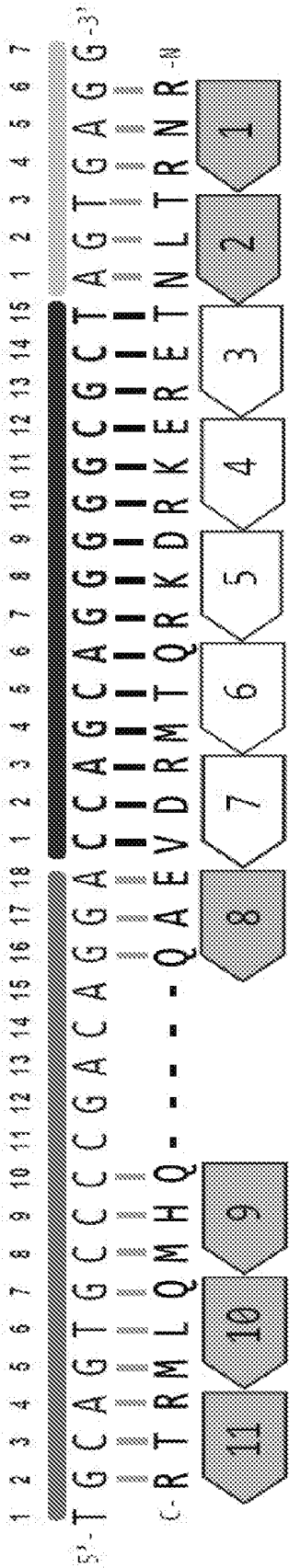


FIG. 1

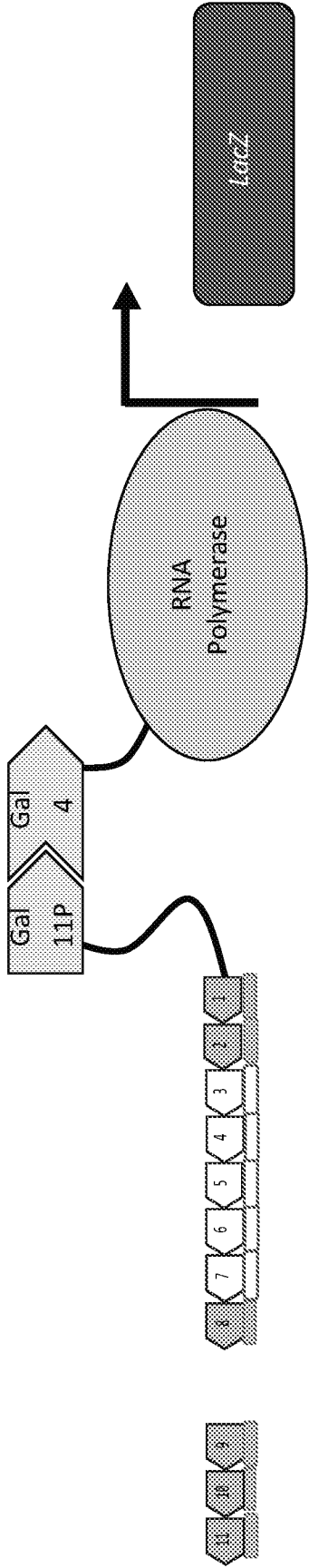


FIG. 2

2/47

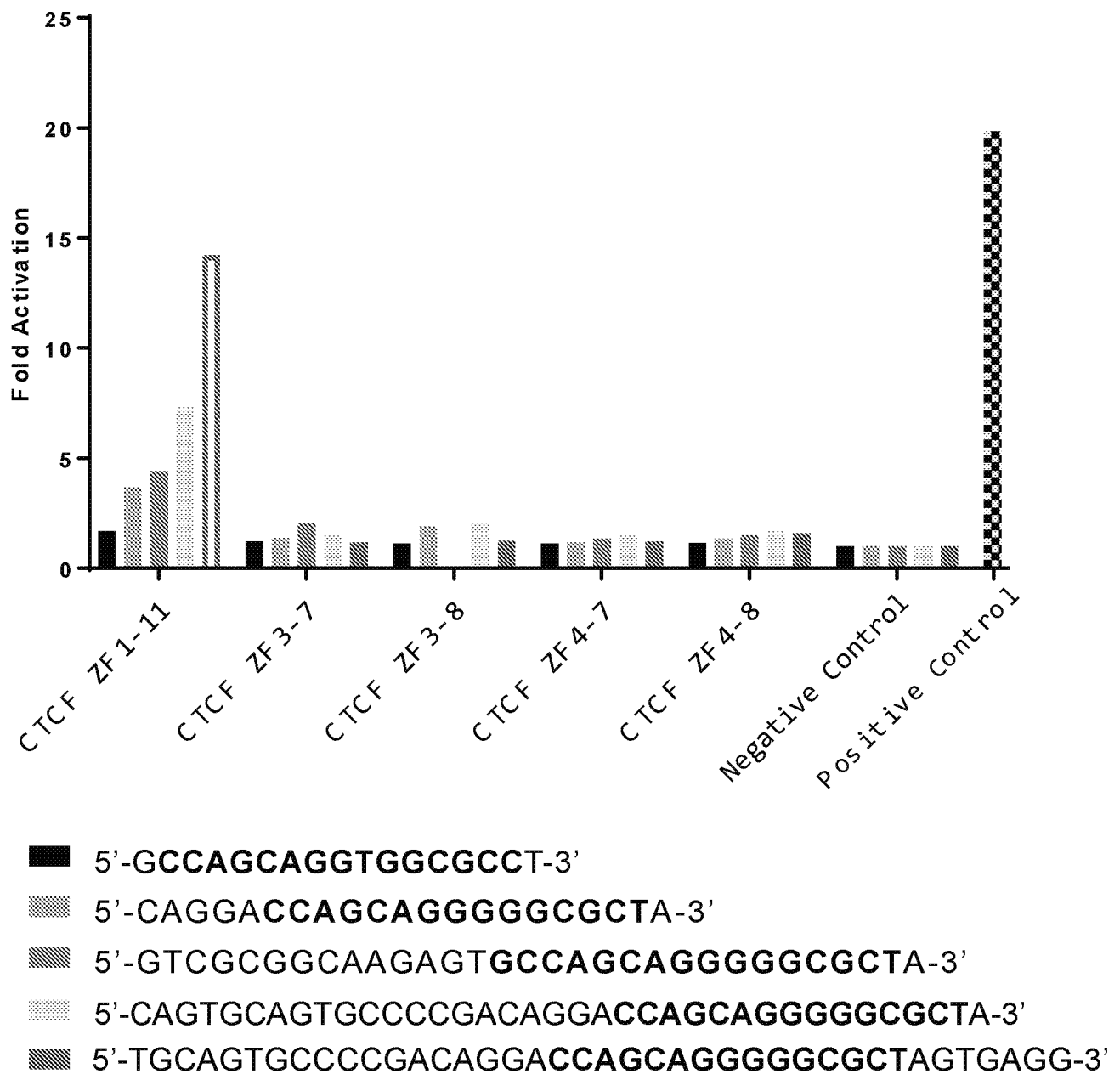
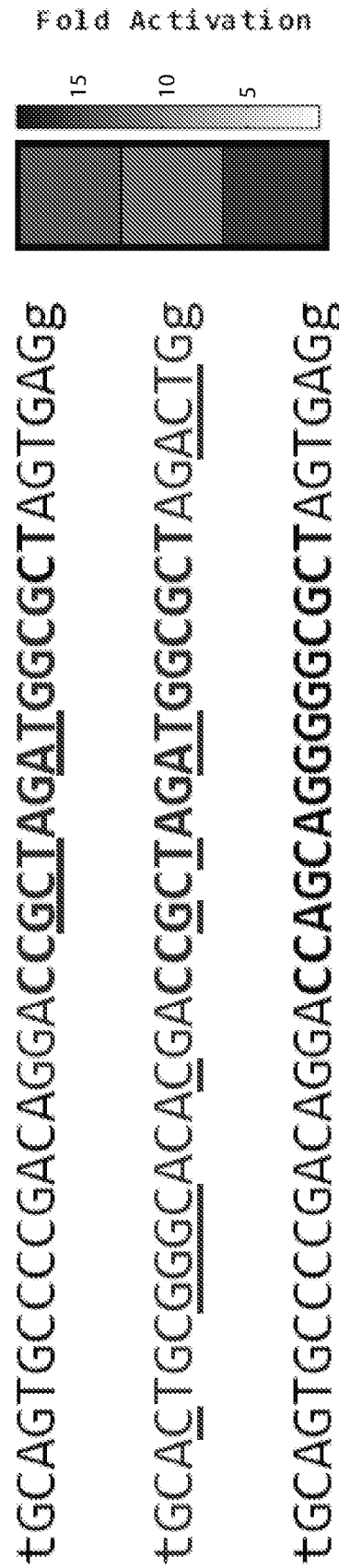
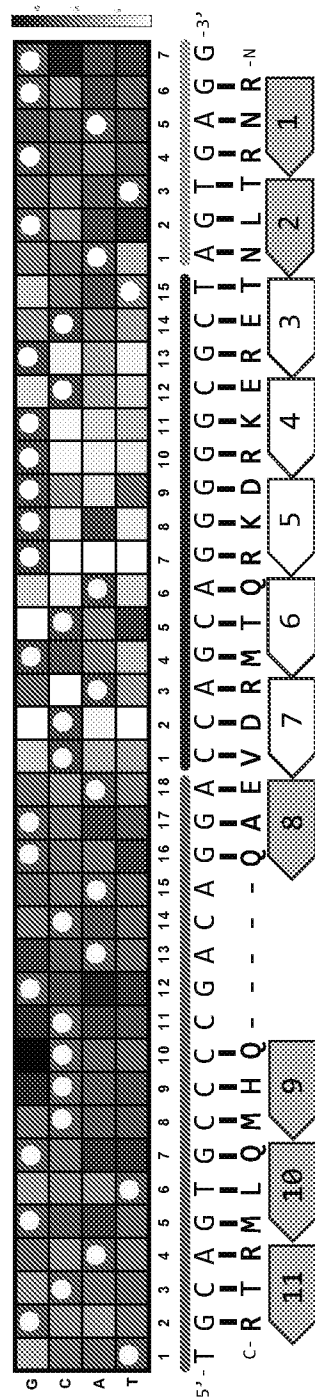


FIG. 3





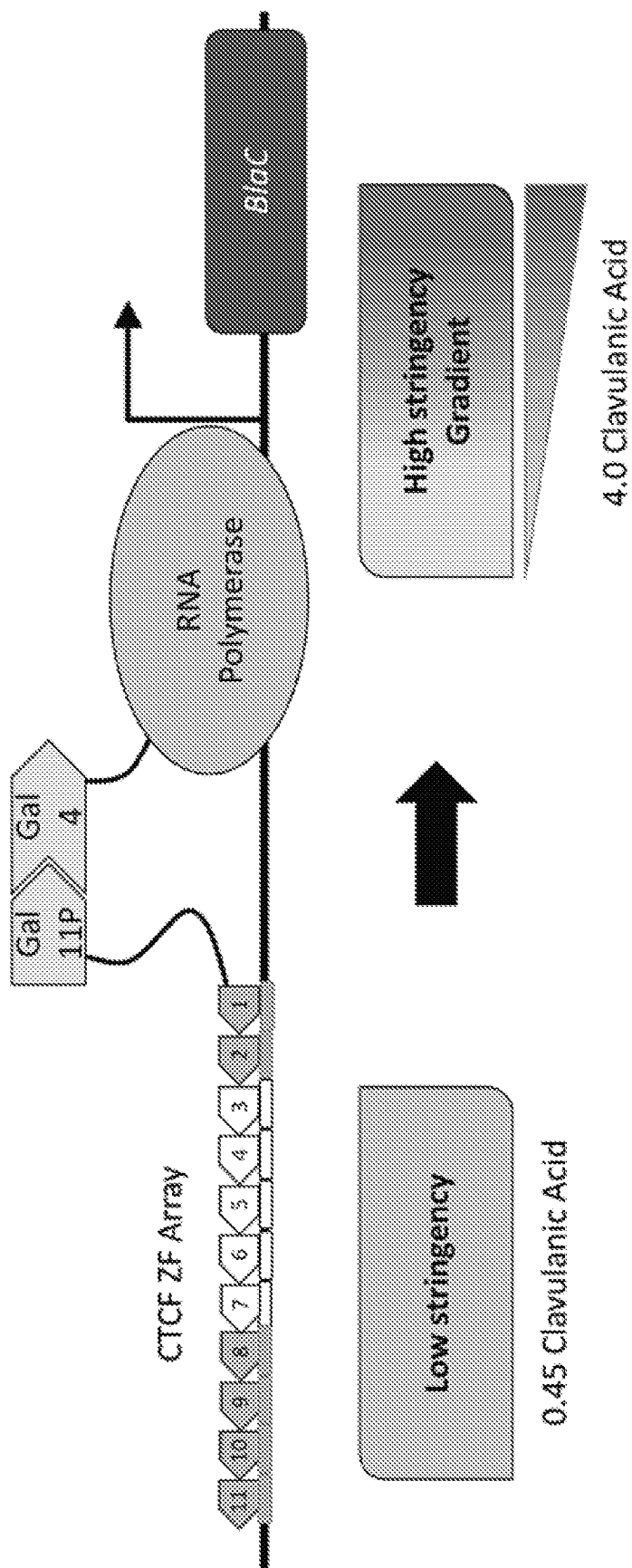


FIG. 6

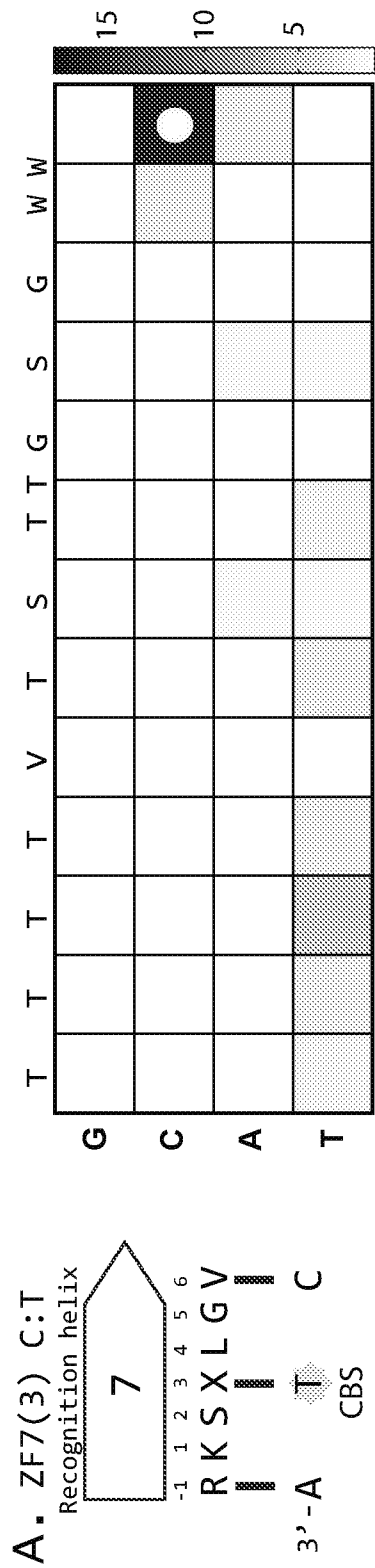


FIG. 7A

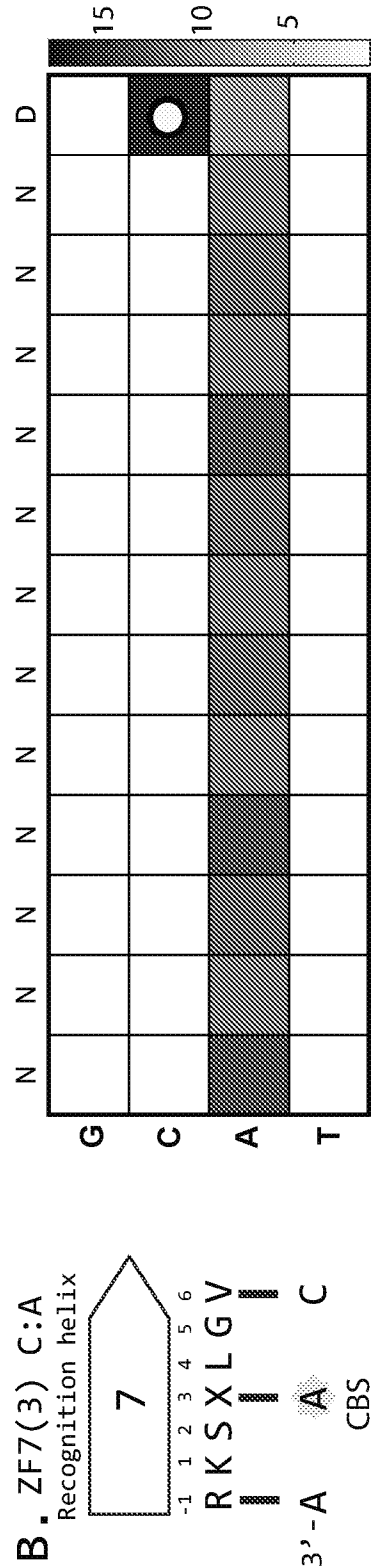


FIG. 7B

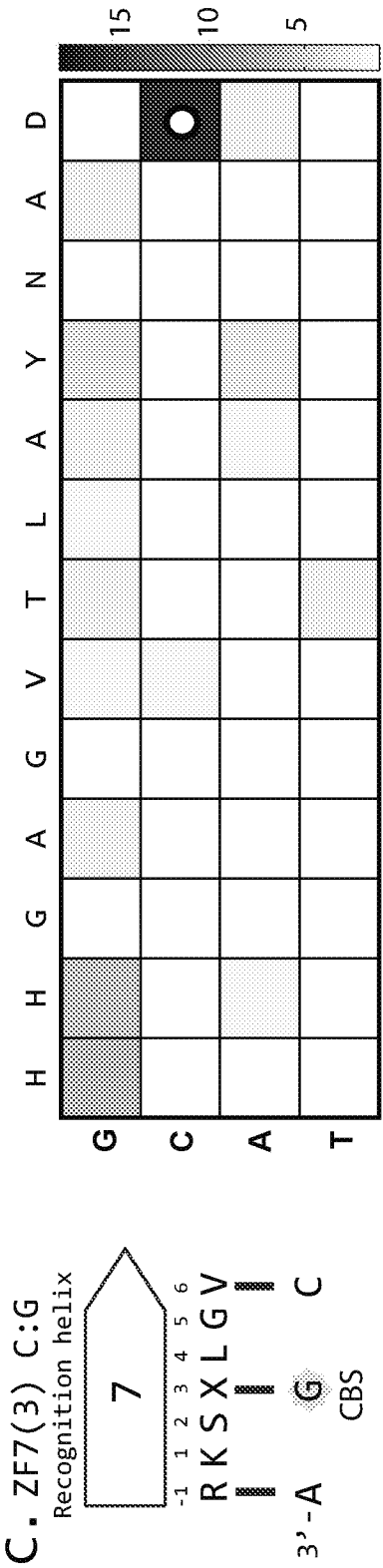


FIG. 7C

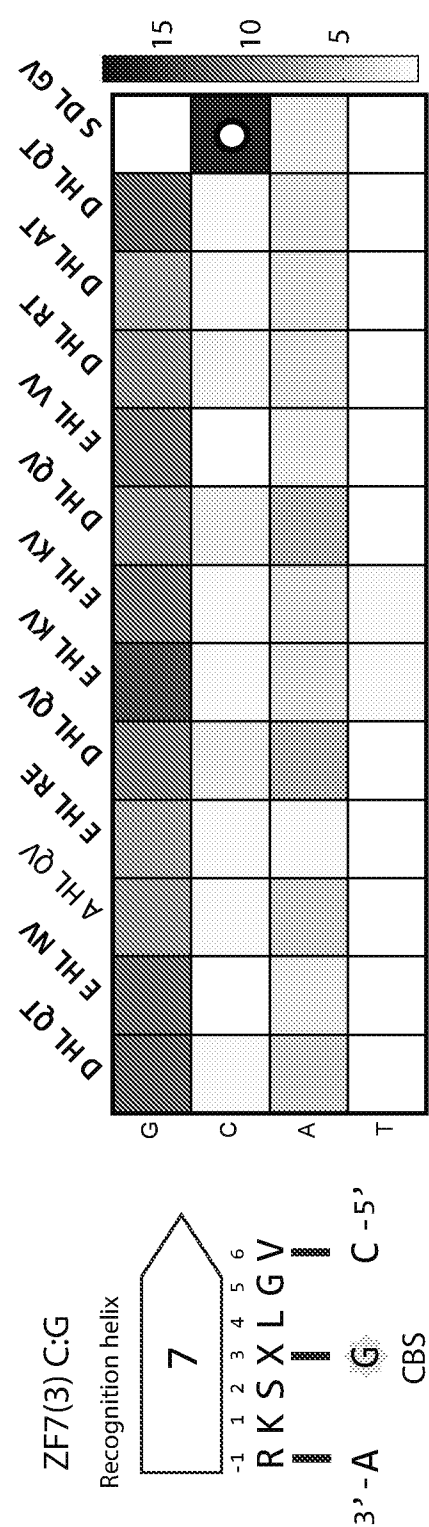


FIG. 8

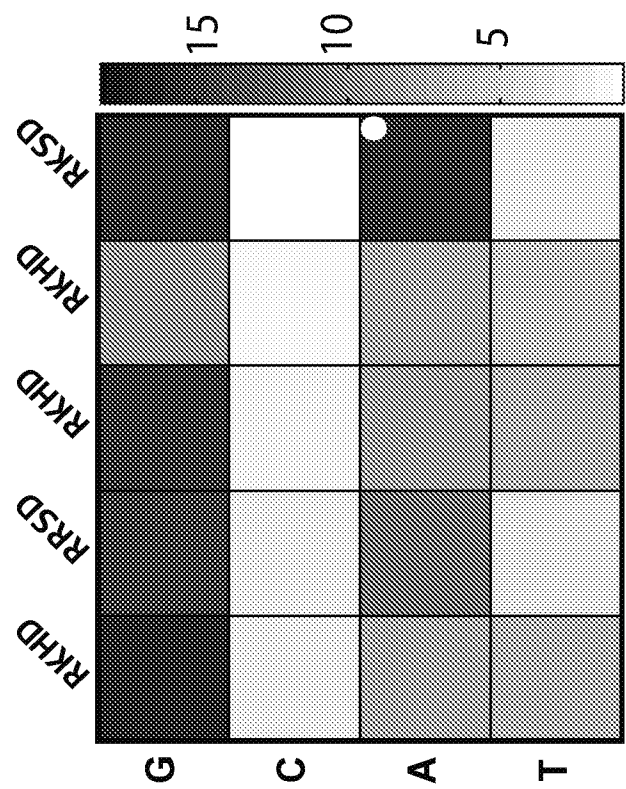
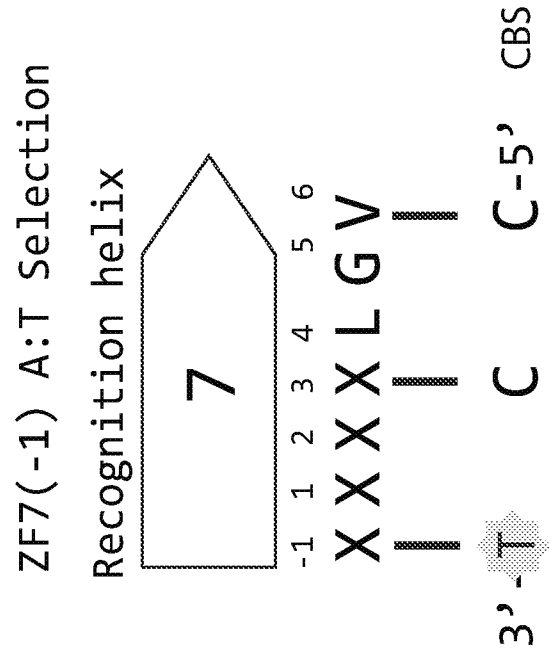


