Generation of Highly Diverse Peptide Library by Linear-double-stranded DNA Based AND Logic Computation in Mammalian Cells

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

1. 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 300060, China;

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

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

Contribute equally.

* To whom correspondence should be addressed. Tel: +86 22 23340123; Fax: +86 22 23340123;

Email: shuaili@tmu.edu.cn & shuai.li2001@gmail.com.

Abstract

Our previous work established linear-double-stranded DNA (ldsDNA) as innovative biological parts for building low noise-signal ratio AND gate circuits in mammalian cells. Here by introducing NNK degenerate codon nucleotides to ldsDNA terminals, we developed a novel methodology for building highly diverse peptide library. Deep sequencing results revealed that our novel method could generate peptide library with both amino acid sequence and peptide length diversities. Here for the first time, our work established ldsDNA as biological parts for building highly diverse peptide library in mammalian cells, which showed great application potential in generating antigen-receptor (antibodies and T-cell receptors) repertoire and functional peptide libraries.

Keywords

AND gate, linear-double-stranded DNA (ldsDNA), peptide library, antigen-receptor, complementarity-determining region (CDR).

Introduction

In mammalian, highly diverse repertoires of antigen-receptors (antibodies and T-cell receptors) are generated by V(D)J recombination mechanism during the early stages of T and B cell maturation [1-3]. In V(D)J recombination process, variable (V), diversity (D, heavy chains only) and joining (J) gene fragments undergo nearly random recombination, making contributions to the diversity of antibodies and T cell receptors (TCRs) repertoires (Figure 1A). Furthermore, during VD or DJ junction formation, indels (nucleotides insertion or deletion) introduced by terminal deoxynucleotide transferase (TdT), Artemis nuclease, DNA polymerase or NHEJ pathway members also contribute to antigen-receptor diversity [4,5]. Diversity (D) gene fragments together with nucleotide sequences surrounding VD and DJ junction encode the hypervariable complementarity determining region 3 (CDR3), which is located in the antigen binding surface and determines antigen-receptor specificity (Figure 1B) [6]. Recently, high-throughput sequencing (HTS) technology was employed to explore the diversity of TCR CDR3 [7]. These sequencing results showed that the CDR3 peptide complexity came from the combination of amino acids sequence diversity with peptide chain length dynamics [8-11].

We previously applied linear-double-stranded DNA (ldsDNA, or PCR amplicon) as novel biological parts to implement Boolean logic AND gates in mammalian cells [12]. 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*. Moreover, we showed ldsDNA with one or two additional terminal nucleotide(s), which leads to reading frame shift, could undergo end-processing and produce in-frame AND-gate output proteins. These characteristics of ldsDNA-based AND gate calculation is similar to V(D)J recombination. In this study, we employed

terminal-NNKs-ldsDNA-based AND-gate calculation to generate highly diverse peptide library (Figure 1C).

Materials and Methods

Cell culture

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

ldsDNA synthesis

KOD DNA polymerase (Toyobo) was used to amplify ldsDNAs (PCR amplicons) taking pEGFP-C1 (Clontech) plasmid as template. To remove plasmid template and free dNTPs, PCR products underwent agarose electrophoresis and gel-purification (Universal DNA purification kit, Tiangen). The amount of PCR products was determined by OD 260 absorption value. PCR primer and ldsDNA sequences were present in 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 1000ng/well) were transfected into cells with Lipofectamine 2000 (Invitrogen) reagent following standard protocol.

RNA extraction, cDNA synthesis and PCR amplification

48 h after transfection, total RNAs were extracted by RNAiso Plus (Takara) reagent. cDNA was synthesized by PrimeScript ^{TM} RT reagent Kit with gDNA Eraser (Takara) following standard protocol.

KOD DNA polymerase (Toyobo) was used to amplify the sequences surrounding the ldsDNA junctions.

Library construction and high-throughput sequencing

Library construction and high-throughput sequencing were conduct by Novogene (China). Sequencing libraries were generated using TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina) following 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 library was sequenced on an Illumina HiSeq 2500 and 250 bp paired-end were generated.

Data analysis

Paired-end reads from the original DNA fragments are merged using FLASH [13], a very fast and accurate analysis tool which is designed to merge paired-end reads when there are overlaps between reads 1 and reads 2. Paired-end reads was assigned to each sample according to the unique barcodes. Reads were filtered by QIIME quality filters. Random nucleotide sequences generated by ldsDNA-based AND gate computation were extract by in-house Perl scripts.

