Generation of highly diverse peptide library by linear-double-stranded DNA based AND gate genetic circuit in mammalian cells

Weijun Su [1], Chunze Zhang [2], Shuai Li [3,4*]

1. School of Medicine, Nankai University, Tianjin 300071, China;

2. Department of Colorectal Surgery, Tianjin Union Medical Center, Tianjin 300121, China;

3. Department of Breast Cancer Pathology and Research Laboratory, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer; Key Laboratory of Cancer Prevention and Therapy, Tianjin; Tianjin's Clinical Research Center for Cancer, 300060, China;

4. State Key Laboratory of Medicinal Chemical Biology, Nankai University, Tianjin 300071, China.

* To whom correspondence should be addressed. Tel: +86 22 23340123; Fax: +86 22 23340123; Email: shuaili@tmu.edu.cn & shuai.li2001@gmail.com.

Abstract

Display technologies, e.g., phage, ribosome, mRNA, bacterial, and yeast-display, combine high content peptide libraries with appropriate screening strategies to identify functional peptide sequences. Construction of large peptide library and display-screen system in intact mammalian cells will facilitate the development of peptide therapeutics targeting transmembrane proteins. Our previous work established linear-double-stranded DNAs (ldsDNAs) as innovative biological parts to implement AND gate genetic circuits in mammalian cell line. In the current study, we employ ldsDNA with terminal NNK degenerate codons as AND gate input to build highly diverse peptide library in mammalian cells. Only PCR reaction and cell transfection experiments are needed to construct the library. High-throughput sequencing (HTS) results reveal that our new strategy could generate peptide library with both amino acid sequence and peptide length diversities. Our work establishes ldsDNA as biological parts for building highly diverse peptide library in mammalian cells, which shows great application potential in developing therapeutic peptides targeting transmembrane proteins.

Keywords

peptide library, linear-double-stranded DNA (ldsDNA), AND gate, degenerate codon.

Introduction

Different state-of-the-art display technologies, e.g., phage, ribosome, mRNA, bacterial, and yeast-display, have been developed to identify functional peptides (1-7). Through construction and screen vast peptide libraries, these display technologies physically couple between phenotype (high affinity high selectivity binding) and genotype (DNA sequence). However, transmembrane proteins such as G protein-coupled receptors (GPCRs), ion channels have almost been excluded from current display screening methods (8). Although a number of different approaches have recently been developed to facilitate the display of transmembrane proteins (9-16), the construction of highly diverse peptide library and the development of functional screen system in intact mammalian cells are the long-term goals of peptide pharmaceutics (8).

Recently, Lerner and colleagues have developed an autocrine-based signaling system in mammalian cell (17-22). Through membrane-tethered peptide library and artificial signal transduction system, this new strategy could screen the agonist/antagonist peptides of GPCRs or ion channels. However, for Lerner's strategy, the diversity information of peptide library is stored in plasmid vectors and is constantly lost during the process of plasmid extraction, lentiviral library generation and cell infection (17-19). As peptide library capacity serves as a determinant factor of a successful screen, the development of novel methodology for building high content peptide library in mammalian cell will greatly accelerate the development of therapeutic peptide targeting transmembrane proteins.

We previously applied linear-double-stranded DNA (ldsDNA, or PCR amplicon) as novel biological parts to implement Boolean logic AND gate genetic circuits in mammalian cells (23,24). Via splitting essential gene expression cassette into signal-mute ldsDNA, our strategy achieves two or three-input AND gate calculation with a low noise-signal ratio both *in vitro* and *in vivo*. This process is similar to V(D)J recombination, which serves as the fundamental molecular mechanism to generate highly diverse repertoires of antigen-receptors (antibodies and T-cell receptors) in mammalian (25,26). In the present study, we employ terminal-NNKs-ldsDNA-based AND-gate genetic circuit to generate high content peptide library.

Materials and Methods

Cell culture

The HEK293T cell line was maintained in Dulbecco's modified Eagle's medium (DMEM) (Thermo Fisher Scientific, Hudson, NH, USA) containing 10% fetal bovine serum (FBS) with 1% penicillin-streptomycin solution at 37 degree with 5% CO_2.

ldsDNA synthesis

KOD-Plus-Ver.2 DNA polymerase (Toyobo, Osaka, Japan) was used to amplify ldsDNAs (PCR amplicons) using the pEGFP-C1 (Clontech, Mountain View, CA, USA) plasmid as the template. PCR products underwent agarose electrophoresis and gel-purification to remove plasmid template and free dNTPs (Universal DNA purification kit, Tiangen, Beijing, China). The amount of PCR products was determined by the OD 260 absorption value. PCR primer

sequences and ldsDNA sequences were present in Supplementary Table 1 and Supplementary Materials, respectively.

ldsDNA transfection

HEK293T cells were seeded into 6-well plates the day before transfection (60% - 70% confluency). 500 ng/well of each input amplicons (total 1000 ng/well) were transfected into cells with Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA) reagent following standard protocol.

RNA extraction, cDNA synthesis and PCR amplification

Forty-eight hours after transfection, total RNAs were extracted by RNAiso Plus (TaKaRa, Beijing, China) reagent. cDNAs were synthesized using the PrimeScript ™ RT reagent Kit with gDNA Eraser (TaKaRa) following standard protocol. KOD-Plus-Ver.2 DNA polymerase (Toyobo) was used to amplify the corresponding cDNA sequences surrounding ldsDNA junctions.

Library construction and high-throughput sequencing

Library construction and high-throughput sequencing were conduct by Novogene (Beijing, China). Sequencing libraries were generated using TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, San Diego, CA, USA) following the manufacturer's recommendations and index codes were added. The library quality was assessed on the Qubit Fluorometer and Agilent Bioanalyzer 2100 system. At last, the libraries were sequenced on an Illumina HiSeq 2500 platform and 250 bp paired-end reads were generated.

Data analysis

Paired-end reads from the original DNA fragments were merged using FLASH, a very fast and accurate analysis tool which was designed to merge paired-end reads when there are overlaps between reads 1 and reads 2. Paired-end reads were assigned to each sample according to the unique barcodes. Reads were filtered by QIIME quality filters. Nucleotide sequences generated by ldsDNA-based AND-gate genetic circuits were extracted by in-house Perl scripts.

Results

Generation of peptide library by ldsDNA-based AND gate genetic circuit

Our previous study demonstrated that ldsDNAs (PCR amplicons), split from intact gene expression cassettes (containing promoter, gene coding sequence and polyA signal), could undergo reconnection to form one output-signal-generating molecule in cultured cells (23,24). Here we hypothesize that ldsDNAs with terminal NNK degenerate codons could generate highly diverse peptide library through AND-gate calculation in mammalian cells (Figure 1).