FIG. 9A



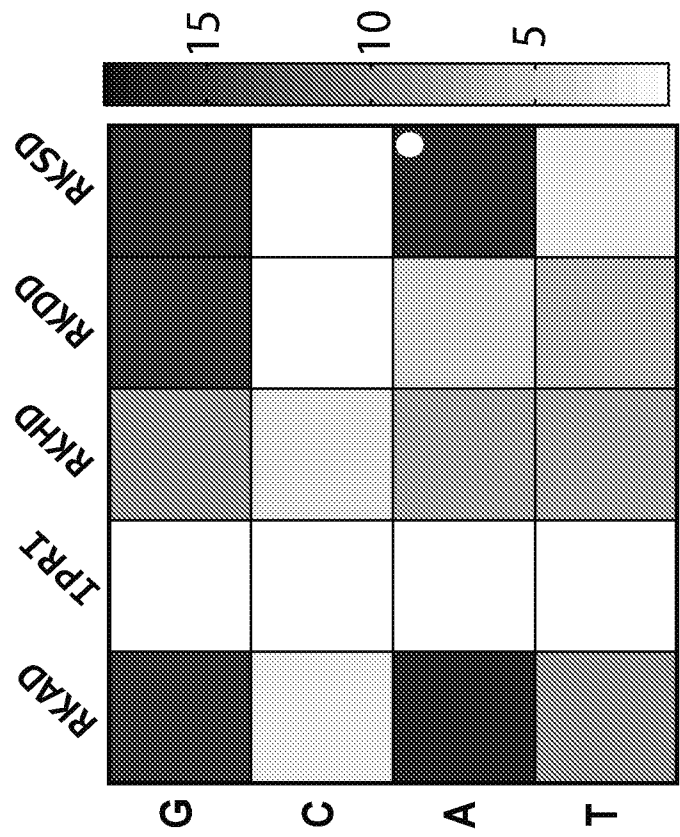
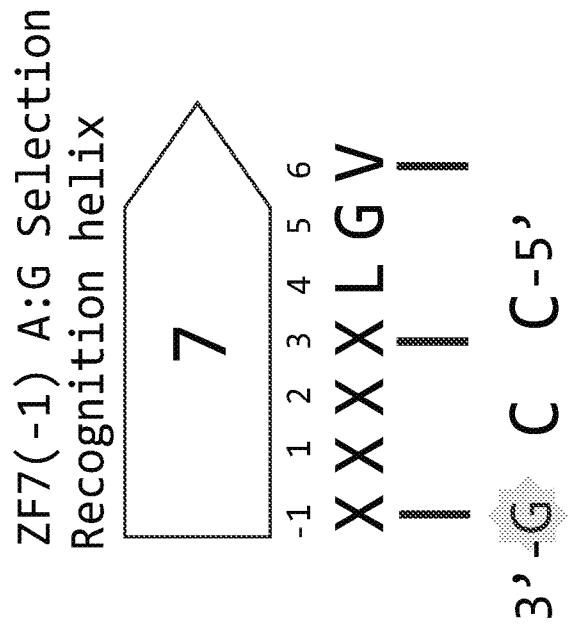


FIG. 9B



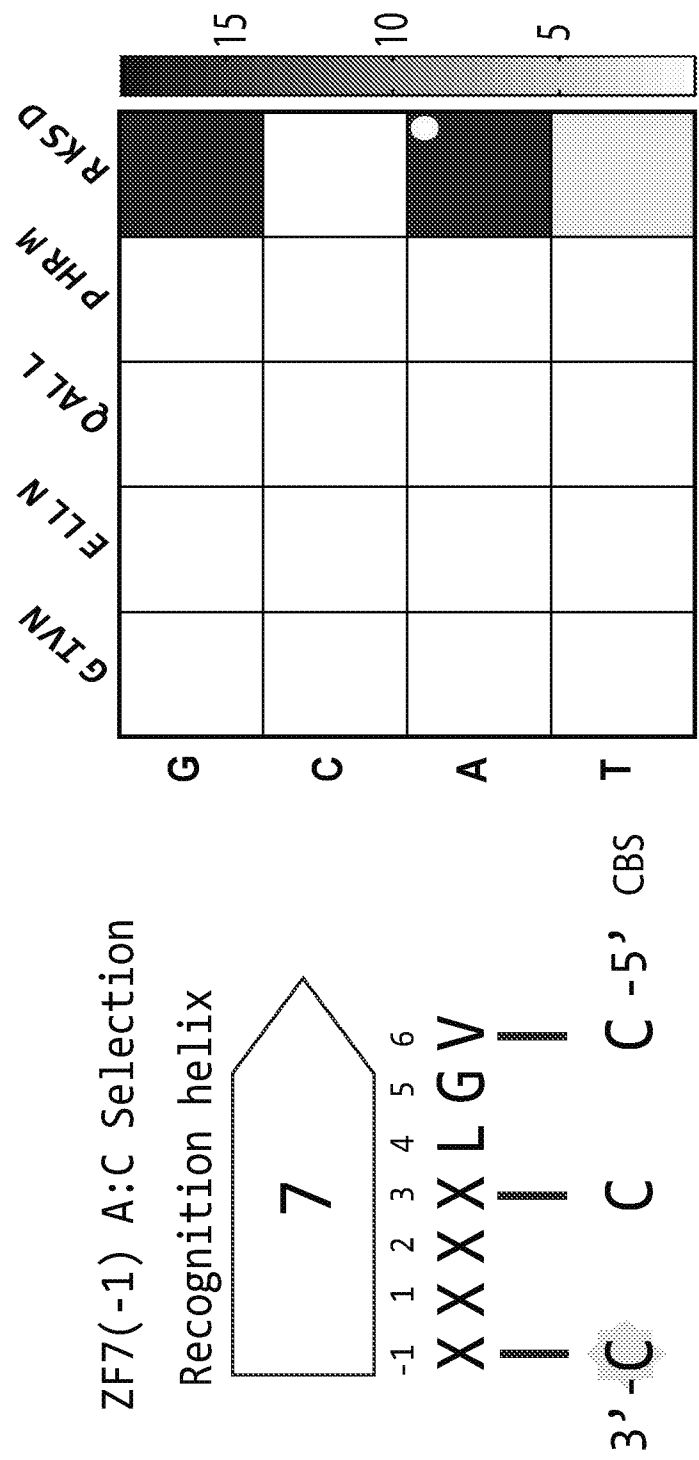


FIG. 9C

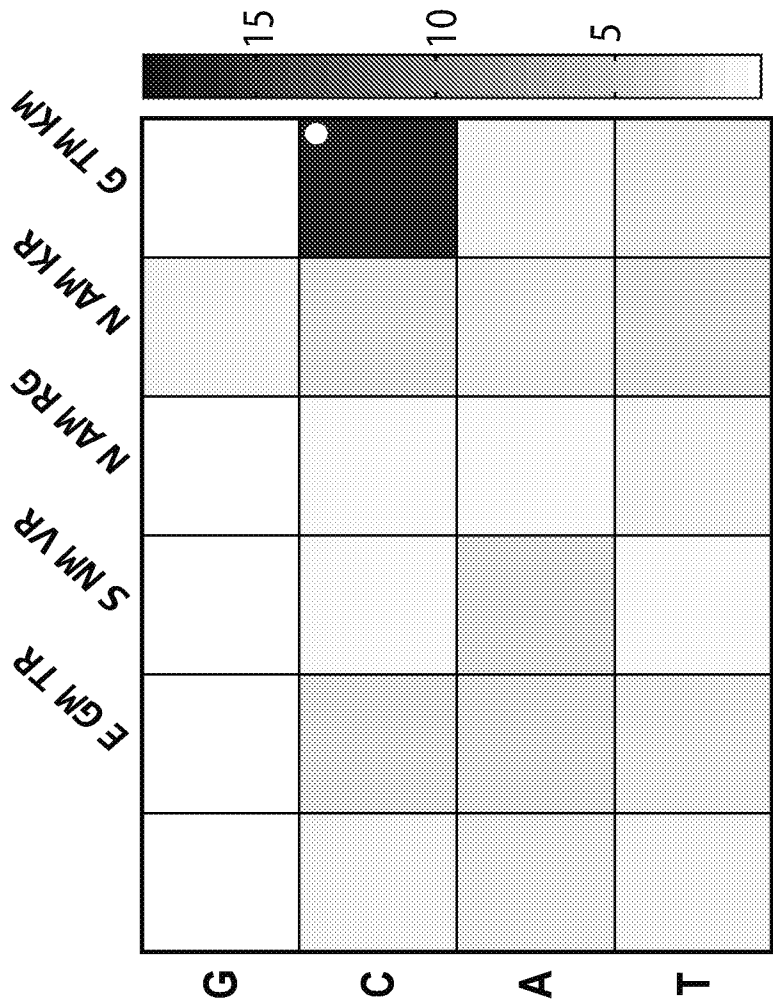
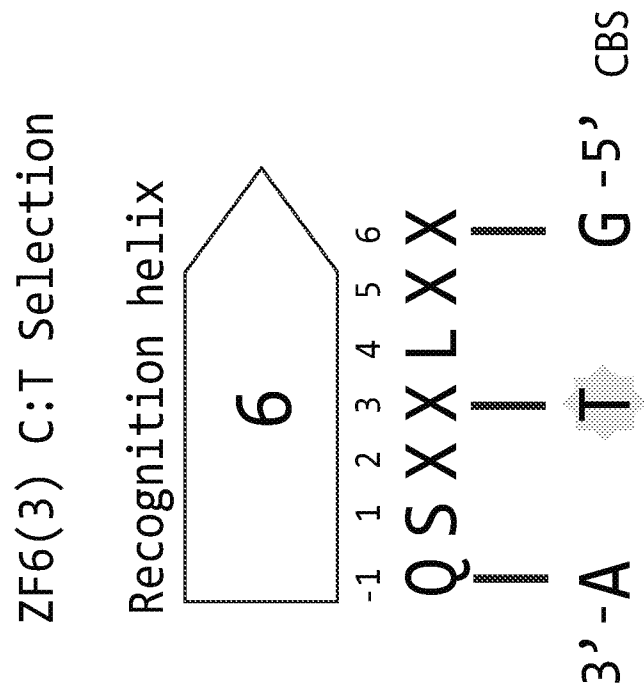


FIG. 10A



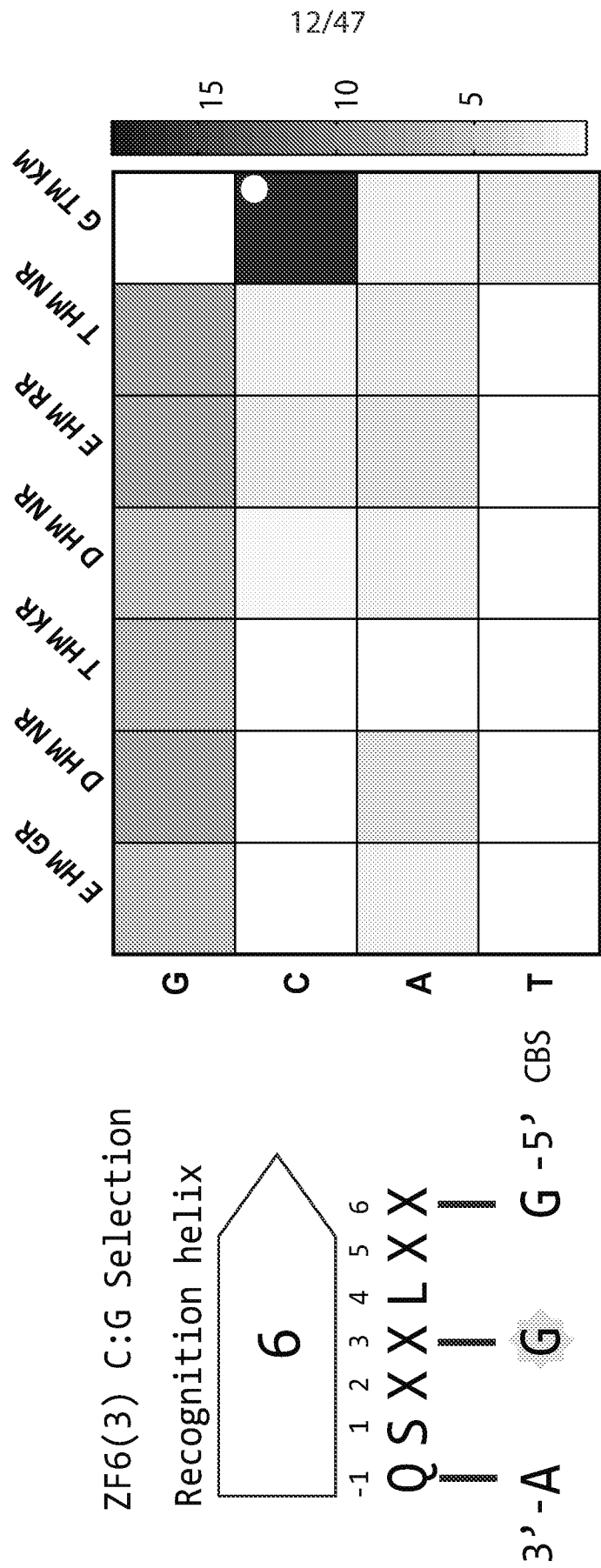


FIG. 10B

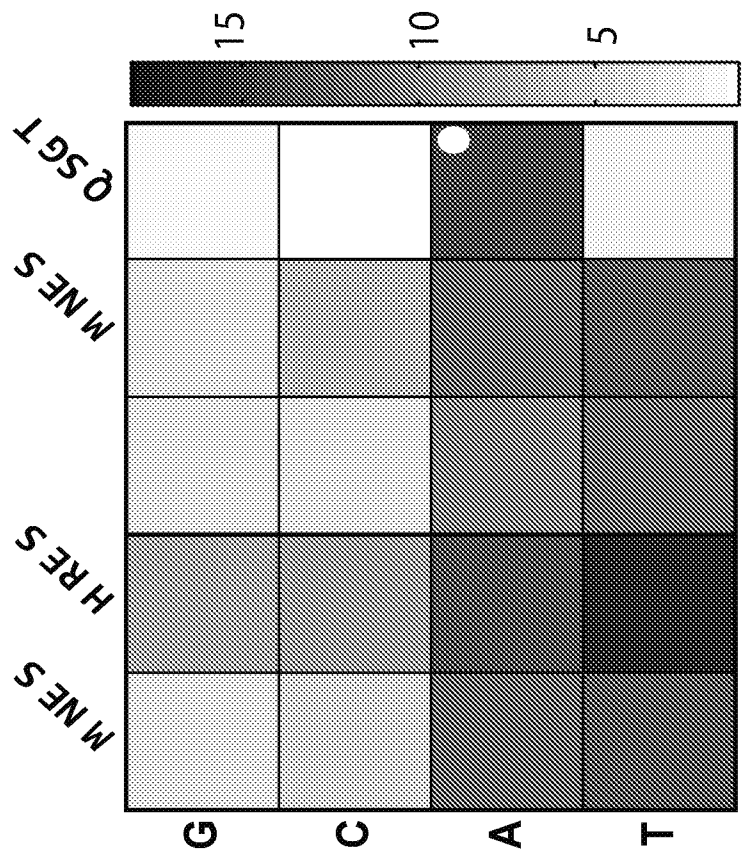
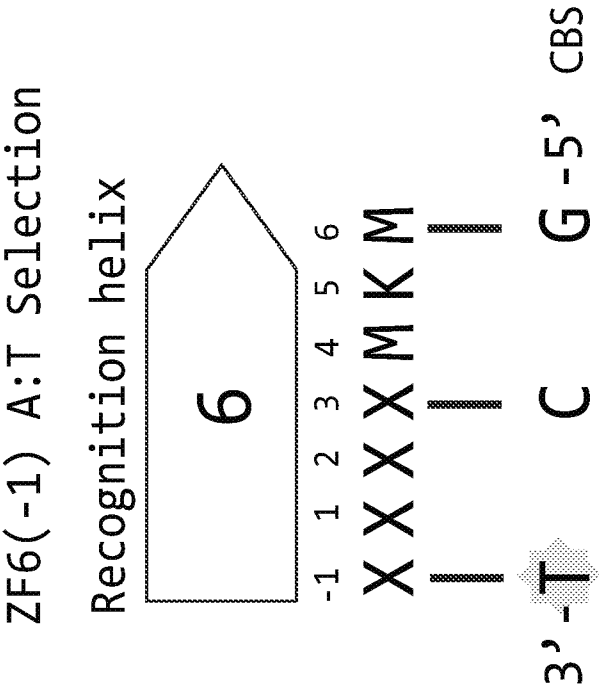


FIG. 11A



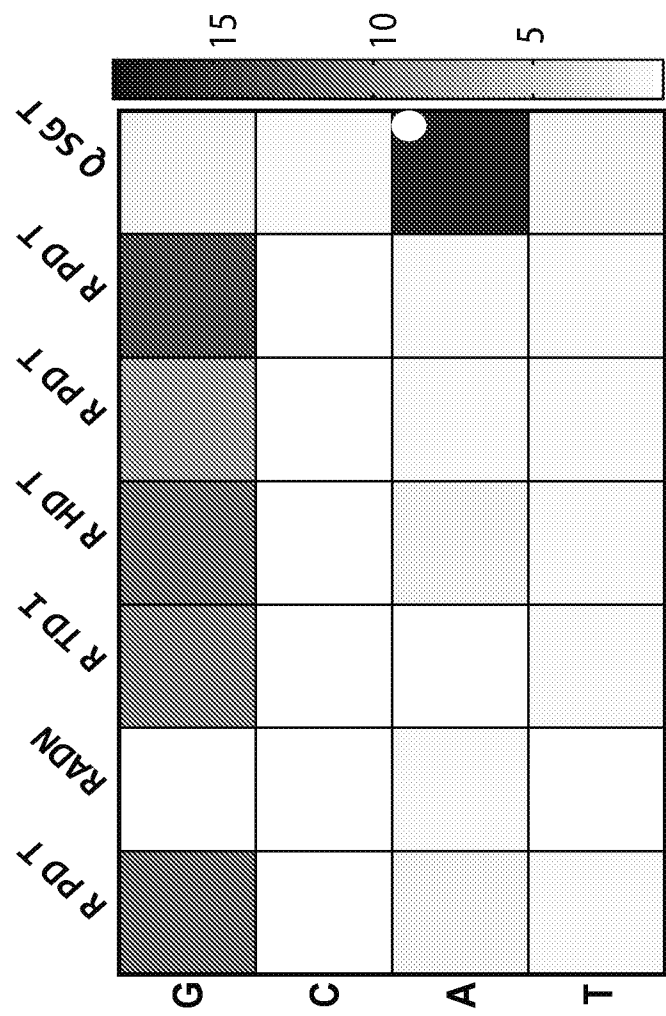
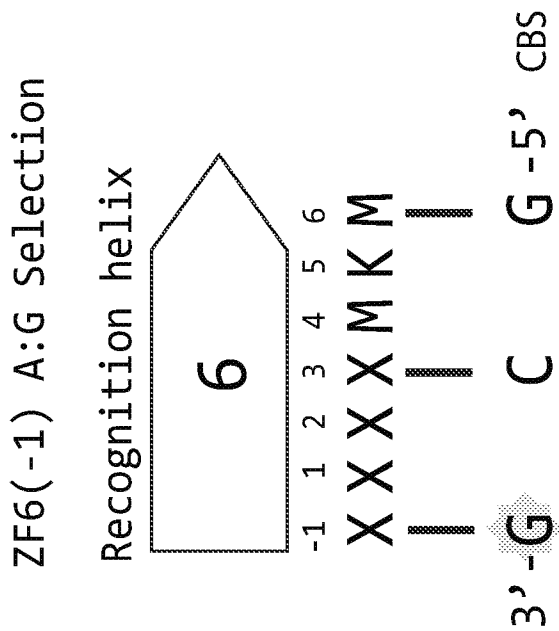