Results

Rational of generation highly diverse peptide library by ldsDNA-based AND gate calculation

Our previous work showed 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 [12]. Here we hypothesized that ldsDNAs with terminal NNK

degenerate codons could generate highly diverse peptide libraries through AND-gate calculation (Figure 1C).

Taking pEGFP-C1 vector as PCR template, primers with multiple NNK-trinucleotides were used to generate ldsDNAs with terminal degenerate codons (Table 1). Three ldsDNAs with CMV promoter, Kozak sequence and different number of NNK-trinucleotide were produced (namely CMV-Kozak-$<NNK>_5$, CMV-Kozak-$<NNK>_{10}$, CMV-Kozak-$<NNK>_{15}$) (for sequences see Supplementary Materials). Meanwhile, two ldsDNAs containing terminal NNK-trinucleotide (or not), GFP coding sequence and SV40 polyA signal were amplified (namely $<NNK>_5$-GFP-SV40 PAS, $<NNK>_0$-GFP-SV40 PAS) (for sequences see Supplementary Materials). When AND-gate calculation was conducted in cells, random peptide library would be generated surrounding the junction formed between these two subtype ldsDNAs. Six different combinations of these two subgroup ldsDNAs were cotransfected into HEK293T cells, respectively. These six combinations were as followings:

1, CMV-Kozak-$<NNK>_5$ with $<NNK>_5$-GFP-SV40 PAS (15N-15N for short);

2, CMV-Kozak-$<NNK>_{10}$ with $<NNK>_5$-GFP-SV40 PAS (30N-15N for short);

3, CMV-Kozak-$<NNK>_{15}$ with $<NNK>_5$-GFP-SV40 PAS (45N-15N for short);

4, CMV-Kozak-$<NNK>_5$ with $<NNK>_0$-GFP-SV40 PAS (15N-0N for short);

5, CMV-Kozak-$<NNK>_{10}$ with $<NNK>_0$-GFP-SV40 PAS (30N-0N for short);

6, CMV-Kozak-$<NNK>_{15}$ with $<NNK>_0$-GFP-SV40 PAS (45N-0N for short).

48h after transfection, total RNA was extracted and cDNAs were synthesized via reverse transcription. Nucleotide sequences surrounding the junctions between two ldsDNA-inputs were PCR amplified and were then subjected to pair-end high-throughput sequencing.

Characteristics of peptide library generated by ldsDNA-based AND gate calculation

Reading frame phase characteristics of terminal-NNKs-ldsDNAs-based AND-gate computation generated nucleotide sequences were present in Figure 2. The in-frame nucleotide sequences (3n) were translated into protein peptides for further analysis. The length distributions of 6 different peptide libraries were showed by Figure 3 (peptides with pre-stop codon were excluded). The amino acid composition of each peptide libraries was present in Figure 4. The theoretical amino acid composition of NNK decency codon was also present (Figure 4). Meanwhile, nucleotide composition feature in each codon position was showed in Figure 5. The distribution of nucleotides was in line with NNK degenerate bases.

One of the key characteristics of a peptide library is its coding capacity. All 6 different peptide libraries cover all 400 possibilities of dual-amino-acids combinations (2AA, Figure 6). For triple-amino-acids (3AA), 15N+15N peptide library covers 91.35% of all possibilities and other 5 libraries cover over 99% of all possibilities (100% for 45N+0N group). However, for quadruple-amino-acids (4AA), 6 libraries showed more different coverage capacities (Figure 6). When it comes to quintuple-amino-acids (5AA), all 6 libraries showed very low coverage levels (< 10%).

Discussion

Here by AND-gate computation of terminal-NNKs-ldsDNAs, we developed a novel method to generate highly diverse peptide library in mammalian cells. This peptide library is characterized by both amino acids sequence diversity and peptide length dynamics, which is similar to hypervariable CDR3. This study provides novel methodology for the generation of highly diverse repertoires of antigen-receptors and functional peptide libraries.