Taking pEGFP-C1 plasmid as PCR template, primers with multiple tandem NNK-trinucleotides were used to generate ldsDNAs with terminal degenerate codons (Supplementary Table 1). Three kinds of ldsDNAs, containing CMV promoter, Kozak sequence and different number of NNK-trinucleotides, were produced (namely CMV-Kozak-$<NNK>_5$, CMV-Kozak-$<NNK>_{10}$, CMV-Kozak-$<NNK>_{15}$) (for sequences see Supplementary Materials). Meanwhile, two kinds of ldsDNAs, containing terminal NNK-trinucleotides (or not), GFP coding sequence and SV40 polyA signal (PAS), were

amplified (namely <NNK>$_5$-GFP-SV40 PAS, <NNK>$_0$-GFP-SV40 PAS). Six different combinations of these two subgroup ldsDNAs were introduced into HEK293T cells to construct the AND-gate genetic circuits. The six combinations are as following: 1, CMV-Kozak-<NNK>$_5$ with <NNK>$_5$-GFP-SV40 PAS (5X+5X for short); 2, CMV-Kozak-<NNK>$_{10}$ with <NNK>$_5$-GFP-SV40 PAS (10X+5X); 3, CMV-Kozak-<NNK>$_{15}$ with <NNK>$_5$-GFP-SV40 PAS (15X+5X); 4, CMV-Kozak-<NNK>$_5$ with <NNK>$_0$-GFP-SV40 PAS (5X+0X); 5, CMV-Kozak-<NNK>$_{10}$ with <NNK>$_0$-GFP-SV40 PAS (10X+0X); 6, CMV-Kozak-<NNK>$_{15}$ with <NNK>$_0$-GFP-SV40 PAS (15X+0X).

When AND-gate genetic circuits are built up in cells, peptide library would be generated surrounding the junctions formed between these two subtype ldsDNAs (Figure 1). Forty-eight hours after transfection, total RNAs were extracted then cDNAs were synthesized via reverse transcription. Nucleotide sequences surrounding the junctions formed by two ldsDNA-inputs were PCR amplified and then subjected to pair-end high-throughput sequencing (HTS). The HTS statistics for all sequencing libraries are shown in Table 1.

Reading frame phase of ldsDNA-based peptide library

To couple phenotype to genotype, display technologies usually fuse oligopeptide to specific anchor proteins. For example in phage display experiment, the peptide library is fused to an open reading frame (ORF) thereby connecting peptide to phage coat protein (1). Our previous research demonstrated that the ldsDNA could be subject to terminal nucleotide(s) deletion which may lead to reading frame shift.

Here we analyze the reading frame phase characteristics of the peptide libraries (Figure 2). Libraries generated by dual-side terminal-NNKs-ldsDNAs, e.g., CMV-Kozak-<NNK>$_5$ with <NNK>$_5$-GFP-SV40 PAS (5X+5X), show a similar reading frame phase distribution. The average in-frame (3n) ratio ranging from 45% to 47%. While libraries generated by single-side terminal-NNKs-ldsDNAs, e.g., CMV-Kozak-<NNK>$_5$ with <NNK>$_0$-GFP-SV40 PAS (5X+0X), show dramatic dynamics of reading frame phase distribution (Figure 2). 5X+0X peptide libraries reach the maximum 69% in-frame rate compared with 63% of 10X+0X and 53% of 15X+0X, respectively. Then, the in-frame nucleotide sequences (3n) are translated into polypeptides for further analysis.

Length dynamic and amino acid composition of ldsDNA-based peptide library

We next analyze the polypeptide length distributions of different peptide libraries (peptides with pre-stop codon are excluded). Dual-side terminal-NNKs-ldsDNAs strategies, including 5X+5X, 10X+5X and 15X+5X, show a more dynamic length distribution compared with that from single-side terminal-NNKs-ldsDNAs strategies (5X+0X, 10X+0X and 15X+0X) (Figure 3). In addition, for single-side terminal-NNKs-ldsDNAs strategies most peptides have a length equal to the theoretical polypeptide length generated by direct ligation (supposing no insertion or deletion happened) (Figure 3D-F). For example, in the 5X+0X group, over 81% of the peptide length is five amino acids.

The amino acid composition of different peptide libraries is in line with the theoretical NNK amino acid composition (pre-stop codon is included in the analysis) (Table 2). However, Proline (P) residue always shows a higher percentage than the theoretical ratio, particularly

when the peptide chain getting longer. Meanwhile, the percentage of each nucleotide at each position of peptide library codons is showed in Supplementary Table 2. The nucleotide distributions are in line with NNK degenerate base ratio.

The capacity of ldsDNA-based peptide library

Peptide library capacity could be represented by its coverage ability on amino acid combination possibilities. Here, all eighteen peptide libraries cover all 400 possible dual-amino-acid combinations (2-mer). For triple-amino-acid (3-mer), the library coverage is ranging from 64% to 98% (Figure 4A). For quadruple-amino-acid (4-mer), the library coverage is ranging from 5.7% to 32.1% (Figure 4B). The peptide library content, defined as the total amino acid number coded by a library, serves as a determinate factor of library capacity (Figure 4).

Discussion

DNA-encoded peptide libraries are ideal peptide-drug discovery platforms for their extremely large capacity. Currently, plasmids and PCR products have been explored as genetic mediums to store highly diverse coding information of the peptide library (8,27). These genetic materials are introduced into phage-host bacteria, yeast, or test tubes to transcript-translate into peptides. However, it’s still a big challenge to build a large peptide library in intact mammalian cells. Here by cell-based AND-gate computation of terminal-NNKs-ldsDNAs, we develop a novel method to generate highly diverse peptide library in mammalian cells. Our protocol is easy to conduct only need PCR-reaction and cell-transfection. High-throughput

sequencing results reveal that ldsDNA-based peptide library is characterized by both amino acids sequence diversity and peptide length dynamics. Our new method will facilitate the development of peptide therapeutics targeting transmembrane proteins.

Transmembrane proteins, such as G protein-coupled receptors (GPCRs) and ion channels, represent important drug targets since they are involved in regulating a wide range of pivotal physiological processes (28-32). Different approaches have been developed to transform transmembrane proteins as suitable screen targets for various display platforms (14-16). However, it's still hard to immobilize transmembrane protein and pan display libraries. In addition, due to the complex structure and sophisticated signal transduction mode of transmembrane proteins, current display techniques could not directly connect peptide-protein affinity to a functional hit. Ideally, the combination of high content peptide library with functional signal screen systems in intact mammalian cells would allow efficient panning of peptide library under physiological condition. Recently, Lerner and colleagues have developed a powerful autocrine-based signaling system which tethered peptide library to cell membrane and employed artificial signal transduction systems to observe peptide-transmembrane protein interaction (17-22). Here, we provide a novel methodology for building high content peptide library in mammalian cells. These progresses highlight mammalian cells as a potential powerful platform for peptide drug development.