FIG. 11B



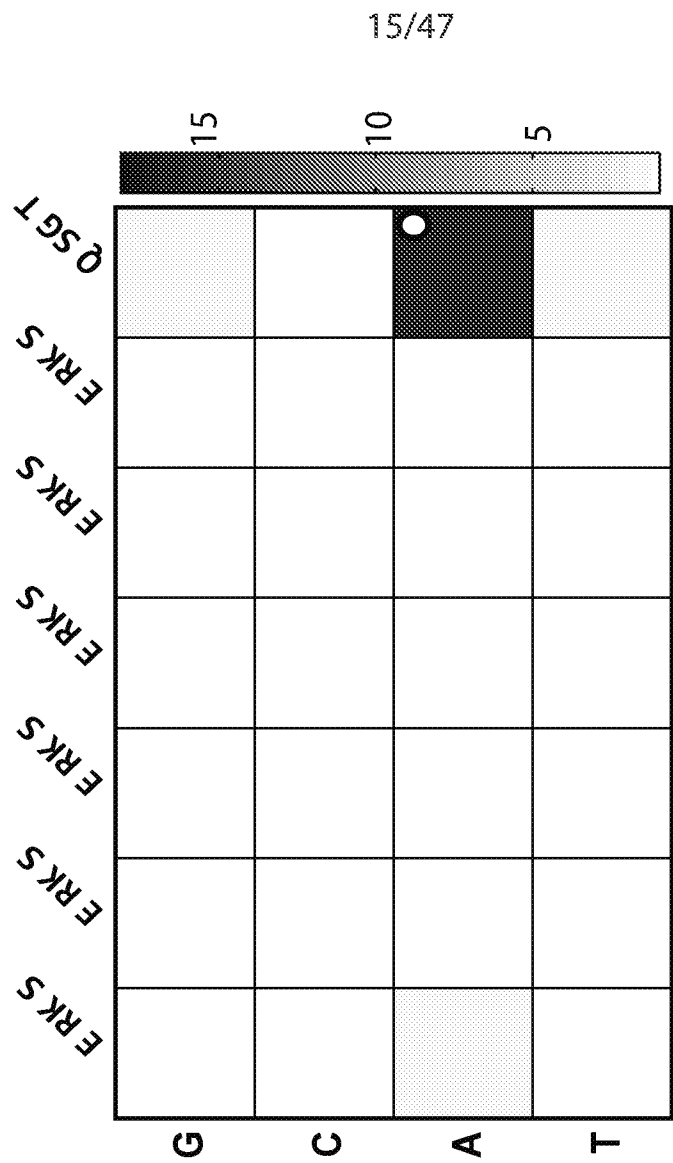
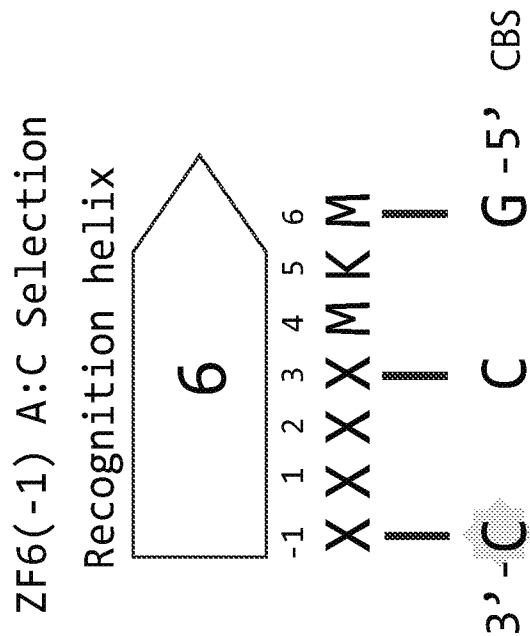


FIG. 11C



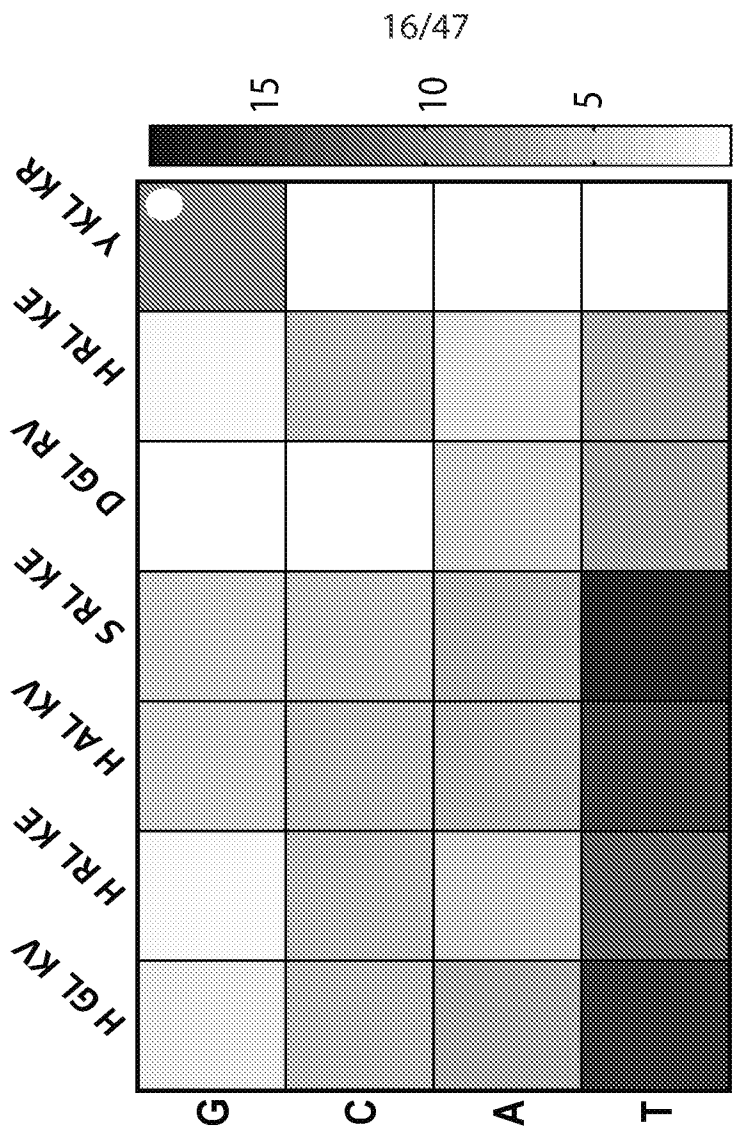
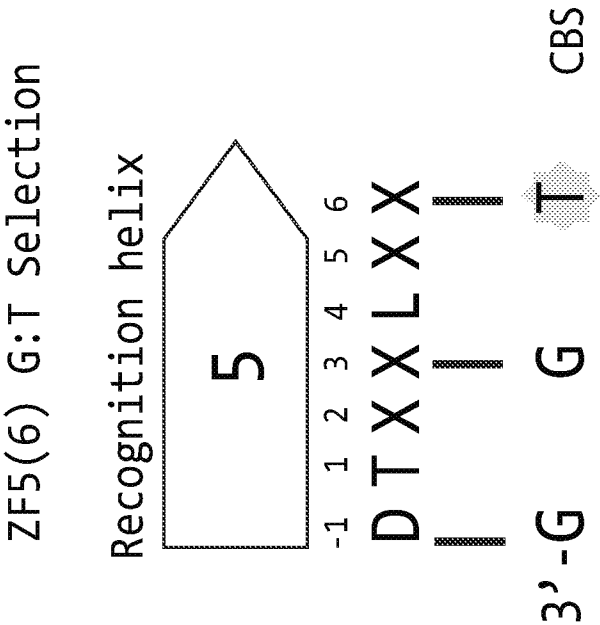


FIG. 12A



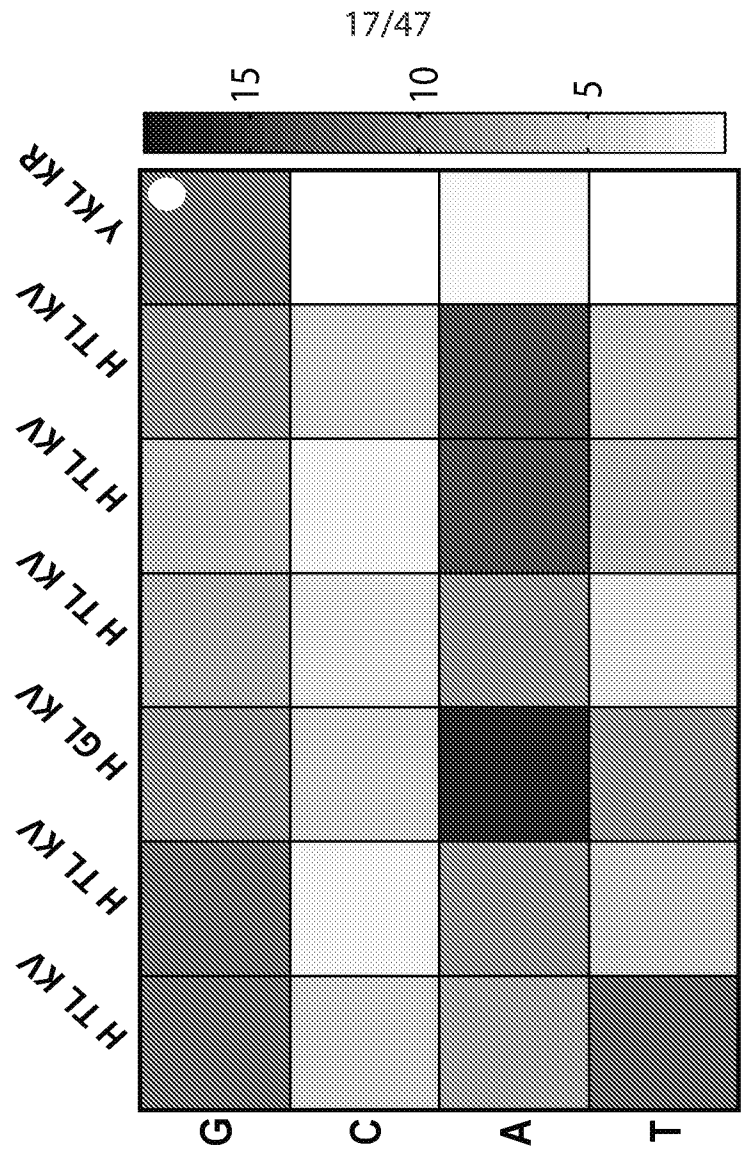
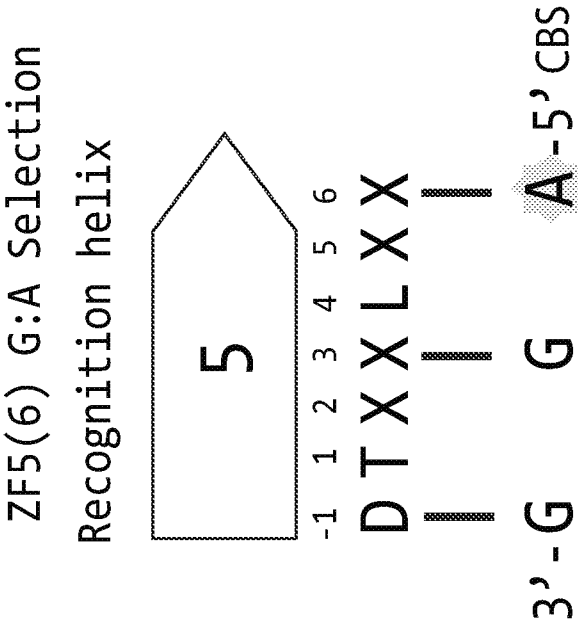


FIG. 12B



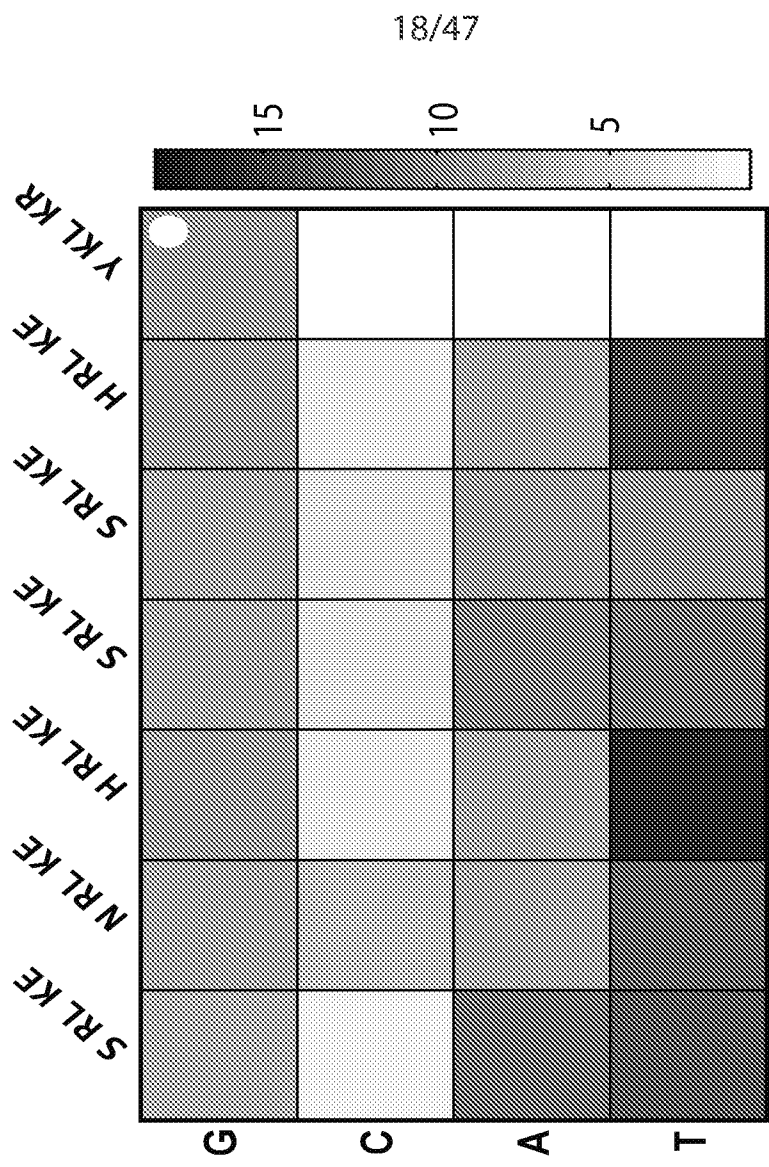
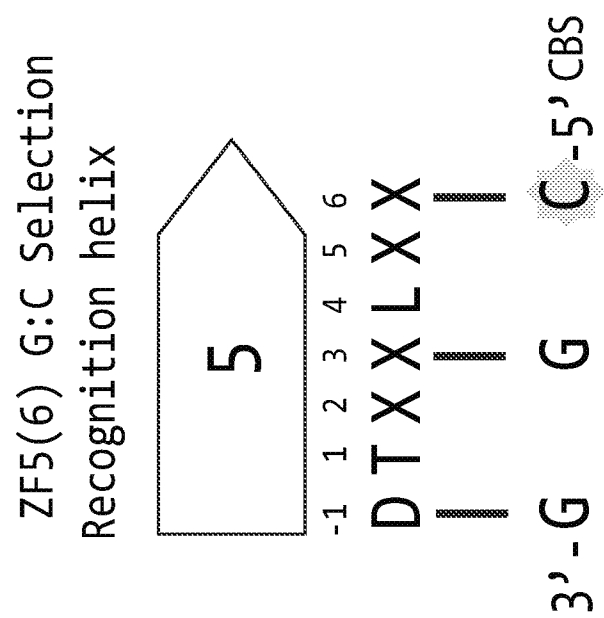