In mammalian, both antibodies and TCRs composed of two different protein chains, heavy chain and light chain. The assortment of heavy and light chains also contributes to antigen-receptor diversity. However, unless employing single cell sequencing methods, it's hard to connect heavy chains to their paired light chains [14]. Single-domain antibodies, referred to as nanobodies, are usually derived from camelids, which make a unique subset of antibodies consisting of heavy-chain homodimers devoid of light chains [15,16]. Nanobodies are extremely stable, smaller in size (about 15 kDa) and could bind antigens with affinities in the nanomolar range [17-20]. Using our ldsDNA-based AND gate strategy, highly diverse nanobody library could be generate without considering the pairing between heavy and light chains. The complexity of this library comes from CDR3 (from ldsDNA-based AND gate calculation) and multiple V gene segments (contribute CDR1 and CDR2) (Figure 1B).

Chimeric antigen receptor (CAR) T cell therapy showed great promise in cancer treatment [21-23]. To date, two CD-19 targeted CAR-T therapies have been approved by FDA for B-cell malignancies. The success of CD-19 targeted CAR-T therapy may come from the following aspects [24]. First, CD19 is specifically expressed on B-cells, this restricts the off-target effects of CAR-T. Second, patients could tolerate the elimination of normal mature and precursor B-cells caused by CD-19 targeted CAR-T therapy. Third, the immune-resistance level is low during the therapy. However, less exciting conclusions can be derived from clinical trials designed for the treatment of solid tumors [25]. Nearly all

these clinical trials were based on targeting sole tumor-antigen. This strategy may lead to poor specificity (raising off-targeting effects) and poor efficacy (caused by tumor heterogeneity). Our novel method, for the first time, enable us to build highly diverse artificial antigen-receptor repertoires in mammalian cells. Using these mammalian cells, we can profile tumor-specific-antigens. For example, T-cells with these artificial repertoires could be introduced into tumor patients and then tumor-antigen-binding antigen-receptors could be explored by tumor sampling and deep sequencing (Figure 7). These antigen-receptors may be used to generate multiple CARs and synergistically applied for solid tumor therapy.

The ldsDNAs generated by this AND gate computation could not be amplified in mammalian cells. Although little portion of these ldsDNA could integrate into chromosomes, most of ldsDNA will be lost during cell proliferation. Recently, several strategies have been developed to promote chromosomal integration of ldsDNAs [26-28]. After integration into the genome, the ldsDNAs could be remembered by cells leading to persistent protein expression.

In summary, we applied ldsDNA-based AND gate calculation to generate highly diverse peptide library, which may have great application potential in building highly diverse antigen-receptor repertoire and functional peptide libraries.

Acknowledgements

This work was supported by National Natural Science Foundation of China [31400673 to S.L., 81402407 to W.S.]; Tianjin Research Program of Application Foundation and Advanced Technology [14JCQNJC09800 to S.L., 15JCQNJC11700 to W.S.]; and 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 experiments. S.L. and W.S. analyzed data. W.S. and S.L. prepared 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. (A) Schematic showing V(D)J recombination process. Individual variable (V), diversity (D) and joining (J) gene segments are randomly selected and joined together. (B) The genesis of antigen-receptor heavy chain complementarity-determining regions (CDRs). CDR1 and CDR2 are encoded by V gene segments. CDR3 is encoded by D gene fragments together with nucleotide sequences surrounding VD and DJ junction. (C) Schematic showing the rational to generate highly diverse peptide library by terminal-NNKs-ldsDNAs-based AND-gate computation.

Figure 2. In-frame (3n) and out-of-frame (3n+1 or 3n+2) characteristics of nucleotide sequences of 6 different libraries. Clean tag number of each libraries was present.

Figure 3. The peptide length distributions of 6 different peptide libraries generated by terminal-NNKs-ldsDNA-based AND-gate calculation.

Figure 4. The amino acid composition of 6 different peptide libraries generated by terminal-NNKs-ldsDNAs-based AND-gate computation (Blue). Theoretical amino acid composition of NNK degenerate codon was also present (Red).

Figure 5. Nucleotide composition feature in each codon position of 6 different peptide libraries.

Figure 6. Peptide libraries coding capacity on covering dual-(2AA), triple-(3AA), quadruple-(4AA) and quintuple-amino acids (5AA) peptide possibilities.

Figure 7. Rational of tumor-specific antigen profiling by T-cell with ldsDNA generated highly diverse antigen-receptor library. T-cells from cancer patients are introduced with terminal-NNKs-ldsDNAs, thereby generating highly diverse antigen-receptor library by AND-gate computation. Then T-cell would be infused into human bodies. Tumor sampling and high-throughput sequencing would be employed to explore tumor-antigen. Antigen-receptors binding to tumor-antigen would be used to develop multiple tumor-antigen-targeting CARs for tumor therapy.