Affinity-based display technologies have been developed to screen peptide libraries to identify protein-binding peptides. Peptide library size is a determinate factor of a success screen. For phage and yeast-display, the complexity of peptide library stores in plasmid (1,2,7). Currently, the phage display library size could reach 10^9 (27,33). While for ribosome

and mRNA-display, PCR products server as genetic medium to store library coding information. Ribosome and mRNA-display could achieve extremely large library size (>10^{13}) (5,34). For mammalian cells, plasmid or lentivirus have been used to introduce peptide library (17-20,35). However, the peptide library capacity is constantly lost during plasmid extraction, lentivirus generation, transfection/infection process. In the current research, we provide a new strategy to build high content peptide library in cultured cell line. The protocol is easy to conduct, only containing two steps: PCR amplification and cell transfection. Since degenerate trinucleotides is present at the 5' end of the primer, the library complexity keeps on increasing during PCR amplification process. Theoretically, no specific nucleotide sequence is enriched during PCR reaction. PCR products are then transfected into cells to conduct AND-gate linkage to generate full gene expression cassette (promoter-CDS-PAS). For our strategy, no particular attention needs to be paid to maintain the library abundance. The size of ldsDNA-based library needs to be further investigated. Determining the peptide sequence diversity in each cell by single-cell-sequencing will provide crucial information for estimating the whole library size.

For current protocol, there are still several issues that need to be further addressed to increase library capacity. First of all, the high in-frame ratio would improve library capacity. Our results reveal that peptide libraries generated by single-side NNKs-ldsDNA show higher in-frame ratios than that from dual-side NNKs-ldsDNAs (Figure 2). Besides, chemical modified primers could be used to generate nuclease resistance ldsDNAs to prevent reading frame shift (36,37). Another major obstacle is the pre-stop codon, which causes premature termination (38). For NNK degenerate codon, the percentage of stop-codon in each amino

acid position is 3.125% (1/32) and will be accumulated when the peptide is getting longer. When the polypeptide length reaches 20 amino acids, about 47% of the sequences contain pre-stop codon. One potential solution to avoid stop codon is to synthesize PCR primer using trimer phosphoramidites (39,40).

In summary, we apply ldsDNA-based AND gate calculation to generate highly diverse peptide library in mammalian cells, which may have great application potential in therapeutic peptide development.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [31870860, 31400673 to S.L., 81402407 to W.S.]; the Tianjin Research Program of Application Foundation and Advanced Technology [14JCQNJC09800 to S.L., 15JCQNJC11700 to W.S.]; the State Key Laboratory of Medicinal Chemical Biology (Nankai University) [2018103 to S.L.]; and the Key Project of National Health and Family Planning Commission of Tianjin [2017057 to C.Z.]. The authors have declared that no competing interest exists.

Author Contributions

S.L. conceived and supervised the study. S.L. and W.S. designed and performed the experiments. S.L. and W.S. analyzed the data. W.S. and S.L. prepared the figures. S.L. and W.S. wrote the paper. All authors discussed the results and reviewed the manuscript.

Competing Interests

The authors have declared that no competing interest exists.

Figure Legends

Figure 1. Schematic showing the rationale of generating highly diverse peptide library by ldsDNA-based AND-gate calculation. ldsDNAs containing CMV promoter and tandem NNK degenerate codons are cotransfected with ldsDNAs containing NNK codons, GFP coding region, and polyA signal sequence. The AND-gate calculation is conducted between two kinds of ldsDNA to form peptide library.

Figure 2. Reading frame character of nucleotide sequences formed by different libraries. Pair-end high-throughput sequencing data is processed to extract the nucleotide sequences formed by ldsDNA-based AND-gate. The nucleotide sequence length is divided by three to show its reading frame character. The average percentage of different phase of different libraries is presented. All data are displayed as the mean ± SD, n = 3.

Figure 3. The peptide length distributions of six different peptide libraries generated by ldsDNA-based AND-gate genetic circuits. In-frame (3n) nucleotide sequences are translated into polypeptide sequences. Then the length of polypeptide sequence in each library is summarized. The average percentage of different polypeptide length is presented. All data are displayed as the mean ± SD, n = 3.

Figure 4. Peptide library coding capacity on covering triple-amino-acid (A) or quadruple-amino-acid (B) peptide possibilities. Peptide libraries are scanned by all triple or quadruple amino-acid combinations, respectively. The library content (defined as the total amino acid number coded by a library) and the percentage of coverage are presented.

References

1. Smith, G.P. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science, 228, 1315-1317.

2. Scott, J.K. and Smith, G.P. (1990) Searching for peptide ligands with an epitope library. Science, 249, 386-390.

3. Mattheakis, L.C., Bhatt, R.R. and Dower, W.J. (1994) An in vitro polysome display system for identifying ligands from very large peptide libraries. Proceedings of the National Academy of Sciences of the United States of America, 91, 9022-9026.

4. Hanes, J. and Pluckthun, A. (1997) In vitro selection and evolution of functional proteins by using ribosome display. Proceedings of the National Academy of Sciences of the United States of America, 94, 4937-4942.

5. Roberts, R.W. and Szostak, J.W. (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. Proceedings of the National Academy of Sciences of the United States of America, 94, 12297-12302.

6. Georgiou, G., Stathopoulos, C., Daugherty, P.S., Nayak, A.R., Iverson, B.L. and Curtiss, R., 3rd. (1997) Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines. Nature biotechnology, 15, 29-34.

7. Boder, E.T. and Wittrup, K.D. (1997) Yeast surface display for screening combinatorial polypeptide libraries. Nature biotechnology, 15, 553-557.

8. Henninot, A., Collins, J.C. and Nuss, J.M. (2018) The Current State of Peptide Drug Discovery: Back to the Future? Journal of medicinal chemistry, 61, 1382-1414.

9. Hardy, D., Bill, R.M., Jawhari, A. and Rothnie, A.J. (2016) Overcoming bottlenecks in the membrane protein structural biology pipeline. Biochemical Society transactions, 44, 838-844.

10. Matar-Merheb, R., Rhimi, M., Leydier, A., Huche, F., Galian, C., Desuzinges-Mandon, E., Ficheux, D., Flot, D., Aghajari, N., Kahn, R. et al. (2011) Structuring detergents for extracting and stabilizing functional membrane proteins. PloS one, 6, e18036.

11. Bayburt, T.H. and Sligar, S.G. (2003) Self-assembly of single integral membrane proteins into soluble nanoscale phospholipid bilayers. Protein science : a publication of the Protein Society, 12, 2476-2481.