FIG. 12C



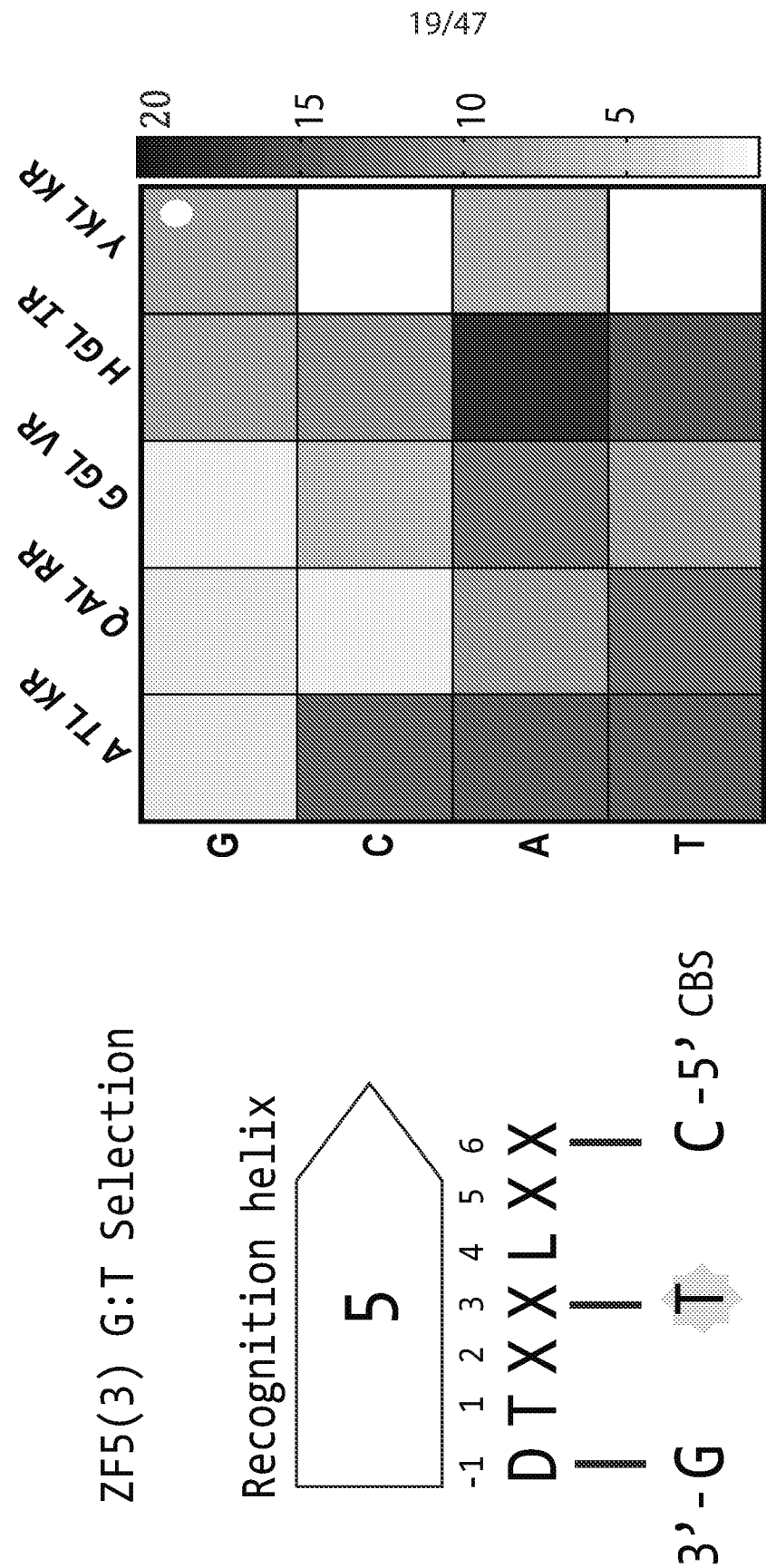


FIG. 13A

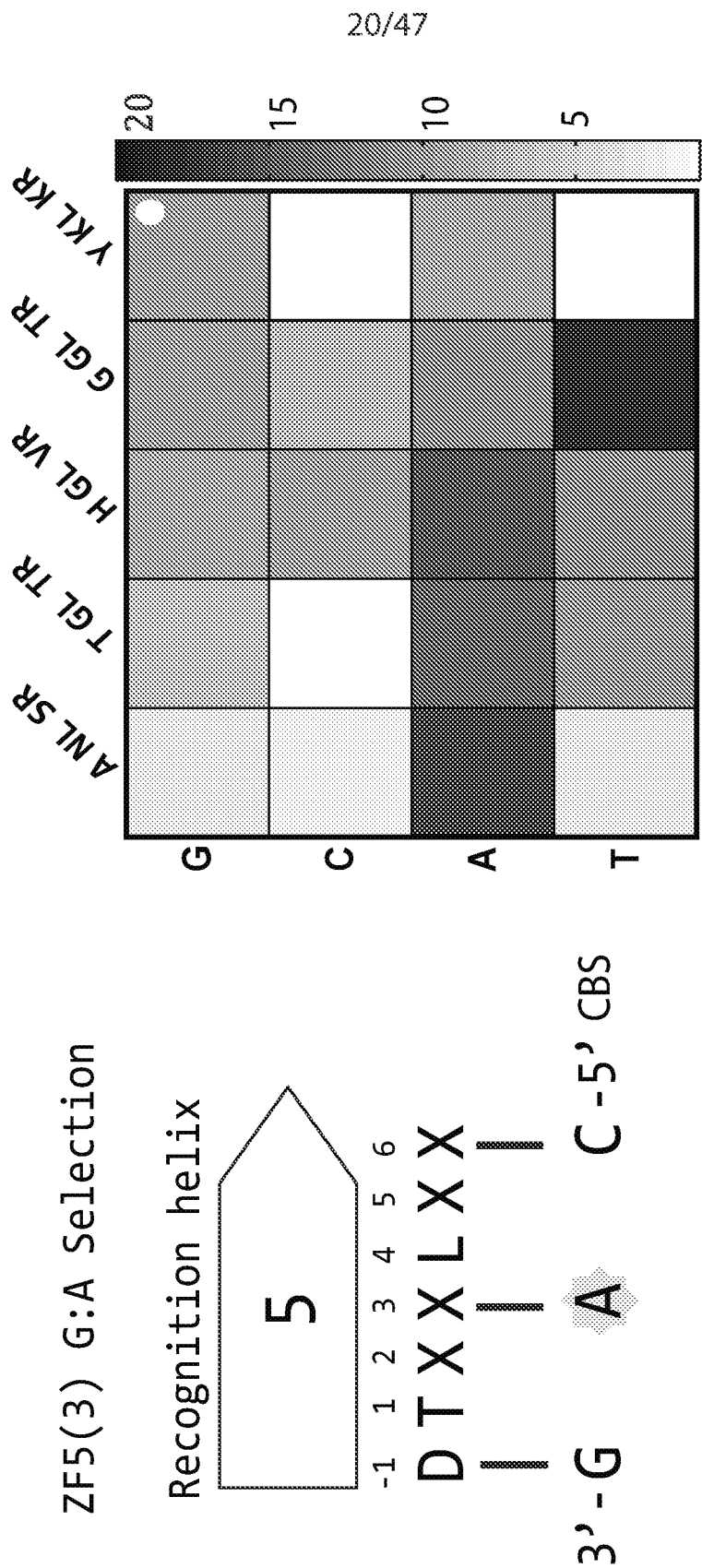


FIG. 13B

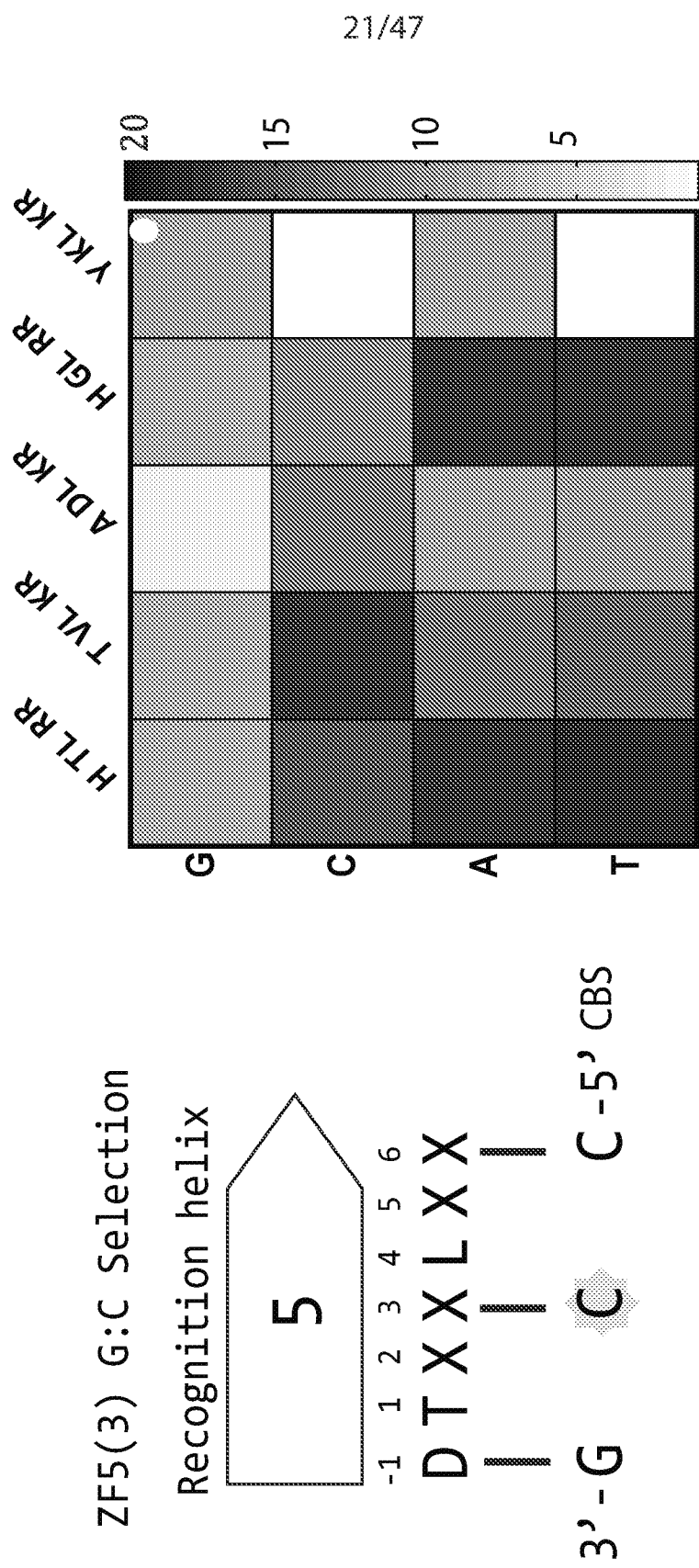


FIG. 13C

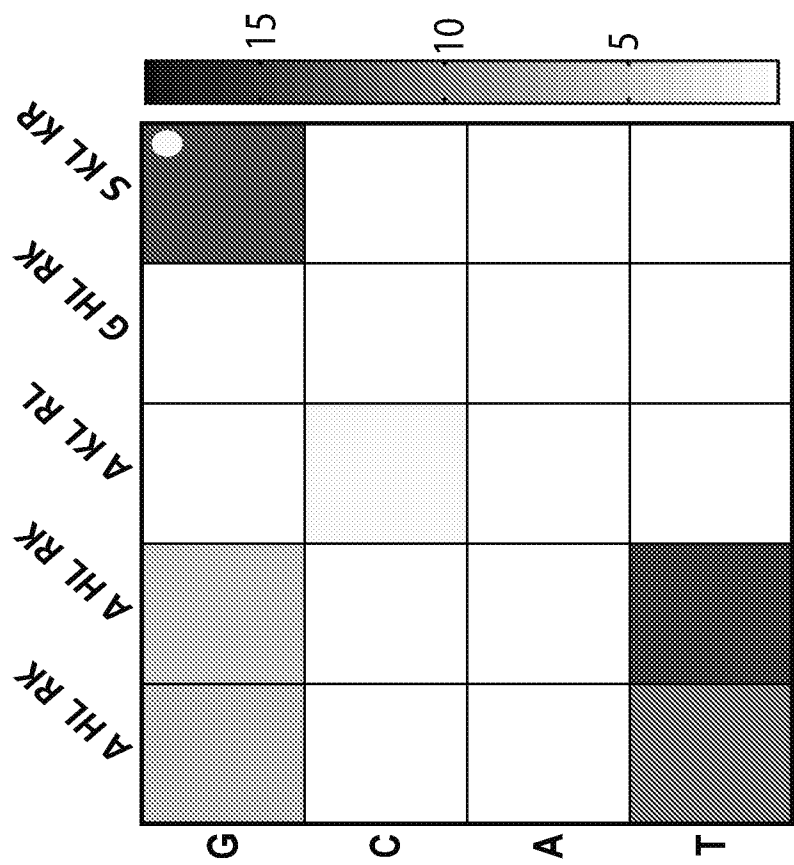
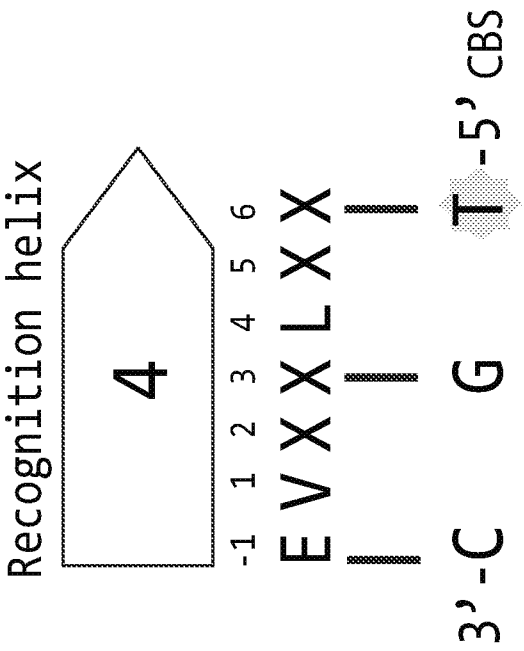


FIG. 13A

ZF4(6) G:T Selection



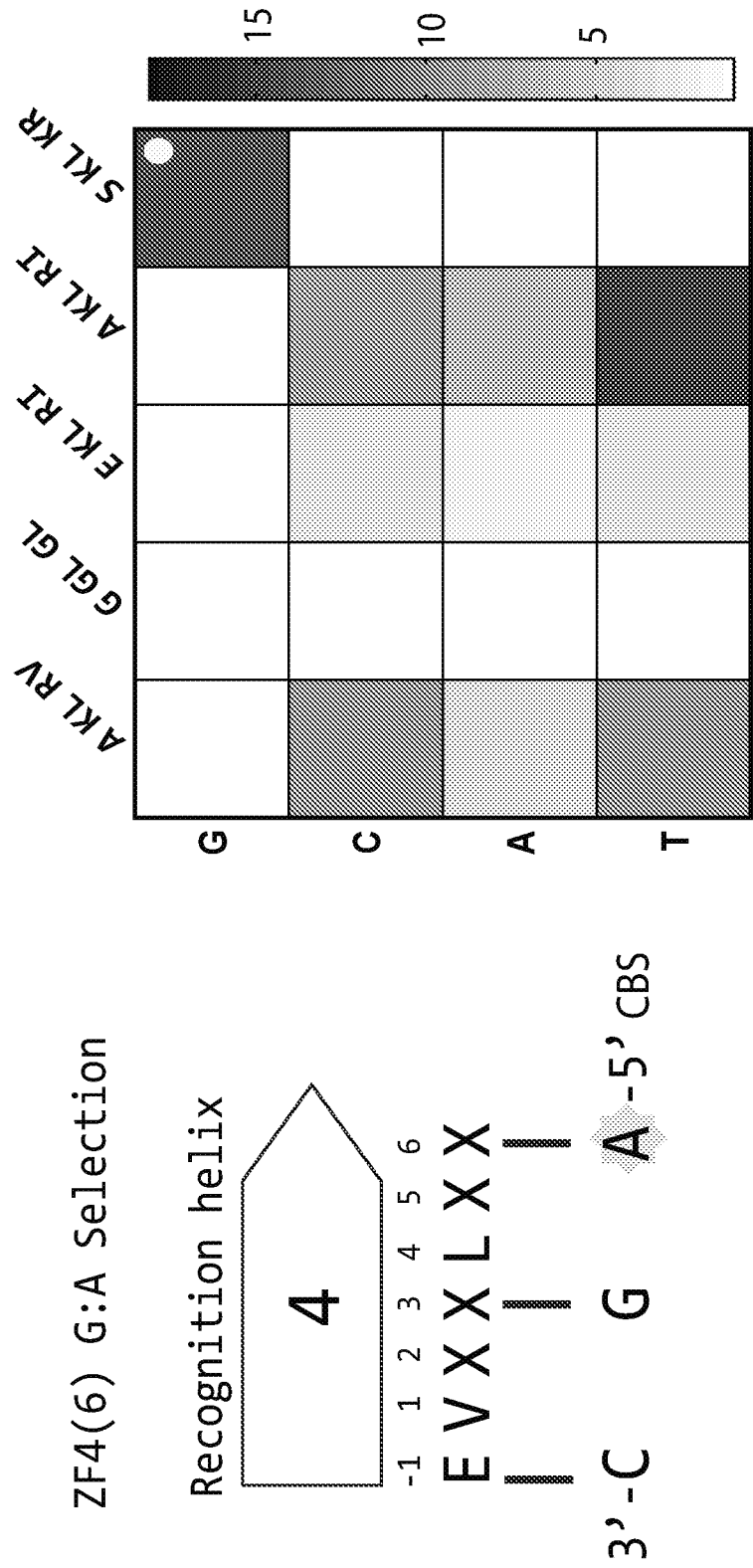


FIG. 14B

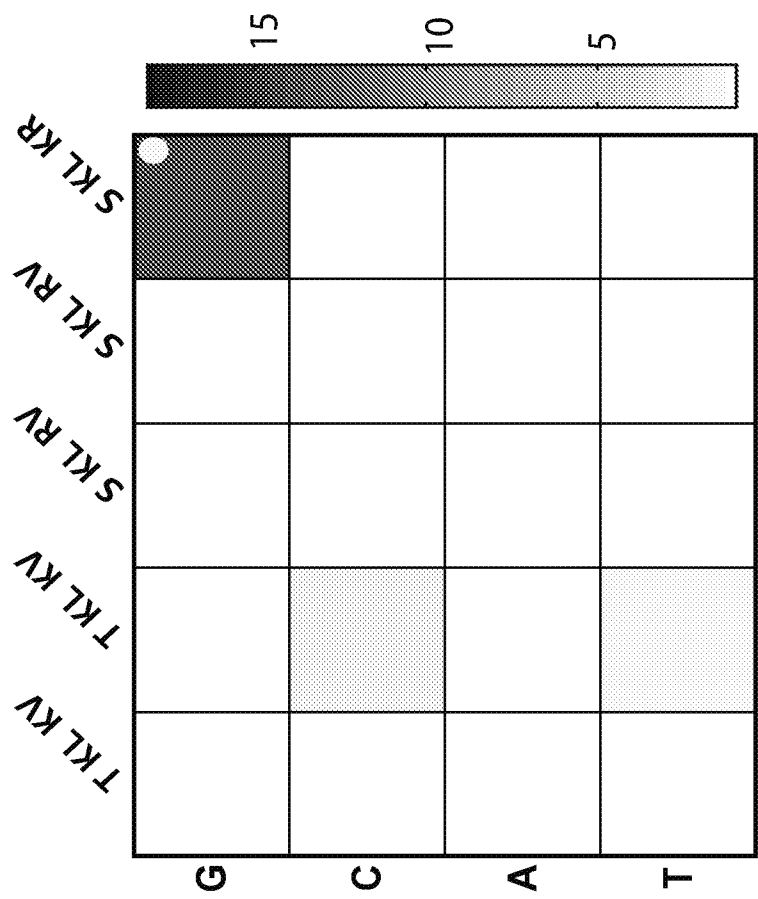
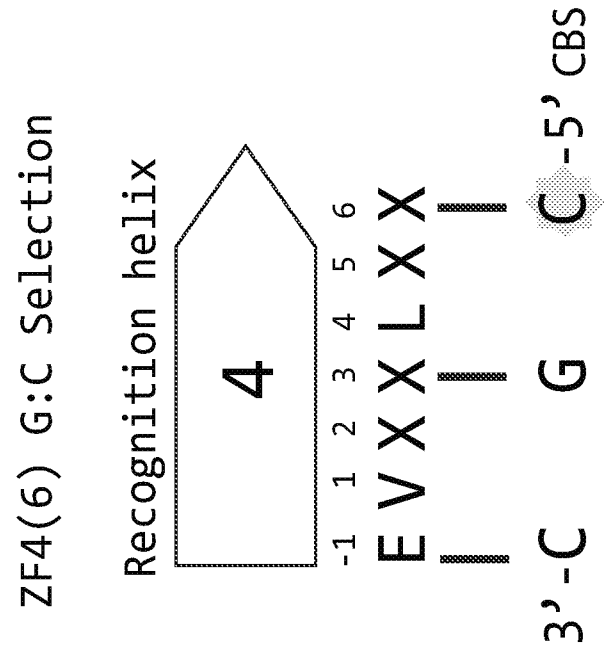


FIG. 14C



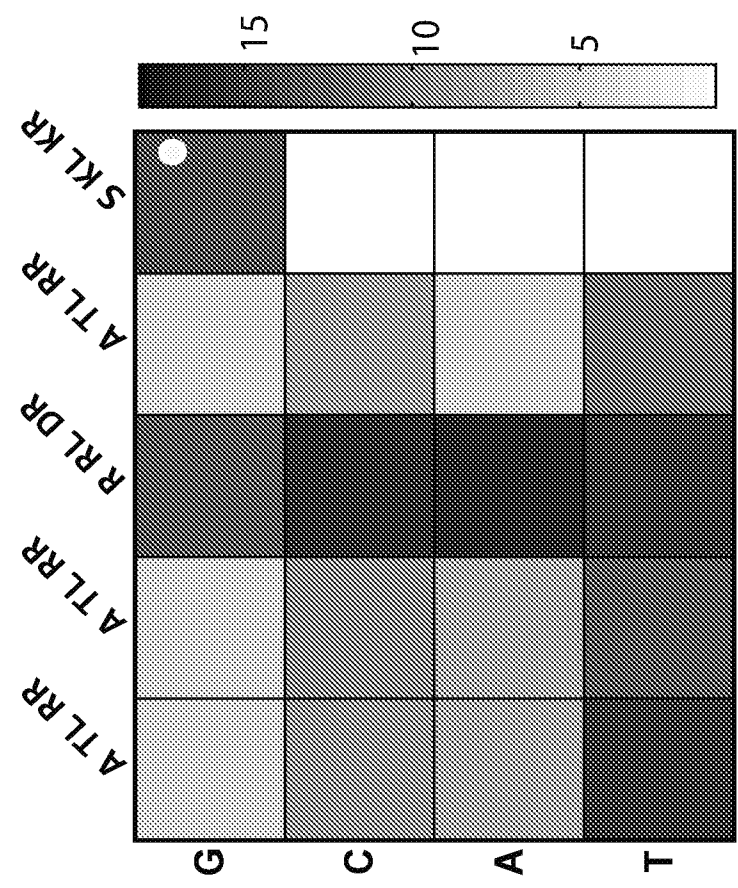
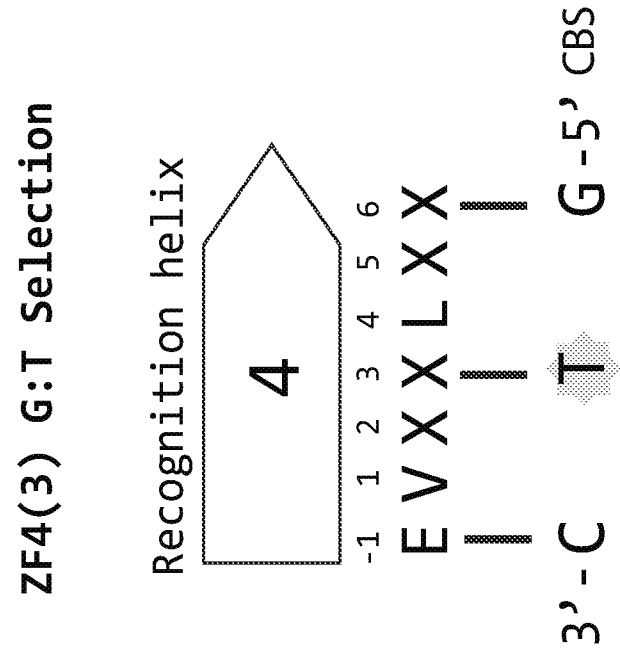


FIG. 15A



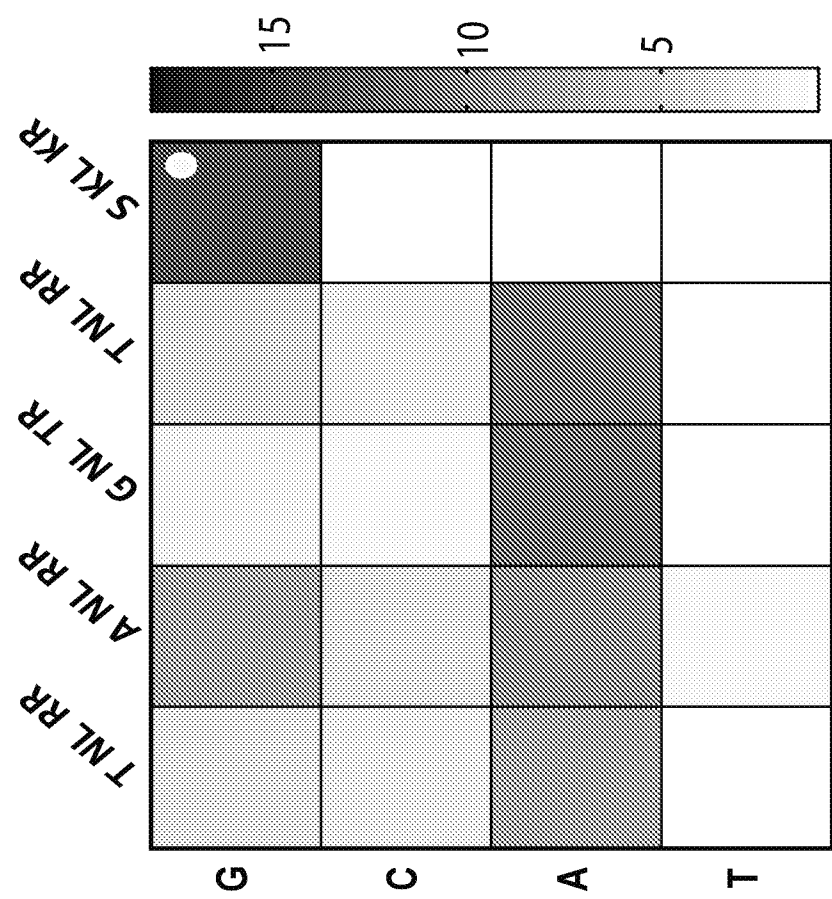
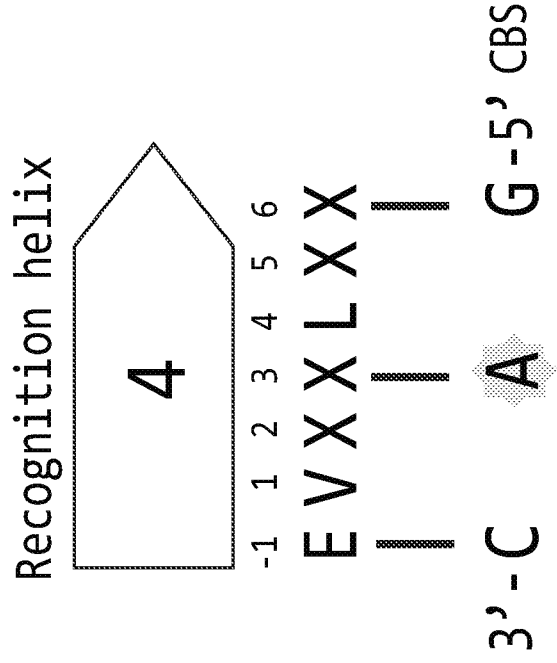