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	CACCATGGTGGCGACCGGTAGCGCT	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 materials

ldsDNA sequences in study:

>CMV-Kozak-<NNK>$_5$

CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCATGCATTAGTTATTA
ATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATA
ACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAAT
AATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGA
GTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCC
CCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTT
ATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATG
CGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGT
CTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCA
AAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGA
GGTCTATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGTCGCC
ACCATGGTGNNKNNKNNKNNKNNK

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

CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCATGCATTAGTTATTA
ATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATA
ACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAAT
AATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGA
GTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCC
CCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTT
ATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATG
CGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGT
CTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCA
AAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGA
GGTCTATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGTCGCC
ACCATGGTGNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK

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

CTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCATGCATTAGTTATTA
ATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATA
ACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAAT
AATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGA
GTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCC
CCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTT
ATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATG
CGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGT
CTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCA

AAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGA
GGTCTATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGTCGCC
ACCATGGTGNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK

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

NNKNNKNNKNNKNNKATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCAT
CCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCG
AGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGC
CCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCT
ACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCC
AGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAG
TTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGA
CGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCAT
GGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGG
ACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCC
GTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAAC
GAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGC
ATGGACGAGCTGTACAAGTCCGGACTCAGATCTCGAGCTCAAGCTTCGAATTCTGCAGTC
GACGGTACCGCGGGCCCGGGATCCACCGGATCTAGATAACTGATCATAATCAGCCATACC
ACATTTGTAGAGGTTTTACTTGCTTTAAAAAACCTCCCACACCTCCCCCTGAACCTGAAAC
ATAAAATGAATGCAATTGTTGTTGTTAACTTGTTTATTGCAGCTTATAATGGTTACAAATA
AAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGT
TTGTCCAAACTCATCAATGTATCTTAACGCGTAAATTGTAAGCGTTAATATTTTGTTAAAA
TTCGCGTTAAATTTTTGTTAAATCAGCTCATTTTTTAACCAATAGGCCG

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

ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGA
CGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCT
ACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCA
CCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGA
AGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCT
TCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACC
CTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGG
GCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGA
AGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAG
CTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGAC
AACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCA
CATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTA
CAAGTCCGGACTCAGATCTCGAGCTCAAGCTTCGAATTCTGCAGTCGACGGTACCGCGGG
CCCGGGATCCACCGGATCTAGATAACTGATCATAATCAGCCATACCACATTTGTAGAGGT
TTTACTTGCTTTAAAAAACCTCCCACACCTCCCCCTGAACCTGAAACATAAAATGAATGC

AATTGTTGTTGTTAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATC
ACAAATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCA
TCAATGTATCTTAACGCGTAAATTGTAAGCGTTAATATTTTGTTAAAATTCGCGTTAAATT
TTTGTTAAATCAGCTCATTTTTTAACCAATAGGCCG

References

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

2 Schatz, D. G. V(D)J recombination. *Immunological reviews* **200**, 5-11, doi:10.1111/j.0105-2896.2004.00173.x (2004).

3 Jung, D. & Alt, F. W. Unraveling V(D)J recombination; insights into gene regulation. *Cell* **116**, 299-311 (2004).

4 Komori, T., Okada, A., Stewart, V. & Alt, F. W. Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. *Science* **261**, 1171-1175 (1993).

5 Ma, Y., Pannicke, U., Schwarz, K. & Lieber, M. R. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell* **108**, 781-794 (2002).

6 Xu, J. L. & Davis, M. M. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37-45 (2000).

7 Matos, T. R., de Rie, M. A. & Teunissen, M. B. M. Research Techniques Made Simple: High-Throughput Sequencing of the T-Cell Receptor. *The Journal of investigative dermatology* **137**, e131-e138, doi:10.1016/j.jid.2017.04.001 (2017).

8 Ruggiero, E. *et al.* High-resolution analysis of the human T-cell receptor repertoire. *Nature communications* **6**, 8081, doi:10.1038/ncomms9081 (2015).