12. Bayburt, T.H. and Sligar, S.G. (2010) Membrane protein assembly into Nanodiscs. FEBS letters, 584, 1721-1727.

13. Scott, D.J., Kummer, L., Tremmel, D. and Pluckthun, A. (2013) Stabilizing membrane proteins through protein engineering. Current opinion in chemical biology, 17, 427-435.

14. Dominik, P.K. and Kossiakoff, A.A. (2015) Phage display selections for affinity reagents to membrane proteins in nanodiscs. Methods in enzymology, 557, 219-245.

15. Congreve, M., Rich, R.L., Myszka, D.G., Figaroa, F., Siegal, G. and Marshall, F.H. (2011) Fragment screening of stabilized G-protein-coupled receptors using biophysical methods. Methods in enzymology, 493, 115-136.

16. Wingler, L.M., McMahon, C., Staus, D.P., Lefkowitz, R.J. and Kruse, A.C. (2019) Distinctive Activation Mechanism for Angiotensin Receptor Revealed by a Synthetic Nanobody. Cell, 176, 479-490 e412.

17. Zhang, H., Sturchler, E., Zhu, J., Nieto, A., Cistrone, P.A., Xie, J., He, L., Yea, K., Jones, T., Turn, R. et al. (2015) Autocrine selection of a GLP-1R G-protein biased agonist with potent antidiabetic effects. Nature communications, 6, 8918.

18. Xie, J., Zhang, H., Yea, K. and Lerner, R.A. (2013) Autocrine signaling based selection of combinatorial antibodies that transdifferentiate human stem cells. Proceedings of the National Academy of Sciences of the United States of America, 110, 8099-8104.

19. Stepanov, A.V., Markov, O.V., Chernikov, I.V., Gladkikh, D.V., Zhang, H., Jones, T., Sen'kova, A.V., Chernolovskaya, E.L., Zenkova, M.A., Kalinin, R.S. et al. (2018) Autocrine-based selection of ligands for personalized CAR-T therapy of lymphoma. Science advances, 4, eaau4580.

20. Zhang, H., Yea, K., Xie, J., Ruiz, D., Wilson, I.A. and Lerner, R.A. (2013) Selecting agonists from single cells infected with combinatorial antibody libraries. Chemistry & biology, 20, 734-741.

21. Zhang, H., Xie, J. and Lerner, R.A. (2014) A proximity based general method for identification of ligand and receptor interactions in living cells. Biochemical and biophysical research communications, 454, 251-255.

22. Zhang, H., Du, M., Xie, J., Liu, X., Sun, J., Wang, W., Xin, X., Possani, L.D., Yea, K. and Lerner, R.A. (2016) Autocrine-Based Selection of Drugs That Target Ion Channels from Combinatorial Venom Peptide Libraries. Angew Chem Int Ed Engl, 55, 9306-9310.

23. Li, S., Su, W. and Zhang, C. (2019) Linear Double-stranded DNAs as Innovative Biological Parts to Implement Genetic Circuits in Mammalian Cells. bioRxiv, 266056.

24. Li, S., Su, W. and Zhang, C. (2019) Linear double-stranded DNAs as innovative biological parts to implement genetic circuits in mammalian cells. The FEBS journal, 286, 2341-2354.

25. Bassing, C.H., Swat, W. and Alt, F.W. (2002) The mechanism and regulation of chromosomal V(D)J recombination. Cell, 109 Suppl, S45-55.

26. Schatz, D.G. (2004) V(D)J recombination. Immunological reviews, 200, 5-11.

27. Obexer, R., Walport, L.J. and Suga, H. (2017) Exploring sequence space: harnessing chemical and biological diversity towards new peptide leads. Current opinion in chemical biology, 38, 52-61.

28. Rask-Andersen, M., Masuram, S. and Schioth, H.B. (2014) The druggable genome: Evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication. Annual review of pharmacology and toxicology, 54, 9-26.

29. Hauser, A.S., Attwood, M.M., Rask-Andersen, M., Schioth, H.B. and Gloriam, D.E. (2017) Trends in GPCR drug discovery: new agents, targets and indications. Nature reviews. Drug discovery, 16, 829-842.

30. de Lera Ruiz, M. and Kraus, R.L. (2015) Voltage-Gated Sodium Channels: Structure, Function, Pharmacology, and Clinical Indications. Journal of medicinal chemistry, 58, 7093-7118.

31. Waxman, S.G. and Zamponi, G.W. (2014) Regulating excitability of peripheral afferents: emerging ion channel targets. Nature neuroscience, 17, 153-163.

32. Zamponi, G.W. (2016) Targeting voltage-gated calcium channels in neurological and psychiatric diseases. Nature reviews. Drug discovery, 15, 19-34.

33. Smith, G.P. and Petrenko, V.A. (1997) Phage Display. Chemical reviews, 97, 391-410.

34. Lamla, T. and Erdmann, V.A. (2003) Searching sequence space for high-affinity binding peptides using ribosome display. Journal of molecular biology, 329, 381-388.

35. Ho, M., Nagata, S. and Pastan, I. (2006) Isolation of anti-CD22 Fv with high affinity by Fv display on human cells. Proceedings of the National Academy of Sciences of the United States of America, 103, 9637-9642.

36. Rahal, E.A., Henricksen, L.A., Li, Y., Turchi, J.J., Pawelczak, K.S. and Dixon, K. (2008) ATM mediates repression of DNA end-degradation in an ATP-dependent manner. DNA repair, 7, 464-475.

37. Chang, H.H., Watanabe, G., Gerodimos, C.A., Ochi, T., Blundell, T.L., Jackson, S.P. and Lieber, M.R. (2016) Different DNA End Configurations Dictate Which NHEJ Components Are Most Important for Joining Efficiency. The Journal of biological chemistry, 291, 24377-24389.

38. Cho, G., Keefe, A.D., Liu, R., Wilson, D.S. and Szostak, J.W. (2000) Constructing high complexity synthetic libraries of long ORFs using in vitro selection. Journal of molecular biology, 297, 309-319.

39. Janczyk, M., Appel, B., Springstubbe, D., Fritz, H.J. and Muller, S. (2012) A new and convenient approach for the preparation of beta-cyanoethyl protected trinucleotide phosphoramidites. Organic & biomolecular chemistry, 10, 1510-1513.

40. Suchsland, R., Appel, B. and Muller, S. (2018) Preparation of trinucleotide phosphoramidites as synthons for the synthesis of gene libraries. Beilstein journal of organic chemistry, 14, 397-406.