FIG. 15B

ZF4(3) G:A Selection



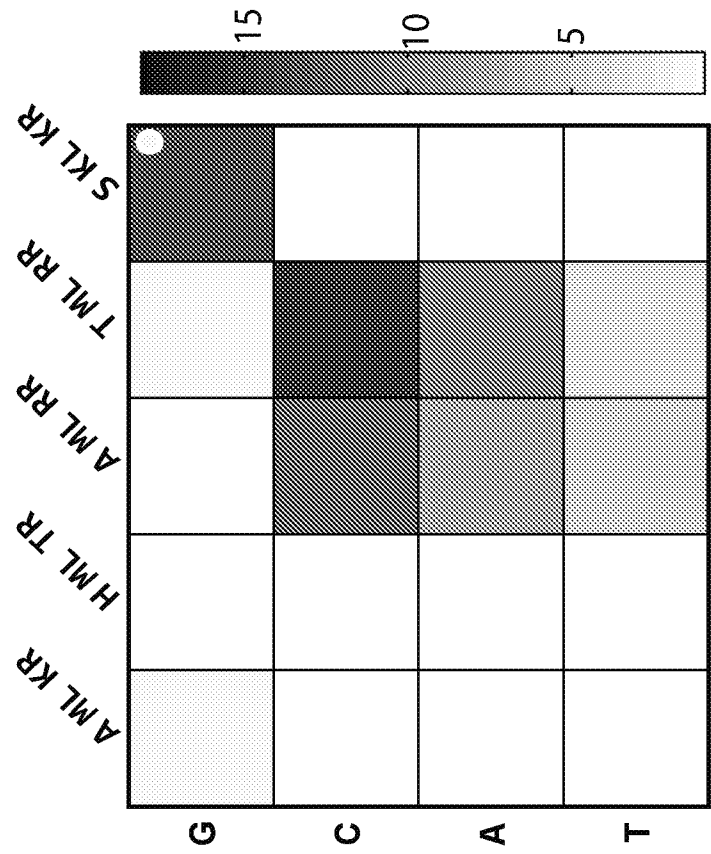
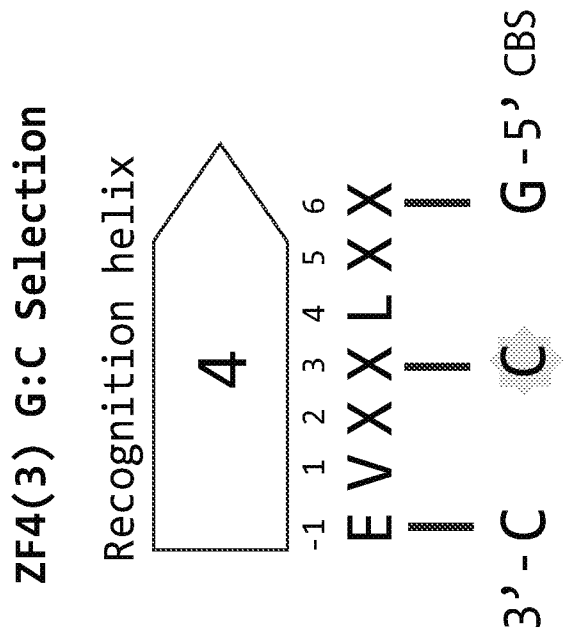


FIG. 15C



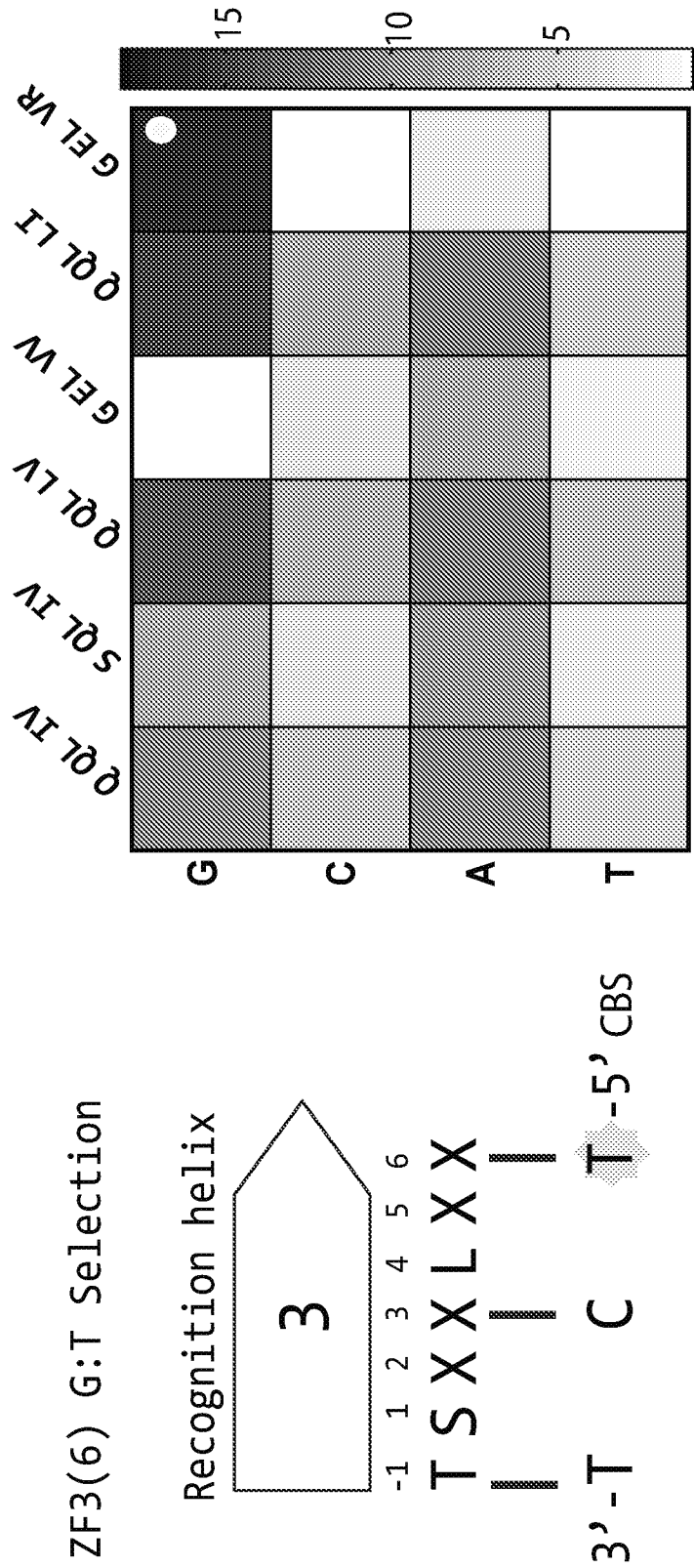


FIG. 16A

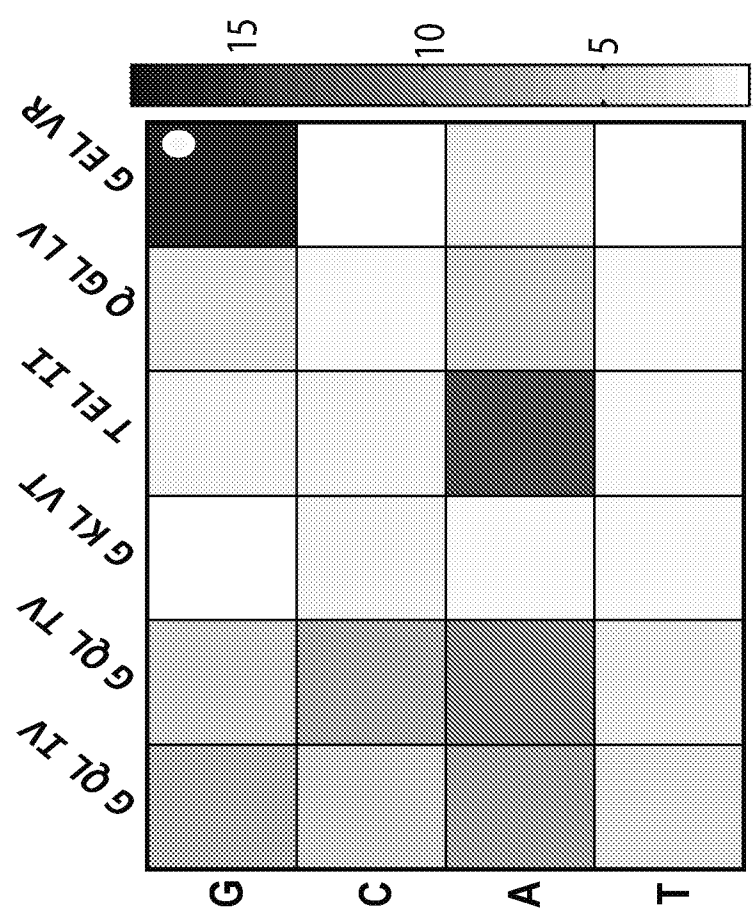
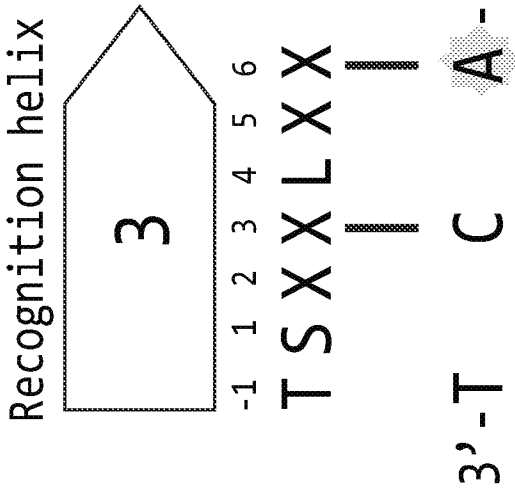


FIG. 16B

ZF3(6) G:A Selection



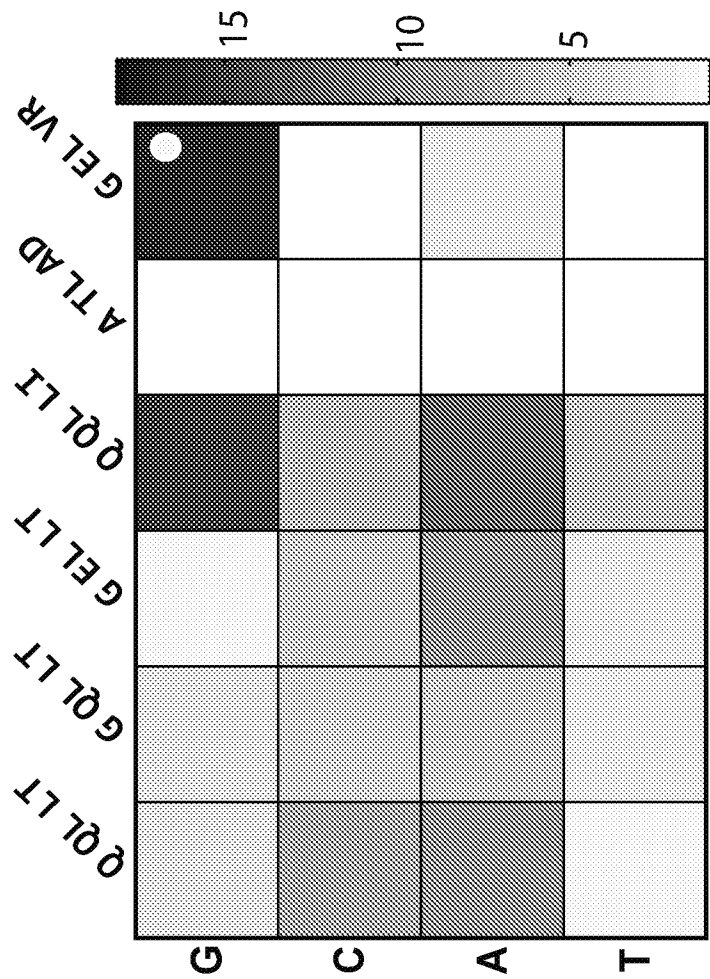
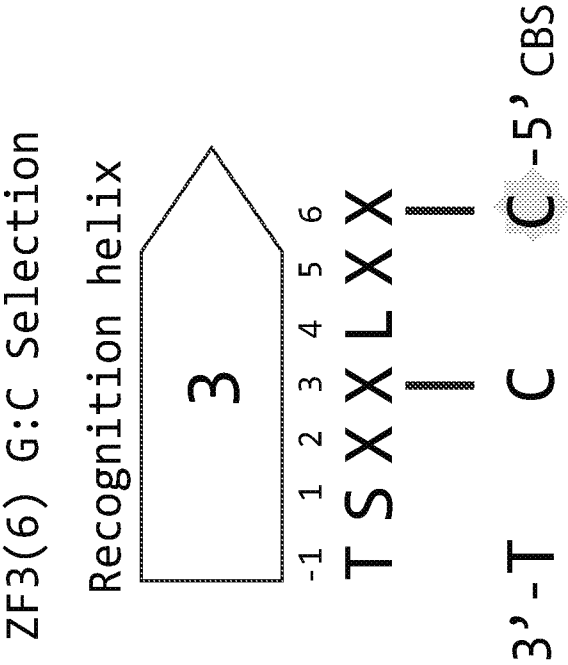


FIG. 16C



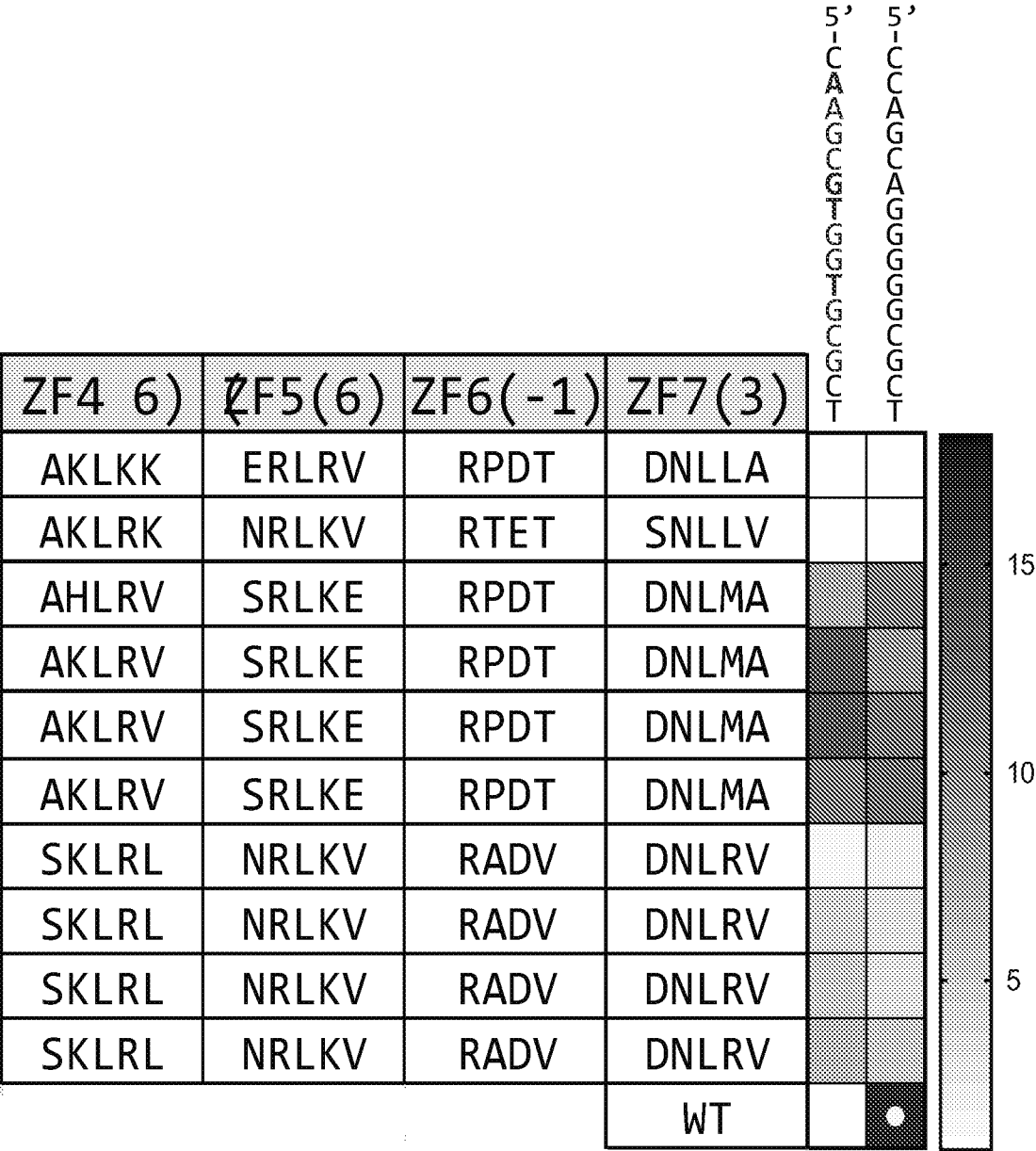
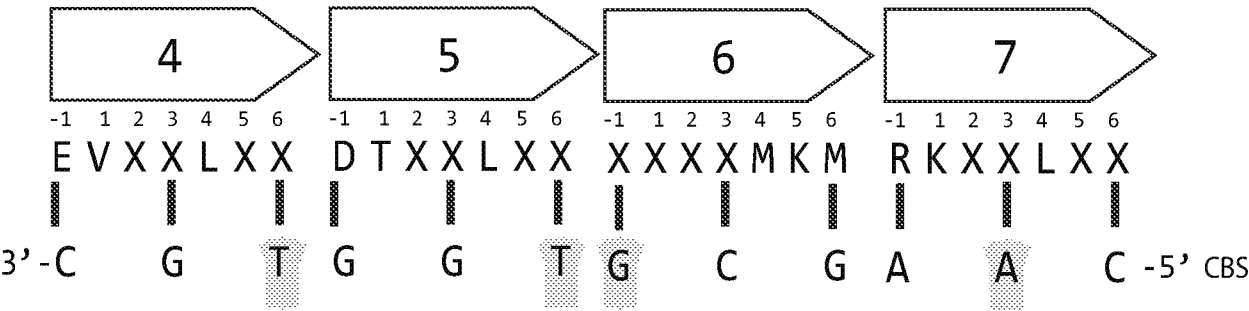
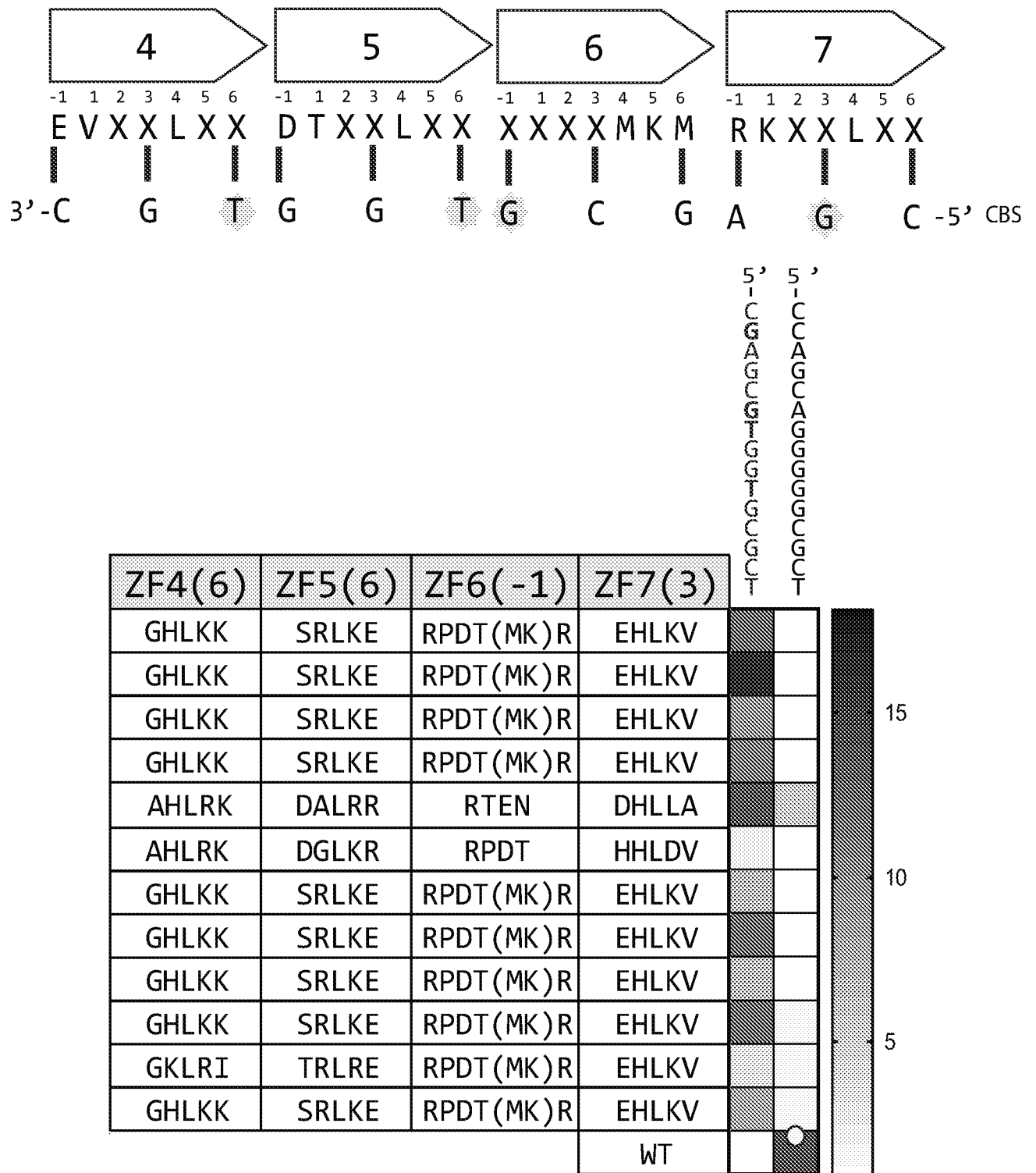