9 Clemente, M. J. *et al.* Deep sequencing of the T-cell receptor repertoire in CD8+ T-large granular lymphocyte leukemia identifies signature landscapes. *Blood* **122**, 4077-4085, doi:10.1182/blood-2013-05-506386 (2013).

10 Kitaura, K., Shini, T., Matsutani, T. & Suzuki, R. A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) alpha and beta repertoires and identifying potential new invariant TCR alpha chains. *BMC immunology* **17**, 38, doi:10.1186/s12865-016-0177-5 (2016).

11 Krell, P. F. *et al.* Next-generation-sequencing-spectratyping reveals public T-cell receptor repertoires in pediatric very severe aplastic anemia and identifies a beta chain CDR3 sequence associated with hepatitis-induced pathogenesis. *Haematologica* **98**, 1388-1396, doi:10.3324/haematol.2012.069708 (2013).

12 Su, W., Zhang, C. & Li, S. Linear-double-stranded DNA (ldsDNA) based AND logic computation in mammalian cells. *bioRxiv*, doi:10.1101/266056 (2018).

13 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

14 Friedensohn, S., Khan, T. A. & Reddy, S. T. Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends in biotechnology* **35**, 203-214, doi:10.1016/j.tibtech.2016.09.010 (2017).

15 Hamers-Casterman, C. *et al.* Naturally occurring antibodies devoid of light chains. *Nature* **363**, 446-448, doi:10.1038/363446a0 (1993).

16 Muyldermans, S. Nanobodies: natural single-domain antibodies. *Annual review of biochemistry* **82**, 775-797, doi:10.1146/annurev-biochem-063011-092449 (2013).

17 Dumoulin, M. *et al.* Single-domain antibody fragments with high conformational stability. *Protein science : a publication of the Protein Society* **11**, 500-515, doi:10.1110/ps.34602 (2002).

18 Rothbauer, U. *et al.* Targeting and tracing antigens in live cells with fluorescent nanobodies. *Nature methods* **3**, 887-889, doi:10.1038/nmeth953 (2006).

19 Worn, A. & Pluckthun, A. Stability engineering of antibody single-chain Fv fragments. *Journal of molecular biology* **305**, 989-1010, doi:10.1006/jmbi.2000.4265 (2001).

20 Fridy, P. C. *et al.* A robust pipeline for rapid production of versatile nanobody repertoires. *Nature methods* **11**, 1253-1260, doi:10.1038/nmeth.3170 (2014).

21 Srivastava, S. & Riddell, S. R. Engineering CAR-T cells: Design concepts. *Trends in immunology* **36**, 494-502, doi:10.1016/j.it.2015.06.004 (2015).

22 Davila, M. L. *et al.* Efficacy and toxicity management of 19-28z CAR T cell therapy in B cell acute lymphoblastic leukemia. *Science translational medicine* **6**, 224ra225, doi:10.1126/scitranslmed.3008226 (2014).

23 Brentjens, R. J. *et al.* CD19-targeted T cells rapidly induce molecular remissions in adults with chemotherapy-refractory acute lymphoblastic leukemia. *Science translational medicine* **5**, 177ra138, doi:10.1126/scitranslmed.3005930 (2013).

24 Maude, S. L., Teachey, D. T., Porter, D. L. & Grupp, S. A. CD19-targeted chimeric antigen receptor T-cell therapy for acute lymphoblastic leukemia. *Blood* **125**, 4017-4023, doi:10.1182/blood-2014-12-580068 (2015).

25 D'Aloia, M. M., Zizzari, I. G., Sacchetti, B., Pierelli, L. & Alimandi, M. CAR-T cells: the long and winding road to solid tumors. *Cell death & disease* **9**, 282, doi:10.1038/s41419-018-0278-6 (2018).

26 Nakai, H. *et al.* Helper-independent and AAV-ITR-independent chromosomal integration of double-stranded linear DNA vectors in mice. *Molecular therapy : the journal of the American Society of Gene Therapy* **7**, 101-111 (2003).

27 Hindmarsh, P. & Leis, J. Retroviral DNA integration. *Microbiology and molecular biology reviews : MMBR* **63**, 836-843, table of contents (1999).

28 Wu, S. C. *et al.* piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 15008-15013, doi:10.1073/pnas.0606979103 (2006).