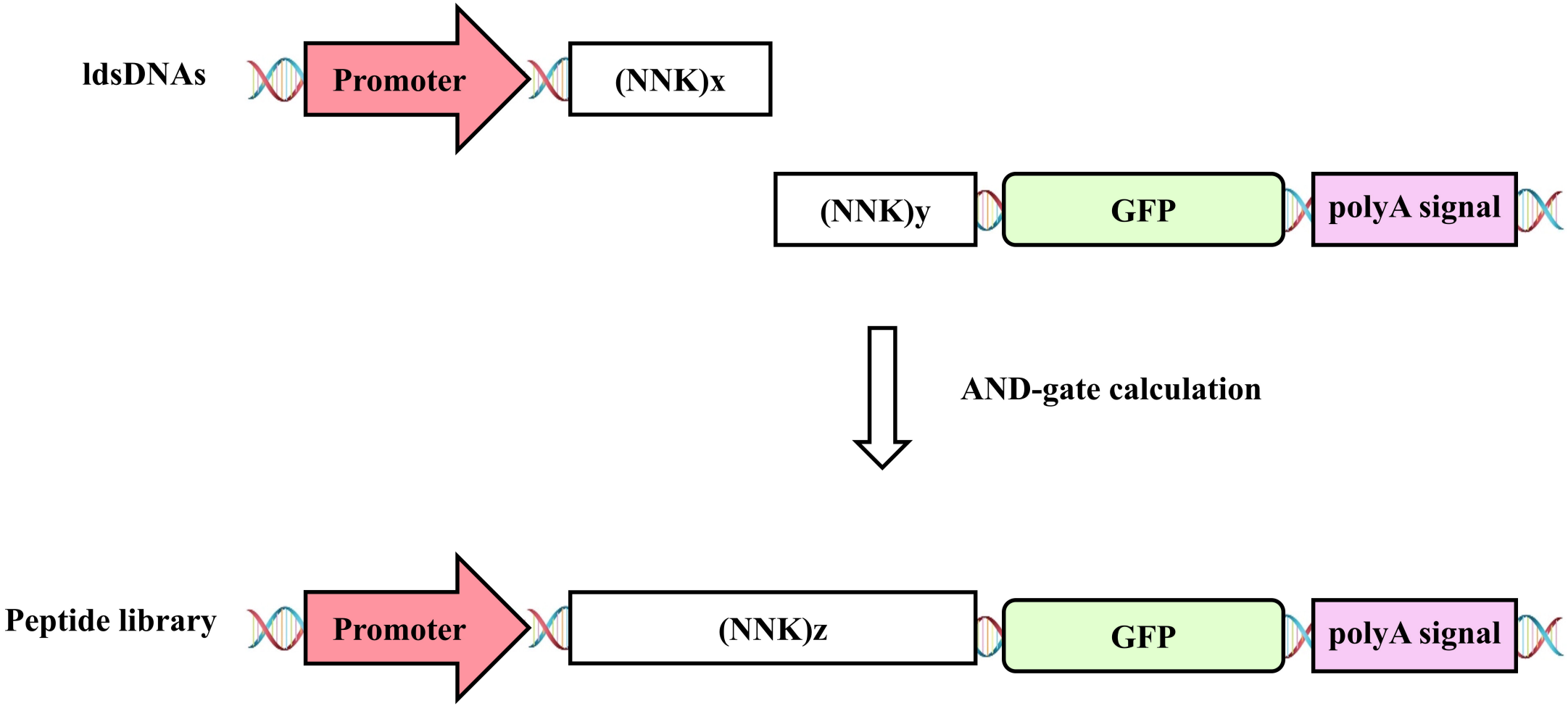
ldsDNAs
Promoter
(NNK)x
(NNK)y
GFP
polyA signal
AND-gate calculation
Peptide library
Promoter
(NNK)z
GFP
polyA signal

Figure 1

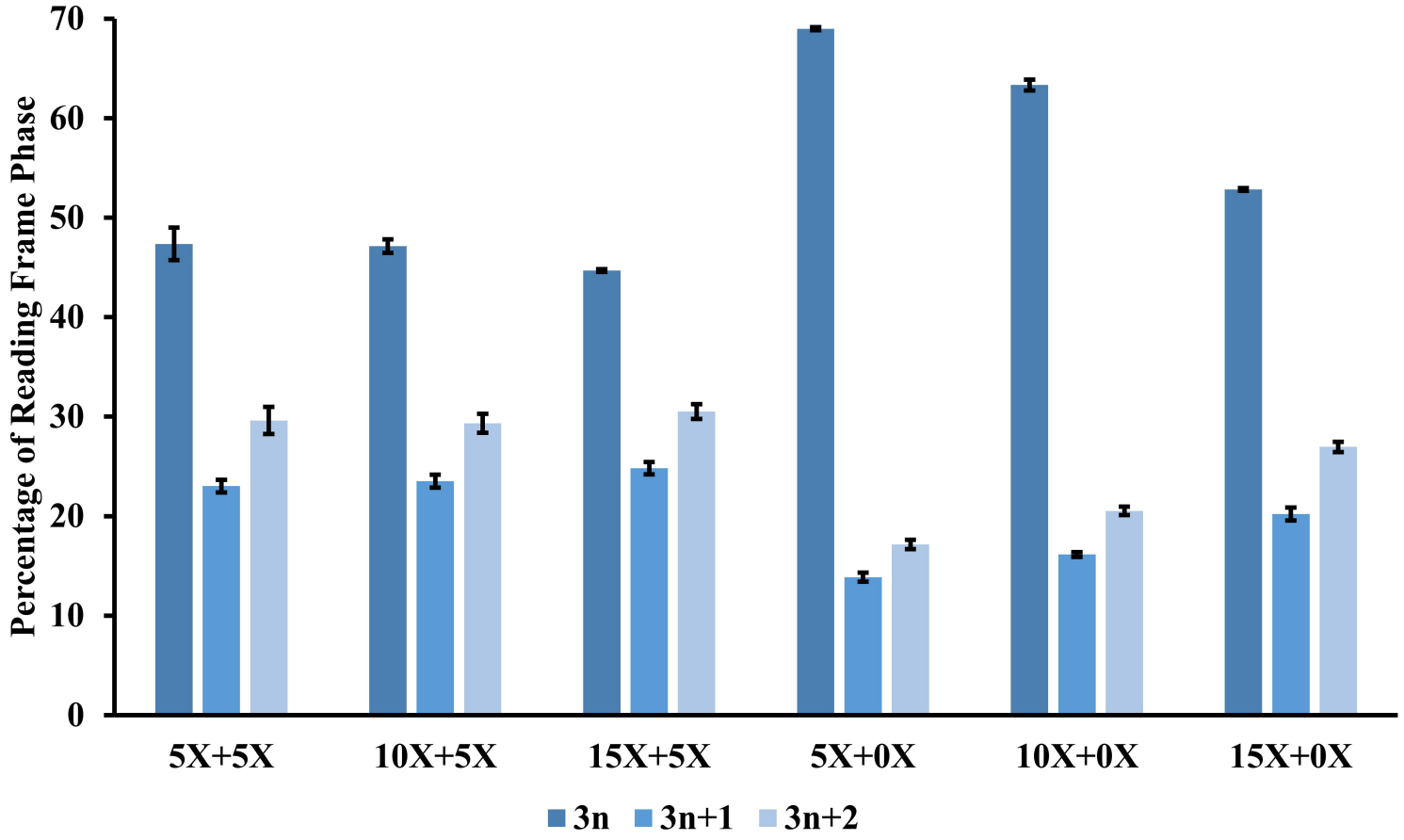
Percentage of Reading Frame Phase
70
60
50
40
30
20
10
0
5X+5X
10X+5X
15X+5X
5X+0X
10X+0X
15X+0X
3n
3n+1
3n+2