FIG. 17

32/47



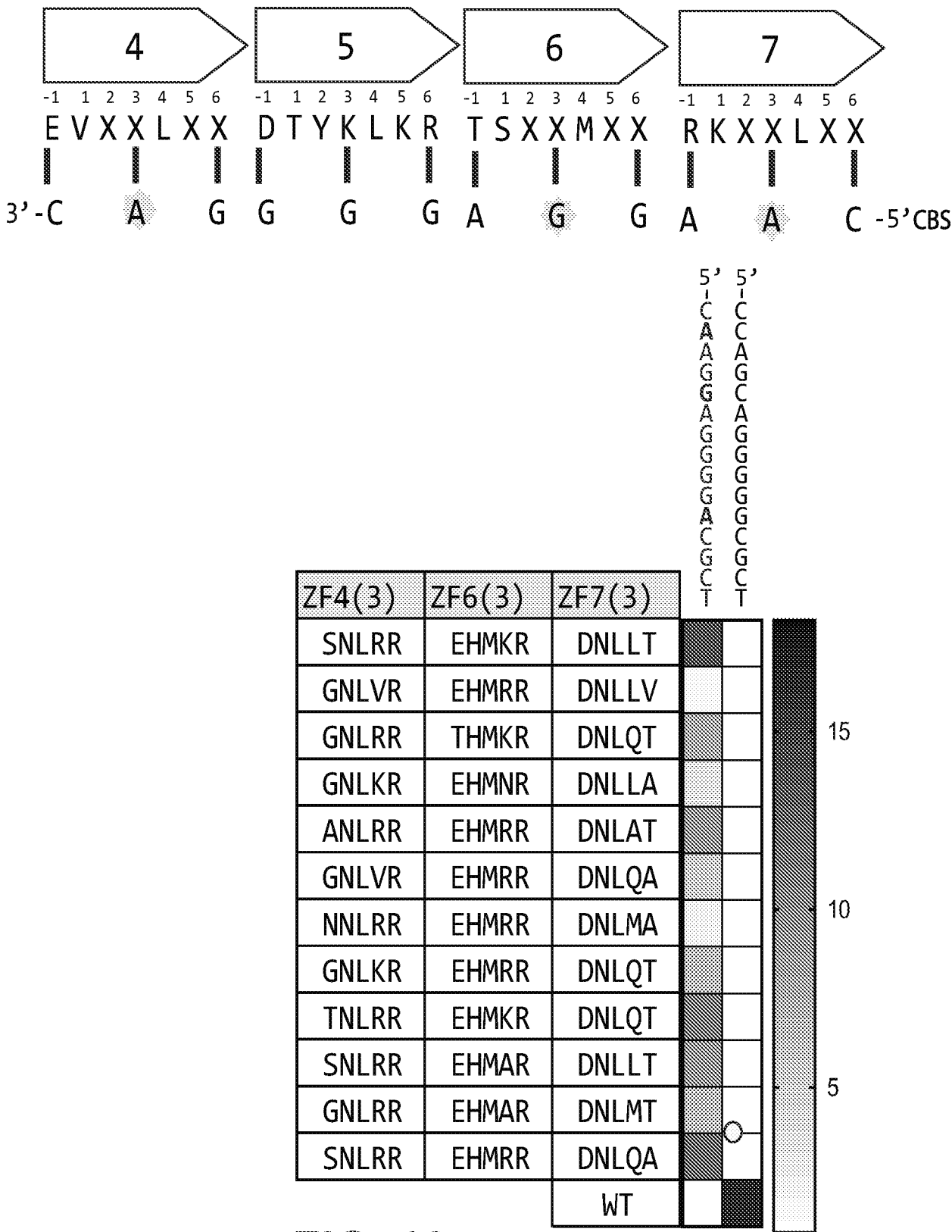


FIG. 19

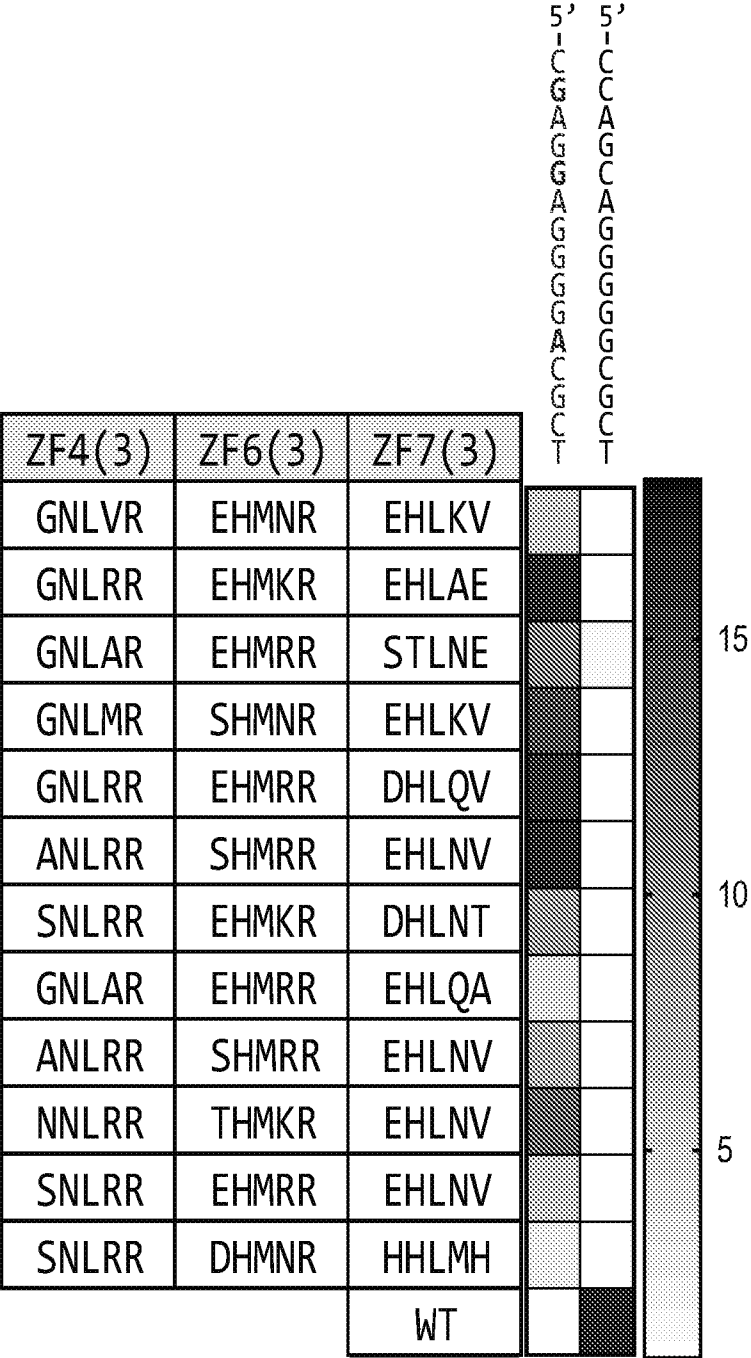
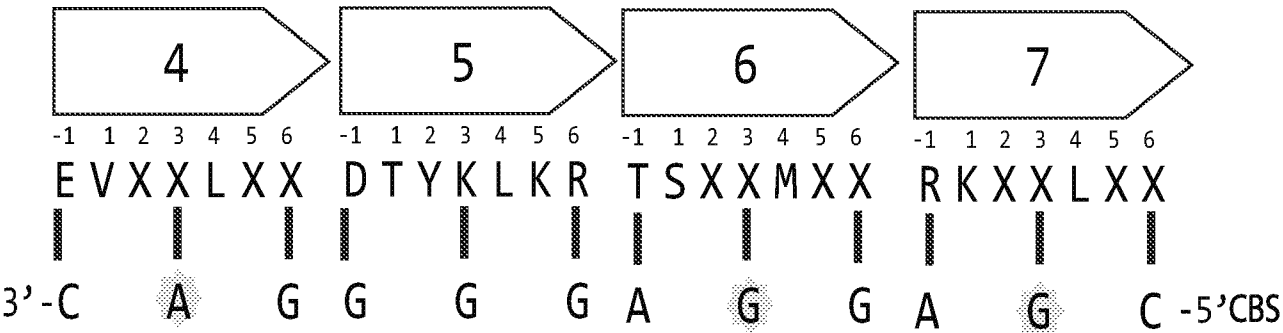