Figure 2

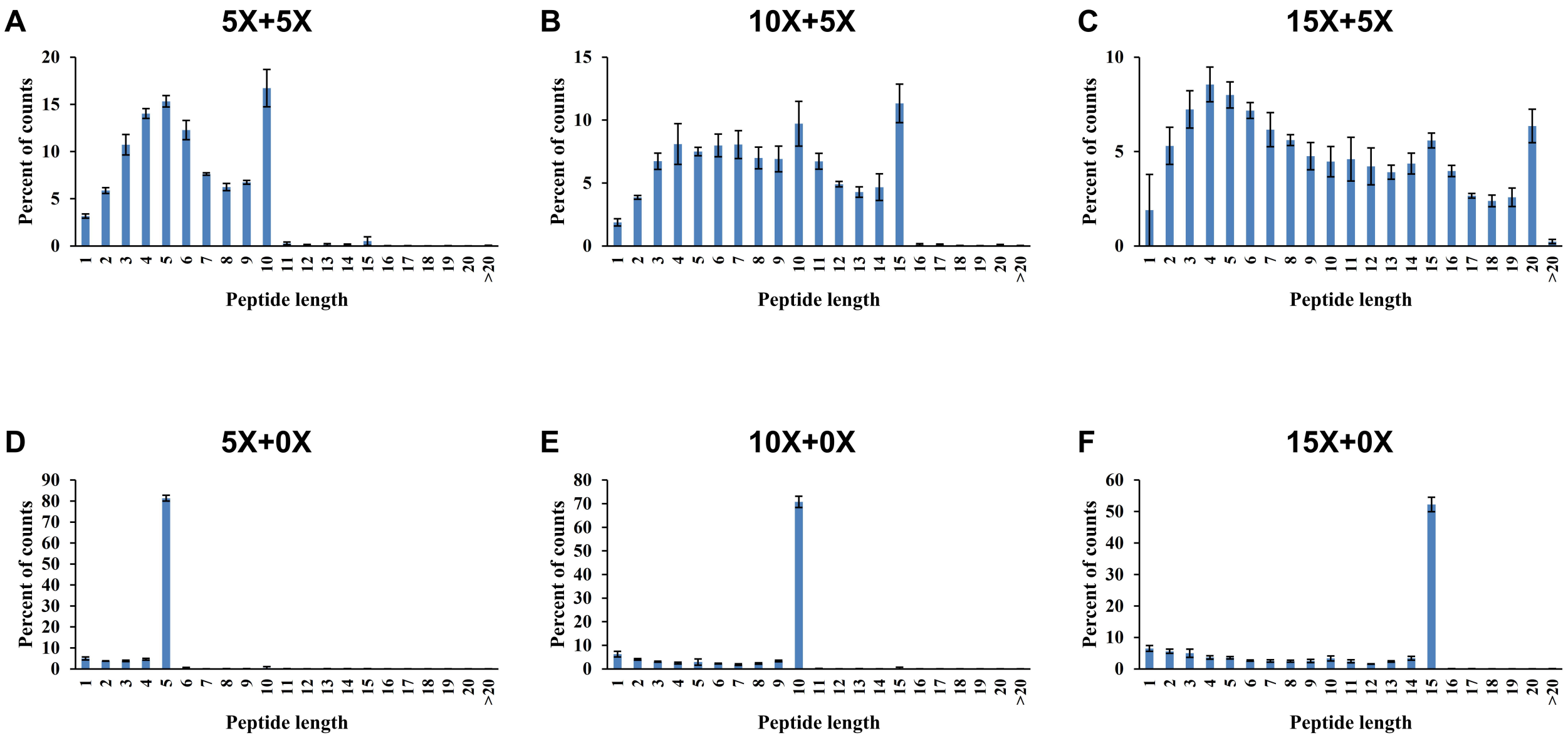
A
5X+5X
B
10X+5X
C
15X+5X
D
5X+0X
E
10X+0X
F
15X+0X
Percent of counts
Peptide length
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 >20

Figure 3

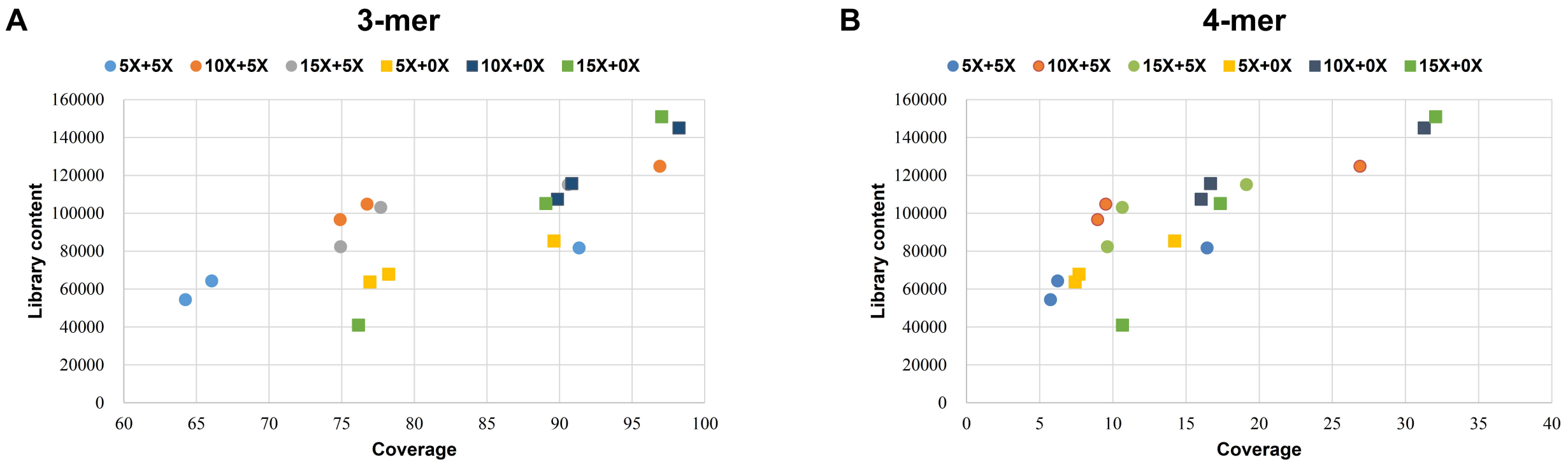
A
3-mer
5X+5X
10X+5X
15X+5X
5X+0X
10X+0X
15X+0X
Library content
Coverage
B
4-mer
5X+5X
10X+5X
15X+5X
5X+0X
10X+0X
15X+0X
Library content
Coverage