FIG. 20

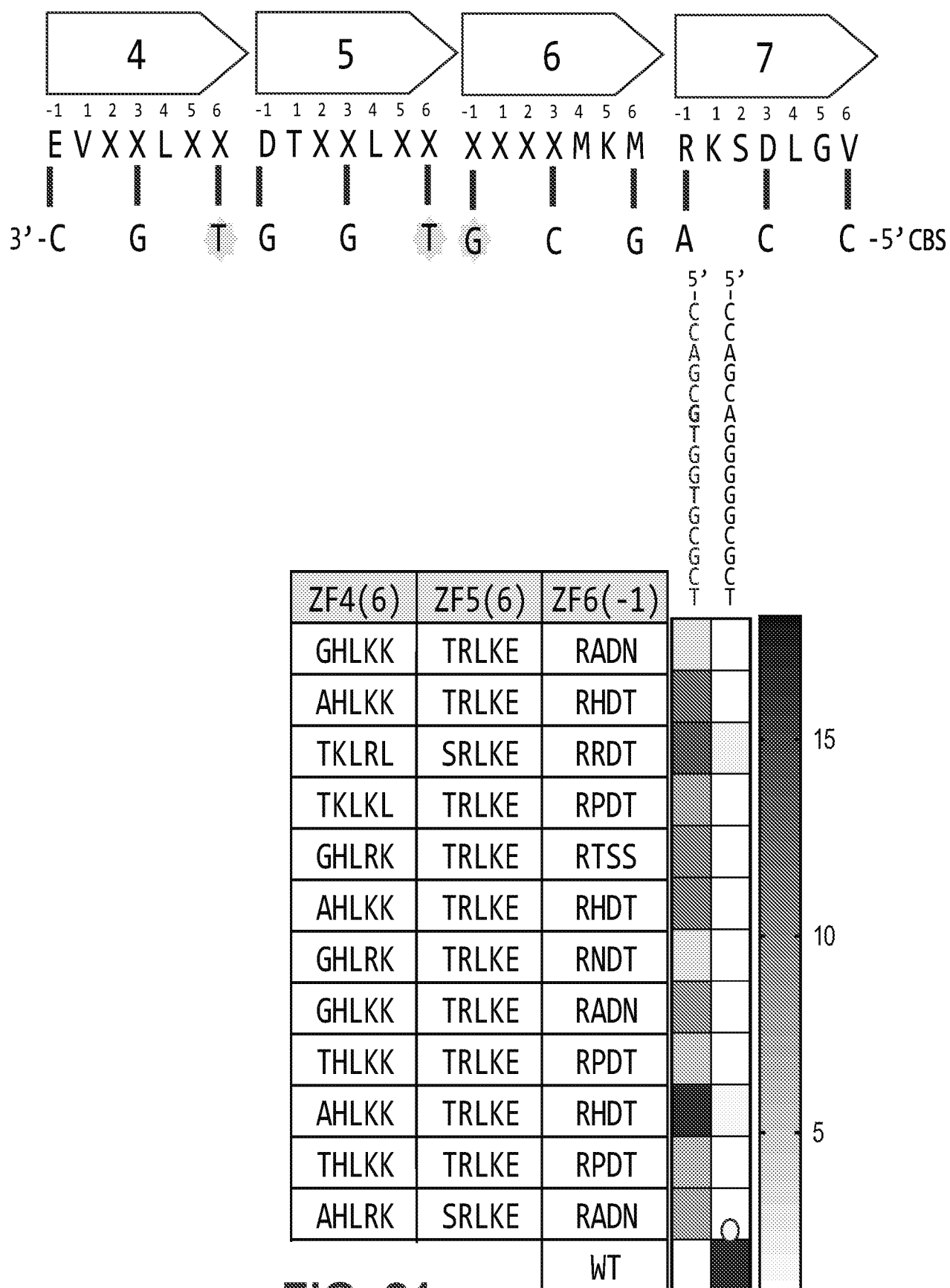


FIG. 21

36/47

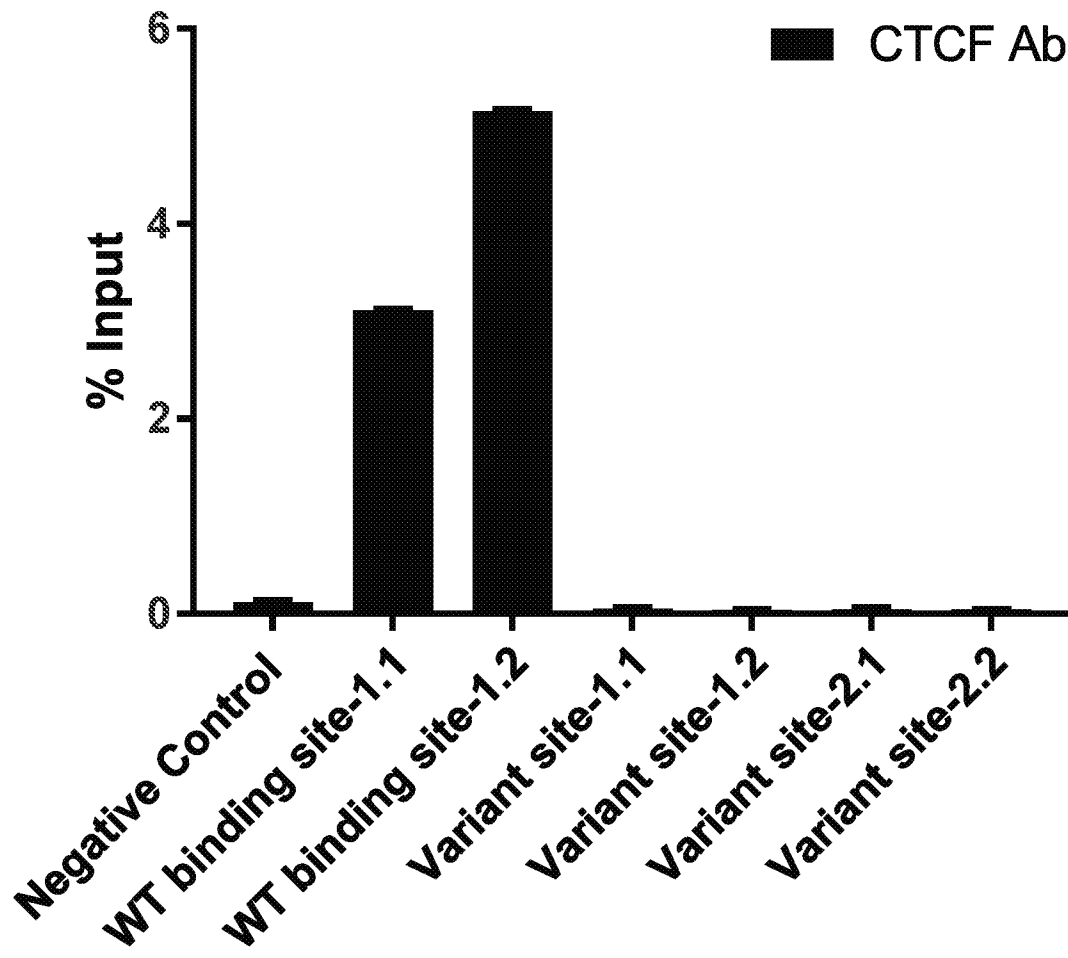


FIG. 22

37/47

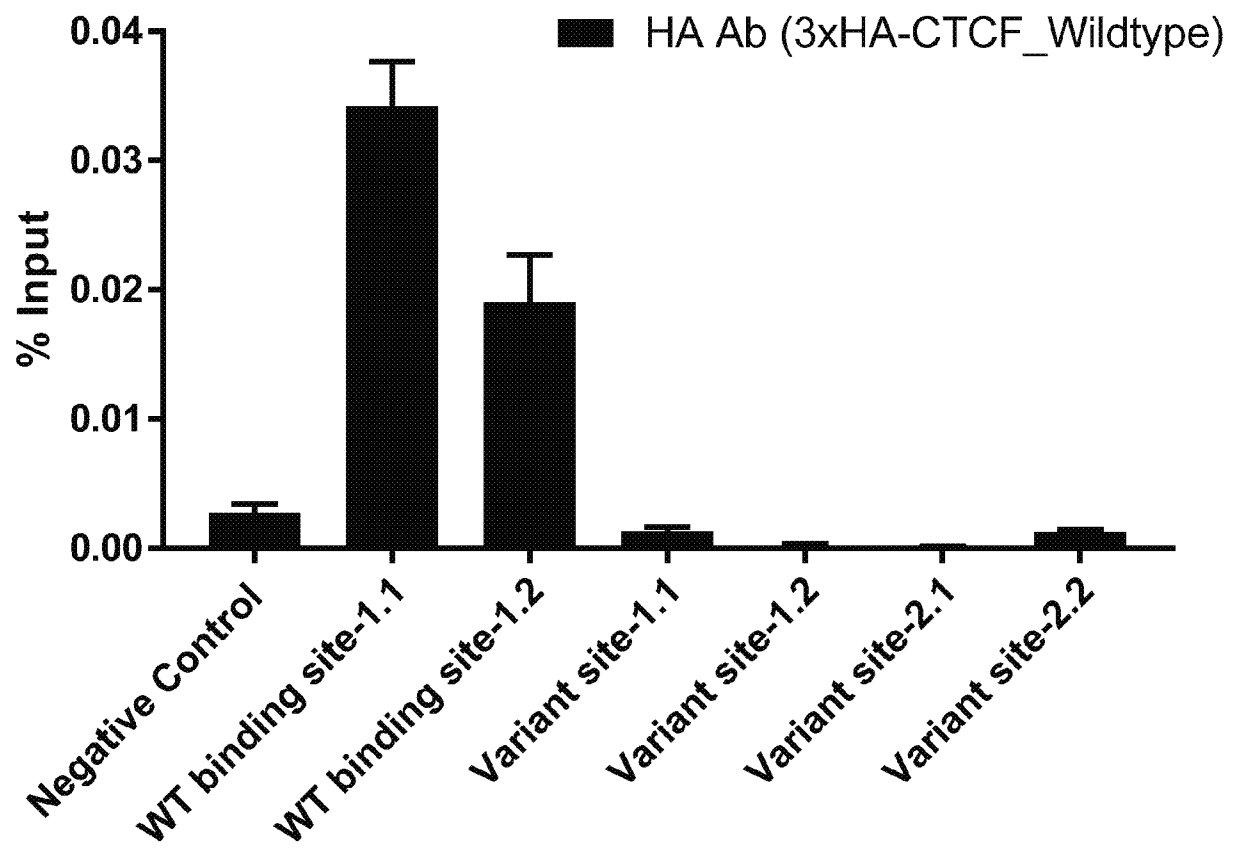


FIG. 23A

38/47

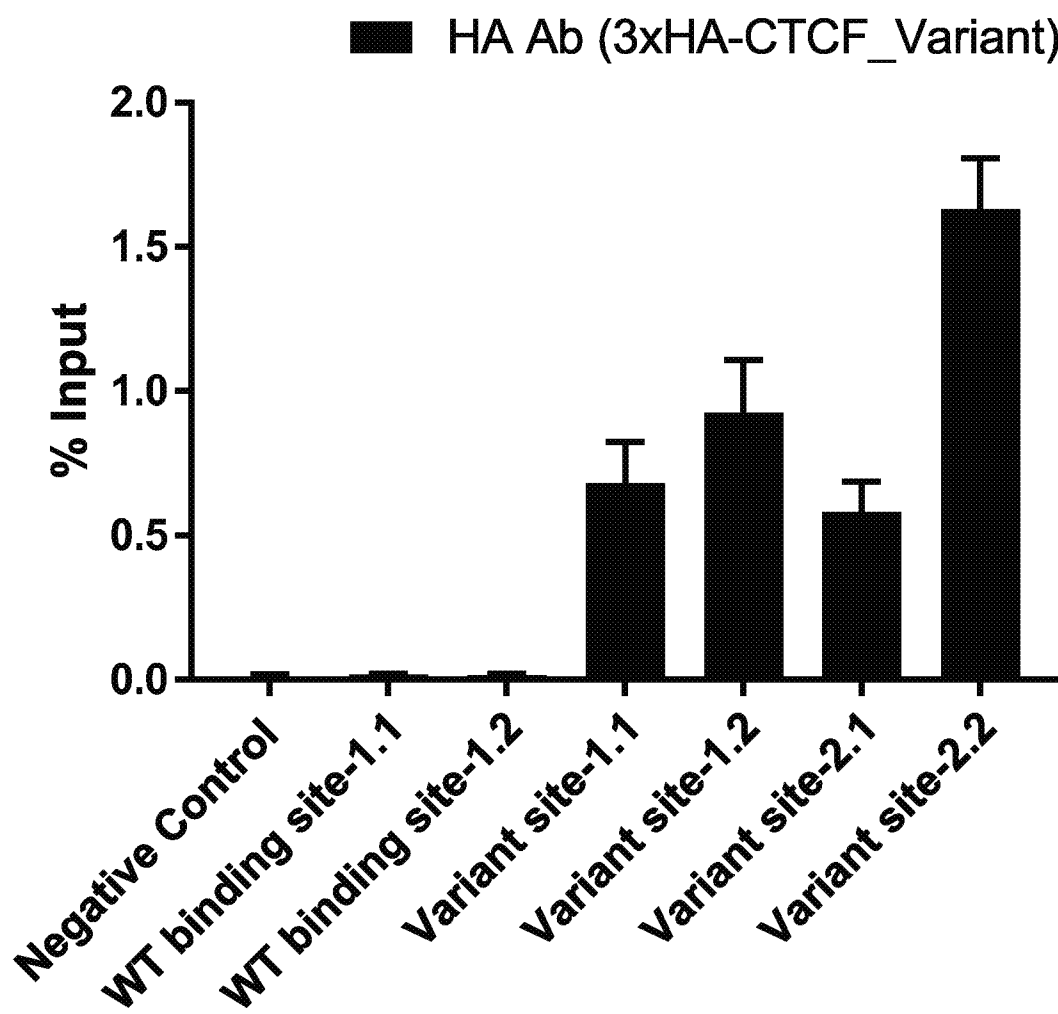


FIG. 23B

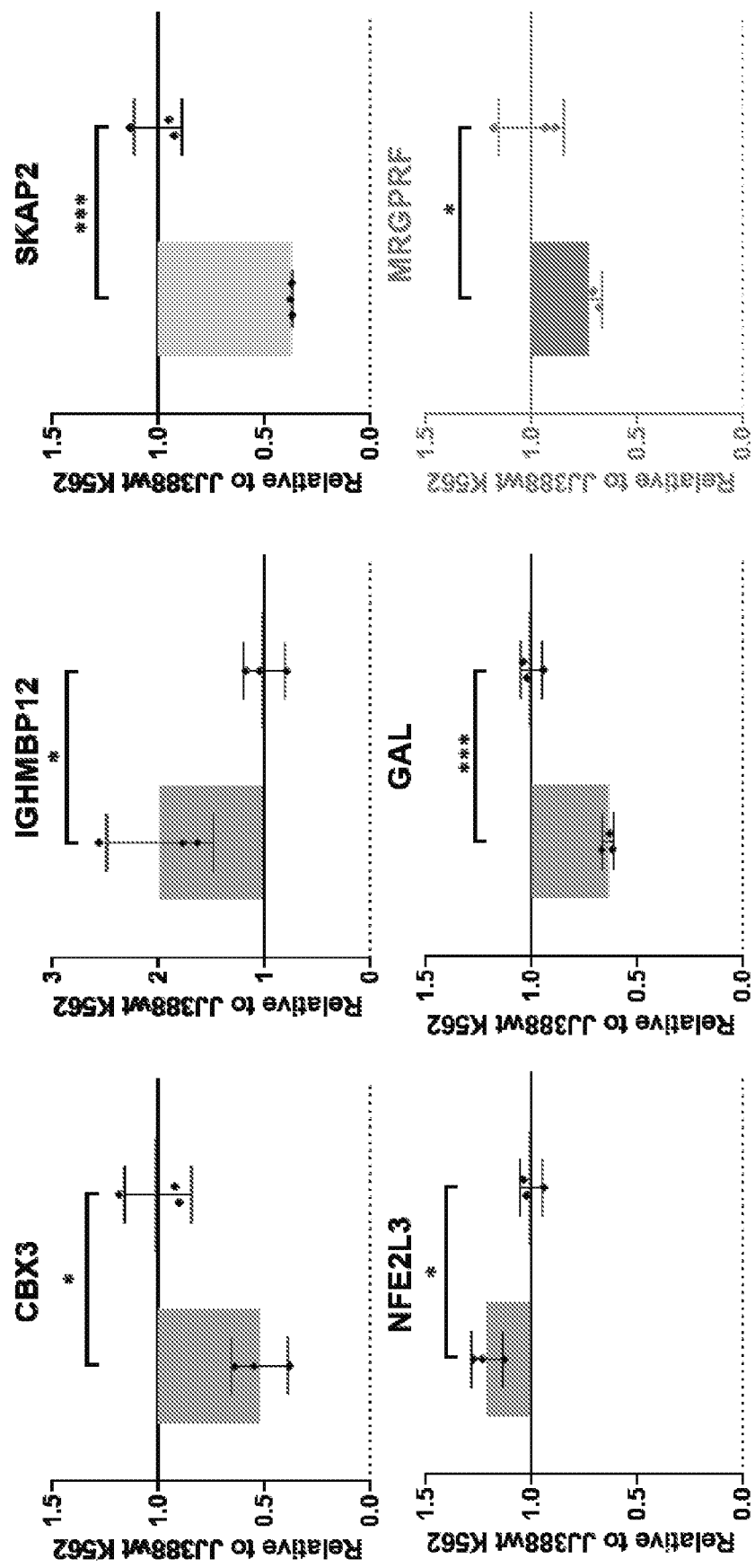


FIG. 24A

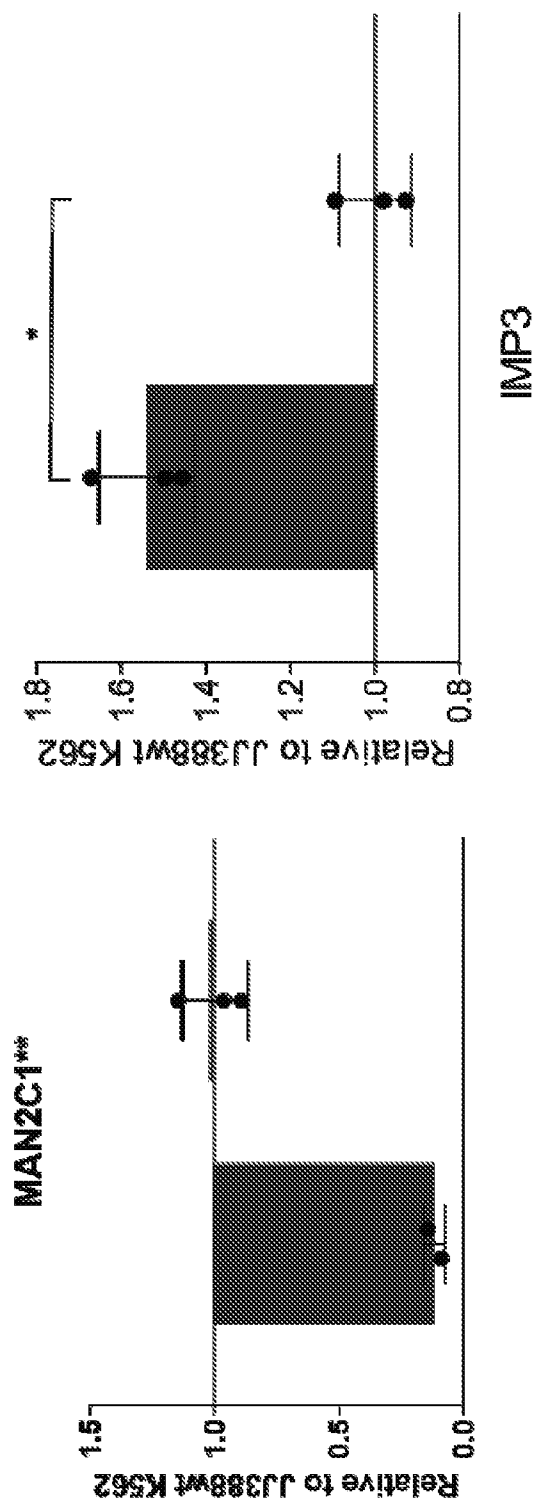


FIG. 24B

Relative to Endo MYC expression of exoMYC.K562

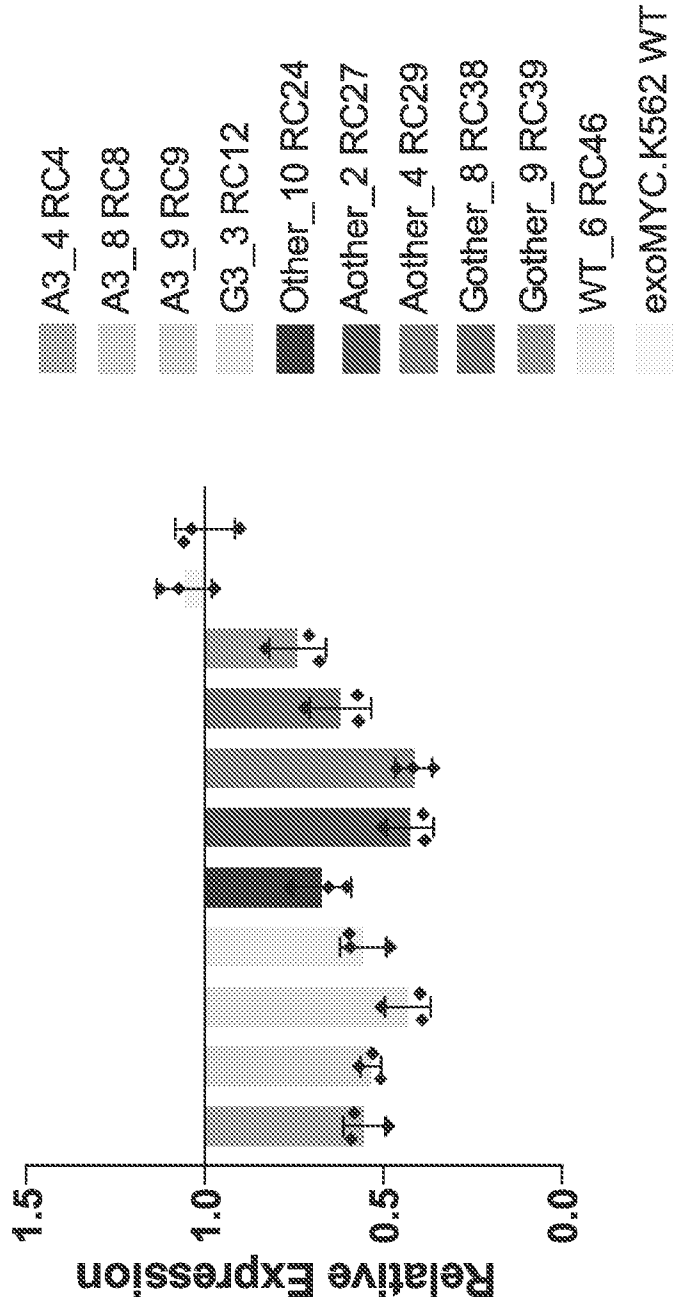
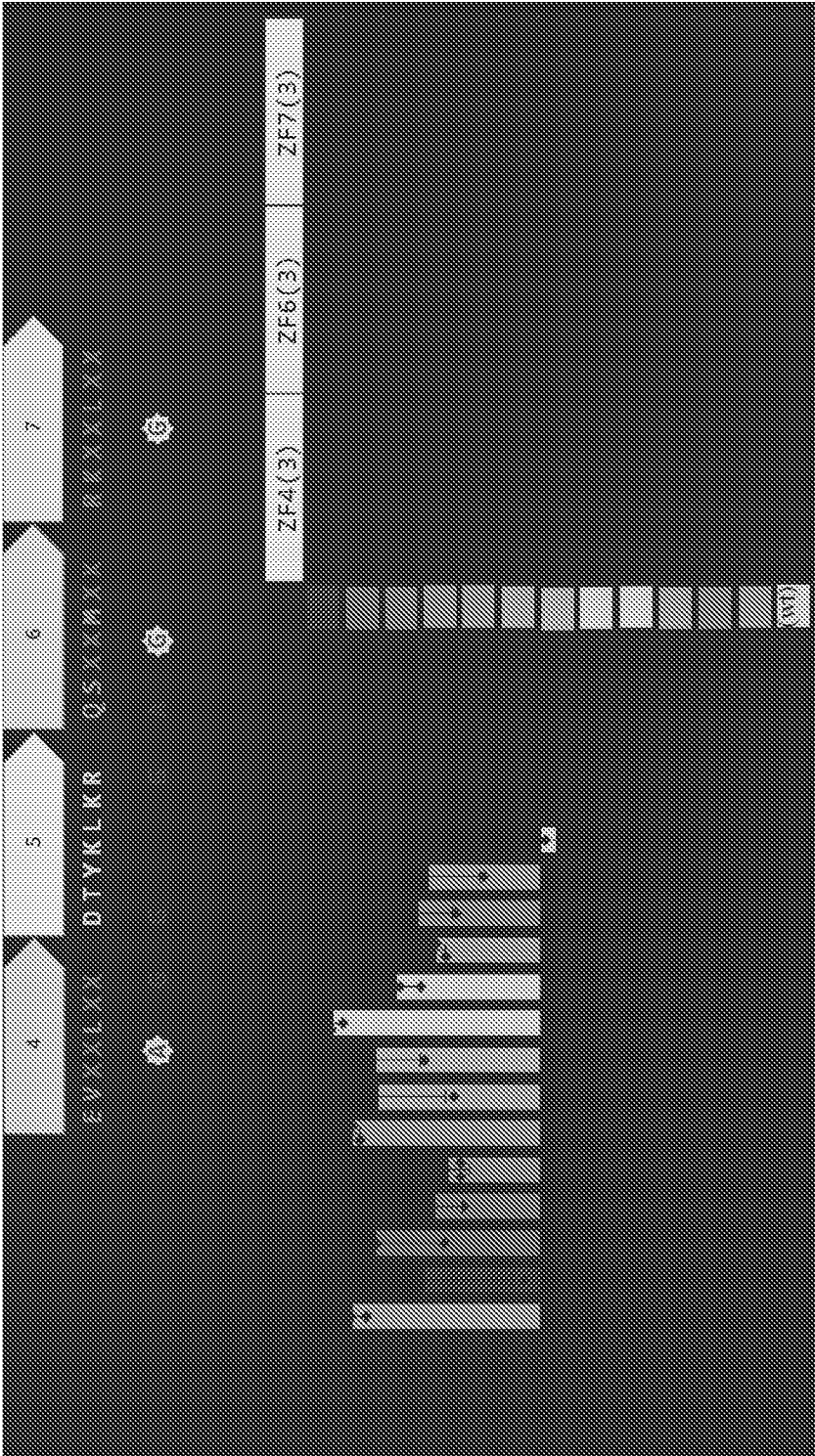
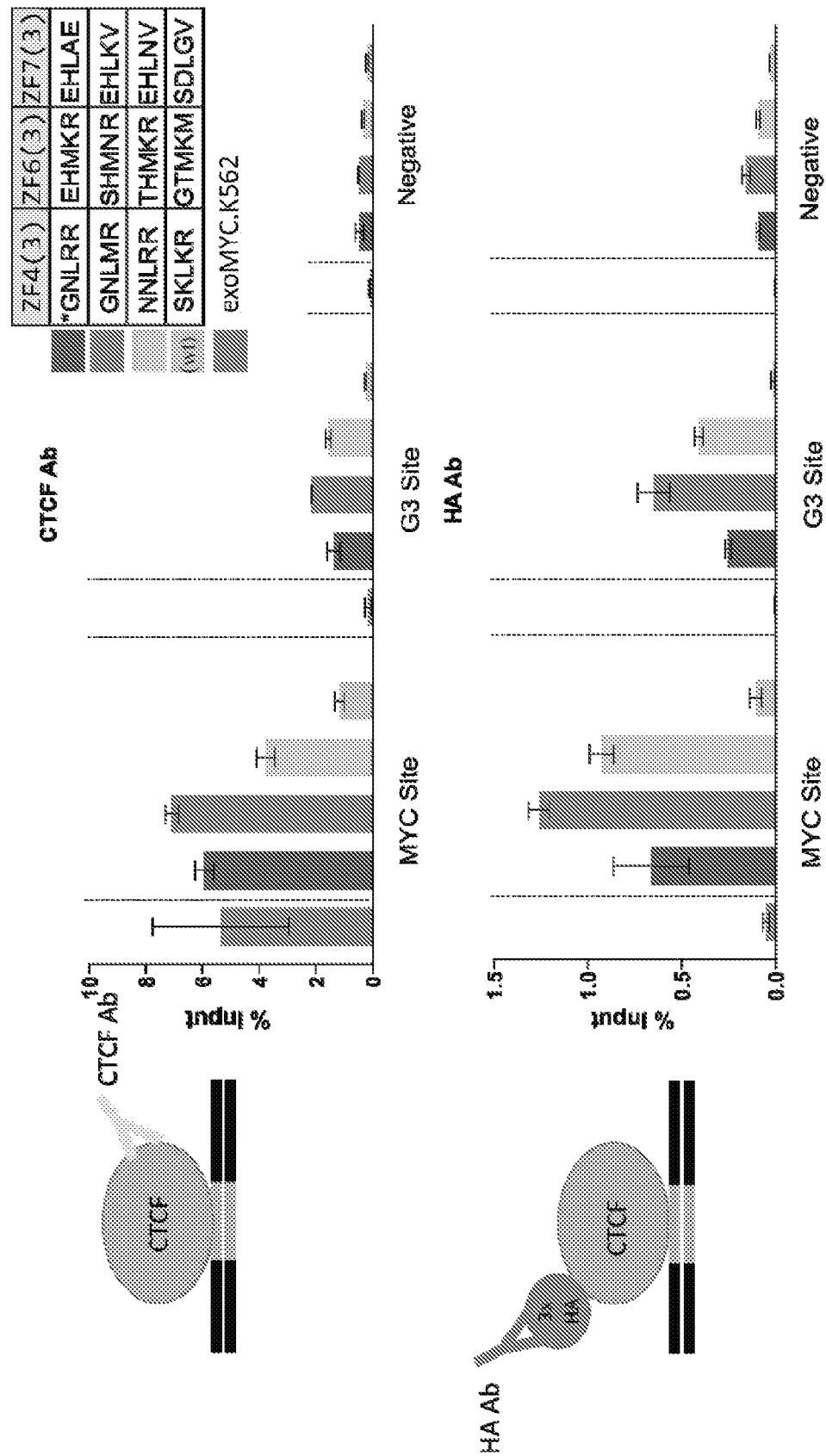


FIG. 25

Binding Site	Sequence	Cell Line
A3	TGCAAGGAGGTGACGCTA	A3_4,8,9
G3	TGCGAGGAGGTGACGCTA	G3_3
Other	TGCCAGCGTGGTGGCTA	Other_10
A Other	TGCAAGCGTGGTGGCTA	Aother_2,4
G Other	TGCGAGCGTGGTGGCTA	Goher_8,9
WT	TGCCAGCAGGTGGCGCTA	WT_6