Figure 4

Table 1. High-throughput sequencing (HTS) statistics

Library	Description	Clean Read count	Read count in analysis
5X+5X (1)	Peptide library generated by ldsDNA CMV-Kozak-<NNK>$_5$ and <NNK>$_5$-GFP-SV40 PAS	79179	32698
5X+5X (2)		75950	25194
5X+5X (3)		61489	23021
10X+5X (1)	Peptide library generated by ldsDNA CMV-Kozak-<NNK>$_{10}$ and <NNK>$_5$-GFP-SV40 PAS	84191	38218
10X+5X (2)		75195	30324
10X+5X (3)		71341	32768
15X+5X (1)	Peptide library generated by ldsDNA CMV-Kozak-<NNK>$_{15}$ and <NNK>$_5$-GFP-SV40 PAS	67121	32273
15X+5X (2)		39153	24083
15X+5X (3)		82312	31784
5X+0X (1)	Peptide library generated by ldsDNA CMV-Kozak-<NNK>$_5$ and <NNK>$_0$-GFP-SV40 PAS	87718	30612
5X+0X (2)		83744	22487
5X+0X (3)		81241	23620
10X+0X (1)	Peptide library generated by ldsDNA CMV-Kozak-<NNK>$_{10}$ and <NNK>$_0$-GFP-SV40 PAS	82228	33130
10X+0X (2)		74850	25512
10X+0X (3)		86618	26676
15X+0X (1)	Peptide library generated by ldsDNA CMV-Kozak-<NNK>$_{15}$ and <NNK>$_0$-GFP-SV40 PAS	85127	33192
15X+0X (2)		30998	9570
15X+0X (3)		75915	23598

Pair-end high-throughput sequencing statistics for all libraries referenced in this study. Sequences formed by ldsDNA-based AND-gate calculation are subject for further analysis.

Table 2. Percentage of each amino acid of different peptide library

Amino acid	NNK	5X+5X	10X+5X	15X+5X	5X+0X	10X+0X	15X+0X
Stop codon	3.125	2.34 ± 0.11	2.46 ± 0.14	2.28 ± 0.15	2.59 ± 0.14	2.47 ± 0.09	2.31 ± 0.13
A (Ala)	6.25	6.16 ± 0.08	6.34 ± 0.18	7.11 ± 0.19	5.43 ± 0.12	6.22 ± 0.13	7.32 ± 0.26
C (Cys)	3.125	2.80 ± 0.10	2.72 ± 0.09	2.74 ± 0.24	2.06 ± 0.17	2.35 ± 0.03	2.16 ± 0.09
D (Asp)	3.125	2.84 ± 0.18	2.96 ± 0.26	2.76 ± 0.07	3.72 ± 0.07	3.13 ± 0.07	2.85 ± 0.05
E (Glu)	3.125	2.74 ± 0.07	2.75 ± 0.05	2.63 ± 0.11	3.67 ± 0.11	3.24 ± 0.12	3.19 ± 0.10
F (Phe)	3.125	3.34 ± 0.18	3.34 ± 0.08	2.88 ± 0.08	1.71 ± 0.05	2.52 ± 0.12	1.88 ± 0.04
G (Gly)	6.25	5.23 ± 0.23	5.28 ± 0.14	5.85 ± 0.08	4.35 ± 0.07	4.89 ± 0.13	5.68 ± 0.14
H (His)	3.125	3.54 ± 0.17	3.41 ± 0.18	3.26 ± 0.20	4.35 ± 0.11	3.72 ± 0.14	3.21 ± 0.10
I (Ile)	3.125	3.45 ± 0.08	3.40 ± 0.38	2.76 ± 0.12	3.49 ± 0.26	3.05 ± 0.12	2.48 ± 0.07
K (Lys)	3.125	3.57 ± 0.16	3.20 ± 0.06	2.70 ± 0.07	5.93 ± 0.17	4.08 ± 0.09	3.32 ± 0.15
L (Leu)	9.375	8.84 ± 0.16	9.35 ± 0.17	8.98 ± 0.28	6.69 ± 0.22	8.63 ± 0.11	7.93 ± 0.12
M (Met)	3.125	2.91 ± 0.09	2.90 ± 0.18	2.68 ± 0.03	3.54 ± 0.10	3.04 ± 0.04	2.59 ± 0.08
N (Asn)	3.125	4.23 ± 0.18	3.55 ± 0.16	2.90 ± 0.07	6.33 ± 0.09	4.37 ± 0.12	3.35 ± 0.12
P (Pro)	6.25	8.16 ± 0.15	9.20 ± 0.52	10.40 ± 0.28	7.34 ± 0.10	9.41 ± 0.16	11.32 ± 0.11
Q (Gln)	3.125	3.13 ± 0.11	3.12 ± 0.04	3.17 ± 0.09	4.13 ± 0.11	3.89 ± 0.02	3.82 ± 0.09
R (Arg)	9.375	8.34 ± 0.28	8.47 ± 0.24	9.79 ± 0.23	8.74 ± 0.08	9.18 ± 0.25	10.68 ± 0.29
S (Ser)	9.375	10.30 ± 0.16	9.76 ± 0.30	9.68 ± 0.20	8.54 ± 0.14	8.63 ± 0.17	8.90 ± 0.32
T (Thr)	6.25	7.24 ± 0.17	6.64 ± 0.27	6.94 ± 0.12	8.88 ± 0.21	7.27 ± 0.17	7.58 ± 0.15
V (Val)	6.25	5.66 ± 0.24	5.56 ± 0.19	5.31 ± 0.21	3.59 ± 0.11	4.58 ± 0.07	4.51 ± 0.21
W (Trp)	3.125	2.30 ± 0.06	2.70 ± 0.31	2.60 ± 0.04	1.96 ± 0.07	2.33 ± 0.05	2.36 ± 0.07
Y (Tyr)	3.125	2.89 ± 0.11	2.89 ± 0.21	2.59 ± 0.07	2.96 ± 0.11	2.99 ± 0.14	2.57 ± 0.16

In-frame (3n) nucleotide sequences generated by ldsDNA-based AND-gate are translated into polypeptide sequences. The total number of each amino acid in each

library is summarized. The average percentage of different amino acid is presented. All data are displayed as the mean ± SD, n = 3. The theoretical amino acid percentage of NNK degenerate codon is also presented.

Supplementary Table 1. List of PCR primers used in study.

Amplicon	Forward primer	Reverse primer
CMV-Kozak-<NNK>$_5$	CTTTCCTGCGTTATCCCCTGATTCT	MNNMNNMNNMNNMNNCACCATGGTGGCGACCGGTAGCGCT
CMV-Kozak-<NNK>$_{10}$	CTTTCCTGCGTTATCCCCTGATTCT	MNNMNNMNNMNNMNNMNNMNNMNNMNNMNNCACCATGGTGGCGACCGGTAGCGCT
CMV-Kozak-<NNK>$_{15}$	CTTTCCTGCGTTATCCCCTGATTCT	MNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNMNNCACCATGGTGGCGACCGGTAGCGCT
<NNK>$_5$-GFP-SV40 PAS	NNKNNKNNKNNKNNKATGGTGAGCAAGGGCGAGGAGCTGT	CGGCCTATTGGTTAAAAAATGAGCT
<NNK>$_0$-GFP-SV40 PAS	ATGGTGAGCAAGGGCGAGGAGCTGT	CGGCCTATTGGTTAAAAAATGAGCT
Amplicon for PE-sequencing library construction	[Barcode]TCAGATCCGCTAGCGCTACC	[Barcode]GAACTTCAGGGTCAGCTTGC

N = T or C or A or G; K = T or G; M = C or A.

Supplementary Table 2. Percentage of each nucleotide at each position of peptide library codons

Codon position	Nucleotide	5X+5X	10X+5X	15X+5X	5X+0X	10X+0X	15X+0X
	T	23.69 ± 0.30	24.31 ± 0.45	22.86 ± 0.35	17.96 ± 0.45	21.21 ± 0.21	19.67 ± 0.36
	C	26.43 ± 0.30	28.23 ± 0.59	30.39 ± 0.29	25.64 ± 0.34	29.42 ± 0.28	31.93 ± 0.41
	A	27.26 ± 0.23	24.58 ± 0.35	23.09 ± 0.13	35.64 ± 0.22	27.31 ± 0.27	24.84 ± 0.15
First	G	22.63 ± 0.18	22.89 ± 0.30	23.66 ± 0.24	20.75 ± 0.11	22.06 ± 0.24	23.55 ± 0.42
	T	24.20 ± 0.40	24.55 ± 0.67	22.60 ± 0.10	19.01 ± 0.02	21.82 ± 0.24	19.39 ± 0.35
	C	28.78 ± 0.19	29.43 ± 0.80	31.66 ± 0.35	26.53 ± 0.31	28.91 ± 0.45	32.50 ± 0.29
	A	25.23 ± 0.38	24.29 ± 0.39	22.19 ± 0.34	33.66 ± 0.07	27.87 ± 0.39	24.55 ± 0.50
Second	G	21.79 ± 0.29	21.74 ± 0.10	23.55 ± 0.09	20.80 ± 0.35	21.40 ± 0.33	23.55 ± 0.13
	T	52.50 ± 0.42	51.13 ± 0.24	47.80 ± 0.04	49.64 ± 0.74	48.56 ± 0.34	45.07 ± 0.44
	C	0.65 ± 0.04	0.66 ± 0.06	1.51 ± 0.03	0.48 ± 0.05	0.31 ± 0.03	0.83 ± 0.08
	A	0.86 ± 0.10	0.61 ± 0.04	1.03 ± 0.08	0.42 ± 0.03	0.36 ± 0.04	0.60 ± 0.05
Third	G	45.98 ± 0.31	47.60 ± 0.29	49.66 ± 0.12	49.46 ± 0.71	50.77 ± 0.33	53.50 ± 0.33

The total number of each nucleotide at each position of peptide library codons is summarized. The average percentage of different nucleotide at each codon base is presented. All data are displayed as the mean ± SD, n = 3.

Supplementary materials

ldsDNA sequences in study:

>CMV-Kozak-<NNK>$_5$

CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCATGCATTAGT
TATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCC
GCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCG
CCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTC
CATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATC
AAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCC
CGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTAC
ATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAA
TGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGAC
GTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTA
ACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCT
ATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGTCGCC
ACCATGGTGNNKNNKNNKNNKNNK

>CMV-Kozak-<NNK>$_{10}$

CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCATGCATTAGT
TATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCC
GCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCG
CCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTC
CATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATC
AAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCC
CGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTAC
ATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAA
TGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGAC
GTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTA
ACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCT
ATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGTCGCC
ACCATGGTGNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK

>CMV-Kozak-<NNK>$_{15}$

CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCATGCATTAGT
TATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCC
GCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCG
CCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTC
CATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATC
AAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCC
CGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTAC

ATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAA
TGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGAC
GTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTA
ACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCT
ATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGTCGCC
ACCATGGTGNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK
NK

>$<NNK>_5$-GFP-SV40 PAS

NNKNNKNNKNNKNNKATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGT
GCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCC
GGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGC
ACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACG
GCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAA
GTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGAC
GGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAAC
CGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCAC
AAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAG
AAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGC
GTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTG
CTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCA
ACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCA
CTCTCGGCATGGACGAGCTGTACAAGTCCGGACTCAGATCTCGAGCTCAAGCTTC
GAATTCTGCAGTCGACGGTACCGCGGGCCCGGGATCCACCGGATCTAGATAACTG
ATCATAATCAGCCATACCACATTTGTAGAGGTTTTACTTGCTTTAAAAAACCTCCC
ACACCTCCCCCTGAACCTGAAACATAAAATGAATGCAATTGTTGTTGTTAACTTG
TTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAA
ATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTA
TCTTAACGCGTAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTTT
GTTAAATCAGCTCATTTTTTAACCAATAGGCCG

>$<NNK>_0$-GFP-SV40 PAS

ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAG
CTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGC
GATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGC
CCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAG
CCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAA
GGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACC
CGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAG
GGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAAC
TACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAG

GTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGAC
CACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACC
ACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATC
ACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGA
GCTGTACAAGTCCGGACTCAGATCTCGAGCTCAAGCTTCGAATTCTGCAGTCGAC
GGTACCGCGGGCCCGGGATCCACCGGATCTAGATAACTGATCATAATCAGCCATA
CCACATTTGTAGAGGTTTTACTTGCTTTAAAAAACCTCCCACACCTCCCCCTGAAC
CTGAAACATAAAATGAATGCAATTGTTGTTGTTAACTTGTTTATTGCAGCTTATAA
TGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCA
CTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTAACGCGTAAATTG
TAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATTTTTGTTAAATCAGCTCATTT
TTTAACCAATAGGCCG