FIG. 26A



B2G²L

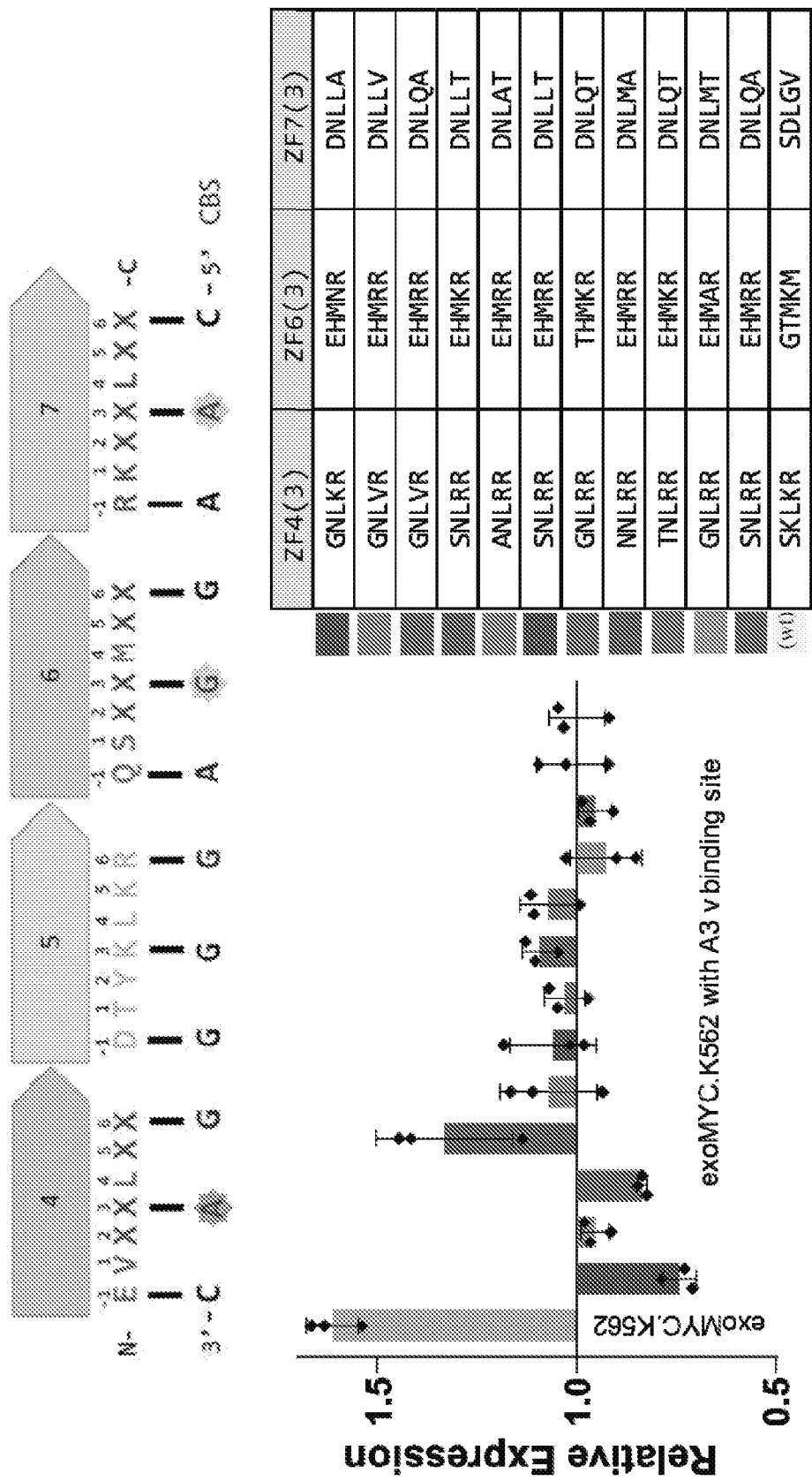


FIG. 27A

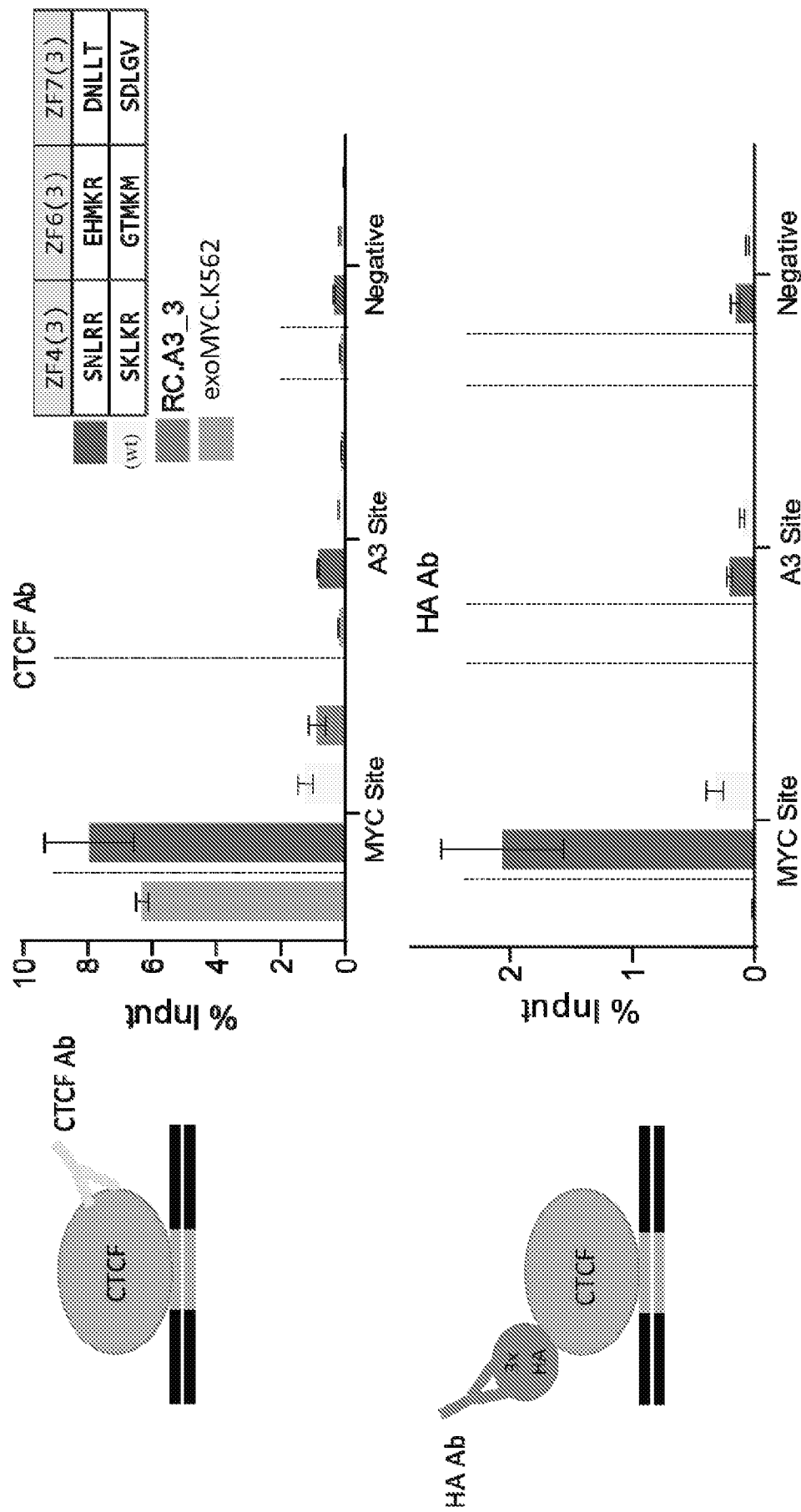
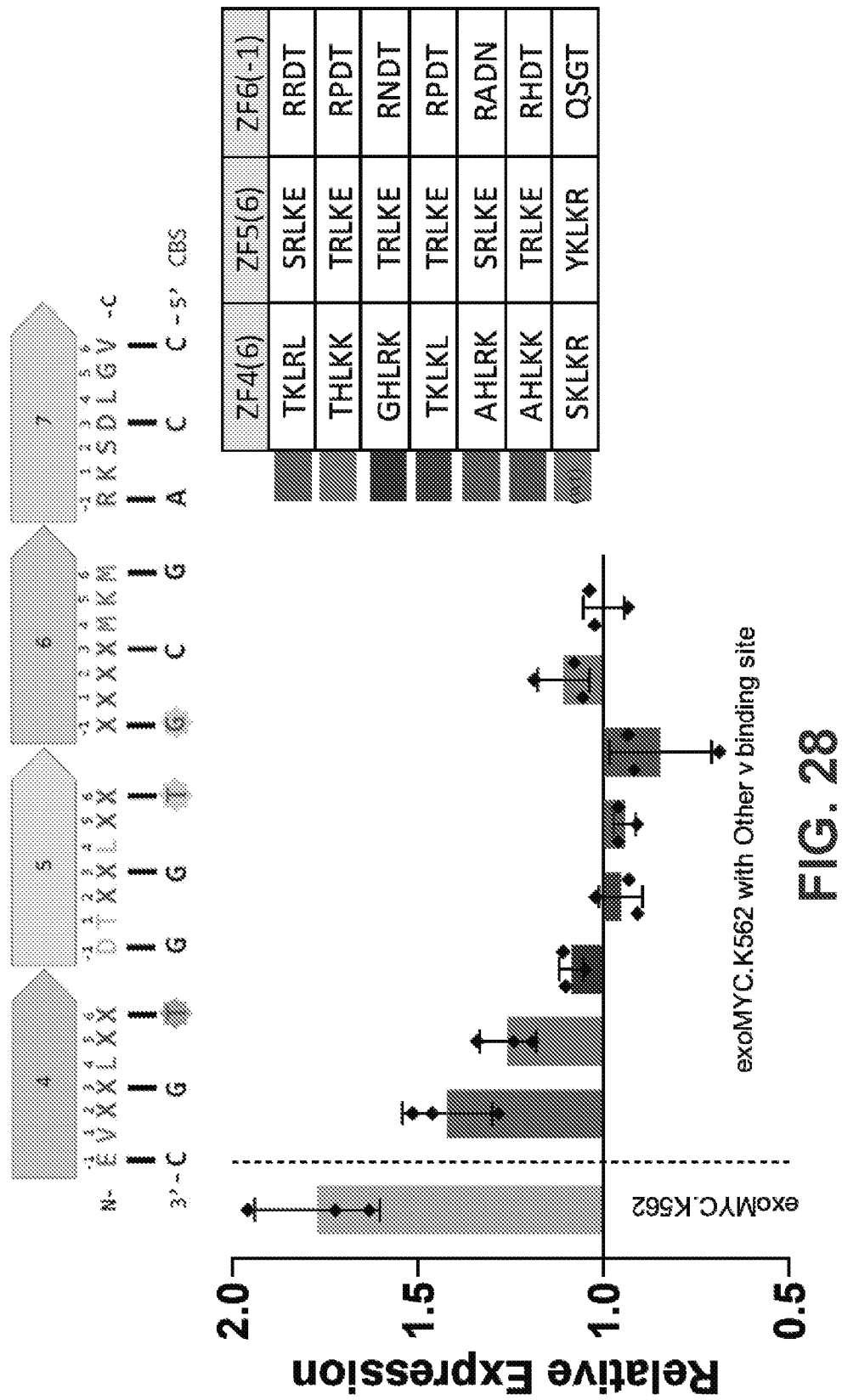
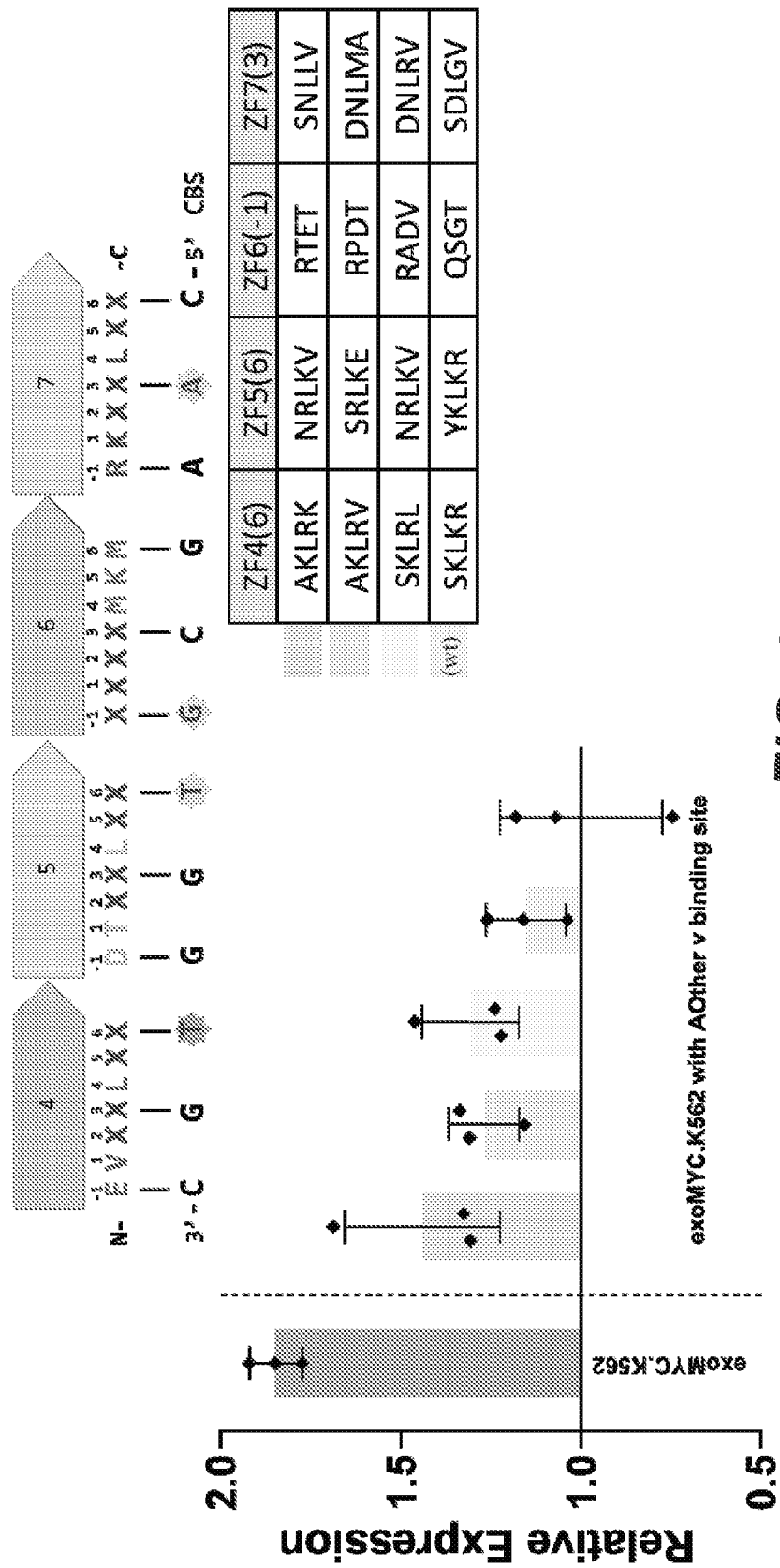


FIG. 27B





29

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/032937

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - A61K 38/16; C07K 19/00; C12N 15/09 (2019.01)
CPC - A61K 38/16; C07K 2319/00; C07K 2319/81 (2019.08)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

USPC - 435/69.1; 514/2; 514/1.1; 530/300 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2007/0213269 A1 (BARBAS, III et al) 13 September 2007 (13.09.2007) entire document	1-3, 17
A	WO 2017/031370 A1 (THE BROAD INSTITUTE, INC. et al) 23 February 2017 (23.02.2017) entire document	1-3, 17
A	US 2012/0115227 A1 (COHEN-HAGUENAUER et al) 10 May 2012 (10.05.2012) entire document	1-3, 17
A	US 2016/0215280 A1 (CSIR) 28 July 2016 (28.07.2016) entire document	1-3, 17
A	✓ LI et al. "Identification of critical base pairs required for CTCF binding in motif M1 and M2," Protein & Cell, 17 March 2017 (17.03.2017), Vol. 8, Iss. 7, Pgs. 544-549. entire document	1-3, 17
A	US 2011/0294873 A1 (MERMOD et al) 01 December 2011 (01.12.2011) entire document	1-3, 17
A	US 2017/0175136 A1 (UNIVERSITY OF WASHINGTON) 22 June 2017 (22.06.2017) entire document	1-3, 17

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search
01 October 2019

Date of mailing of the international search report

17 OCT 2019

Name and mailing address of the ISA/US
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, VA 22313-1450
Facsimile No. 571-273-8300

Authorized officer
Blaine R. Copenheaver
PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2019/032937

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:

a. ☒ forming part of the international application as filed:

☒ in the form of an Annex C/ST.25 text file.

☐ on paper or in the form of an image file.

b. ☐ furnished together with the international application under PCT Rule 13ter.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.

c. ☒ furnished subsequent to the international filing date for the purposes of international search only:

☒ in the form of an Annex C/ST.25 text file (Rule 13ter.1(a)).

☐ on paper or in the form of an image file (Rule 13ter.1(b) and Administrative Instructions, Section 713).

2. ☒ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

SEQ ID NOs: 8-13, 15-18, and 30-39 were searched.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2019/032937

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☒ Claims Nos.: 18-26
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

See extra sheet(s).

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-3 and 17 to the extent that they read on an engineered CCCTC-binding factor (CTCF) variant, wherein the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS), wherein the mutant CBS has a G residue at position 2 of the consensus CBS motif, the engineered CTCF comprises the amino acid sequence DHLQT at ZF7 positions +2 to +6.

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/032937

Continued from Box No. III Observations where unity of invention is lacking

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2019/032937

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees need to be paid.

Group I+: claims 1-17 are drawn to engineered CCCTC-binding factor variants.

The first invention of Group I+ is restricted to an engineered CCCTC-binding factor (CTCF) variant, wherein the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS), wherein the mutant CBS has a G residue at position 2 of the consensus CBS motif, the engineered CTCF comprises the amino acid sequence DHLQT at ZF7 positions +2 to +6. It is believed that claims 1-3 and 17 read on this first named invention and thus these claims will be searched without fee to the extent that they read on the above embodiment.

Applicant is invited to elect additional engineered CCCTC-binding factor variants, each with specified consensus CBS motif and corresponding mutant amino acid sequence, to be searched in a specific combination by paying an additional fee for each set of election. An exemplary election would be an engineered CCCTC-binding factor (CTCF) variant, wherein the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS), wherein the mutant CBS has a T residue at position 3 of the consensus CBS motif, the engineered CTCF comprises the amino acid sequence RKHD at ZF7 positions -1 to +3. Additional engineered CCCTC-binding factor variants will be searched upon the payment of additional fees. Applicants must specify the claims that read on any additional elected inventions. Applicants must further indicate, if applicable, the claims which read on the first named invention if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched/examined.

The inventions listed in Groups I+ do not relate to a single general inventive concept under PCT Rule 13.1, because under PCT Rule 13.2 they lack the same or corresponding special technical features for the following reasons:

The Groups I+ formulas do not share a significant structural element responsible for improved CTCF binding sequence affinity, requiring the selection of alternatives for the amino acid sequence of the mutant CCCTC-binding factors and the corresponding mutant consensus CBS motifs, where "the engineered CTCF comprising the amino acid sequence DHLQT, ELINV AHLQV, F-LRE, DHLQIV, ELKT, EHIVV, DHL-RT, or DHTLAT at ZF7 positions +2 to +6" and "the engineered CTCF at ZF7 positions -1 to +3 comprising: the amino acid sequence RKHD or RRS, wherein the mutant CBS has a T residue at position 3 of the consensus CBS motif; the amino acid sequence RKAD, IPRI, RKHD, or RKDD, wherein the mutant CBS has a G residue at position 3 of the consensus CBS motif; or the amino acid sequence GIVN, ELLN, QALL or PHRM, wherein the mutant CBS has a C residue at position 3 of the consensus CBS motif" and "the engineered CTCF at ZF6 positions +2 to +6 comprising: the amino acid sequence NAMKR, GNMAR, EGMTR, SNMVR, or NAMRQ wherein the mutant CBS has a T residue at position 5 of the consensus CBS motif; or the amino acid sequence EHIMGR, DHMNR, THMKR, EH-MRR, or THINR, wherein the mutant CBS has a G residue at position 5 of the consensus CBS motif" and "the engineered CTCF at ZF6 positions -1 to +3 comprising: the amino acid sequence MNES or HRES, wherein the mutant CBS has a T residue at position 6 of the consensus CBS motif; or the amino acid sequence RPD, RTDI, or RHD, wherein the mutant CBS has a G residue at position 6 of the consensus CBS motif" and "the engineered CTCF at ZF5 positions -2 to +6 comprising: the amino acid sequence HGLKV, HRLKE, HALKV, SRLKE, or DGLRV, wherein the mutant CBS has a T residue at position 7 of the consensus CBS motif, the amino acid sequence HTLKV, or HGLKV, wherein the mutant CBS has an A residue at position 7 of the consensus CBS motif; or the amino acid sequence SRLKE, HRLKE or N RLKE, wherein the mutant CBS has a C residue at position 7 of the consensus CBS motif" and "the engineered CTCF at ZF5 positions +2 to +6 comprising: the amino acid sequence AJLKR, QALRR, GGLVR, or HGLIR, wherein the mutant CBS has a T residue at position 8 of the consensus CBS motif; the amino acid sequence ANLSR, TIGLTR HGLNR, or GGLTR, wherein the mutant CBS has an A residue at position 8 of the consensus CBS motif; the amino acid sequence HTLRR, TVLTR, ADLKR, or HGLRR, wherein the mutant CBS has a C residue at position 8 of the consensus CBS motif" and "the engineered CTCF at ZF4 positions +2 to +6 comprising: the amino acid sequence AHIRK, wherein the mutant CBS has a T residue at position 10 of the consensus CBS motif; the amino acid sequence AKLRM, EKLRI, or AKLRI, wherein the mutant CBS has an A residue at position 10 of the consensus CBS motif; or the amino acid sequence TKLK, wherein the mutant CBS has a C residue at position 10 of the consensus CBS motif" and "the engineered CTCF at ZF4 positions +2 to +6 comprising: the amino acid sequence AHIRK, wherein the mutant CBS has a T residue at position 10 of the consensus CBS motif; the amino acid sequence AKLRM, EKLRI, or AKLRI, wherein the mutant CBS has an A residue at position 10 of the consensus CBS motif; or the amino acid sequence TKLK, wherein the mutant CBS has a C residue at position 10 of the consensus CBS motif" and "the engineered CTCF comprising: (i) the amino acid sequence AKLKK, AKLRK, AHLRV, AKLRV5 or SKLRL at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence ERLRV, NRLKV SRLKE, or NRLKV at ZF5 positions +2 to +6 of the engineered CTCF; (iii) the amino acid sequence RPD, RTET, or RADV at ZF6 positions -1 to +3 of the engineered CTCF; and (iv) the amino acid sequence DNLLA, SNLLV, DNLMA, or DNLR at ZF7 positions +2 to +6 of the engineered CTCF" and "the engineered CTCF comprising: (i) the amino acid sequence GHLKK, AHLRK, or GKLRI at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence SRLKE, DALRR, DGLKR, or TRLRE at ZF5 positions +2 to +6 of the engineered CTCF; (iii) the amino acid sequence at RPD, RTET, or RADV at ZF6 positions -1 to +3 of the engineered CTCF; and (iv) the amino acid sequence DNLLA, SNLLV, DNLMA, or DNLR at ZF7 positions +2 to +6 of the engineered CTCF" and "the engineered CTCF comprising: (i) the amino acid sequence EHLK, EHLAE, STLNE, DHLQV, EHLNV, DHLNT, EIHTLQA, or HKI-LM- at ZF7 positions +2 to +6 of the engineered CTCF" and "the engineered CTCF comprising: (i) the amino acid sequence GFHLKK, AHLKK, TKLRL, TKLKL, GHLRK, THLKK, or AHLRK at ZF4 positions +2 to +6 of the engineered CTCF; (ii) the amino acid sequence TRLKE or SRLKE at ZF5 positions +2 to +6 of the engineered CTCF; and (iii) the amino acid sequence RADN, RHDT, RRD, RPD, RTSS, or RNDT at ZF6 positions -1 to +3 of the engineered CTCF".

Additionally, even if Groups I+ were considered to share the technical features of an engineered CCCTC-binding factor (CTCF) variant comprising at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF, wherein the engineered CTCF variant binds to a mutant CTCF binding sequence (CBS) with a higher affinity than wild-type CTCF, the mutant CBS comprising at least one nucleotide base that differs in sequence from the nucleotide sequence of a

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2019/032937

consensus CBS; these shared technical features do not represent a contribution over the prior art.

Specifically, WO 2017/031370 A1 to The Broad Institute, Inc. discloses an engineered CCCTC-binding factor (CTCF) variant (we mutated three of the zinc-finger domains in CTCF proteins, Para. [0636]) comprising at least one amino acid residue in at least one zinc finger that differs in sequence from the amino acid sequence of a wild-type CTCF (we mutated three of the zinc-finger domains in CTCF proteins, Para. [0636]).

"Identification of critical base pairs required for CTCF binding in motif M1 and M2" to Li et al. discloses an engineered CCCTC-binding factor (CTCF) variant binds to a mutant CTCF binding sequence (CBS) with a higher affinity than wild-type CTCF (The binding of M1 to GST-CTCF-ZFs became stronger, Pg. 544, right-hand column, top of second paragraph), and a mutant CBS comprising a nucleotide base that differs in sequence from the nucleotide sequence of a consensus CBS (Genome-wide studies have identified that the majority of CTCF binding sites belongs to a set of nonpalindromic CTCF binding sites with a consensus sequence referred to as M1, Pg. 544, left-hand column, middle of first paragraph; we have looked into CTCF binding motif M1 and M2 and designed multiple point mutations according to the position weight matrix... [f]or this purpose, we synthesized M1 with a series of 3 bp mutation, Pg. 547, left-hand column, bottom of page).

The inventions listed in Groups I+ therefore lack unity under Rule 13 because they do not share a same or corresponding special technical features.