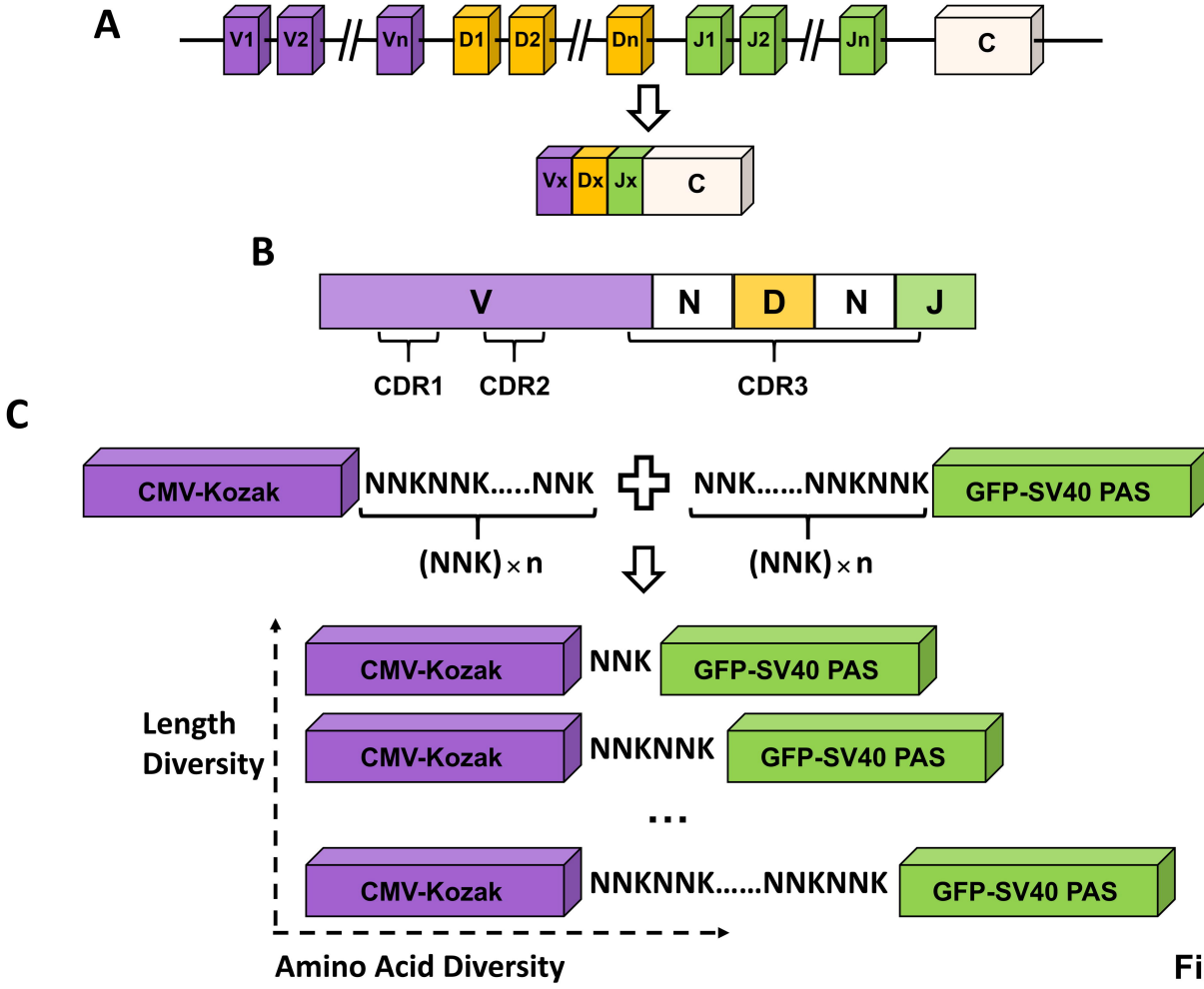
A
V1
V2
Vn
D1
D2
Dn
J1
J2
Jn
C
Vx
Dx
Jx
C
B
V
N
D
N
J
CDR1
CDR2
CDR3
C
CMV-Kozak
NNKNNK.....NNK
NNK.....NNKNNK
GFP-SV40 PAS
(NNK)×n
(NNK)×n
Length Diversity
CMV-Kozak
NNK
GFP-SV40 PAS
CMV-Kozak
NNKNNK
GFP-SV40 PAS
...
CMV-Kozak
NNKNNK.....NNKNNK
GFP-SV40 PAS
Amino Acid Diversity

Figure 1

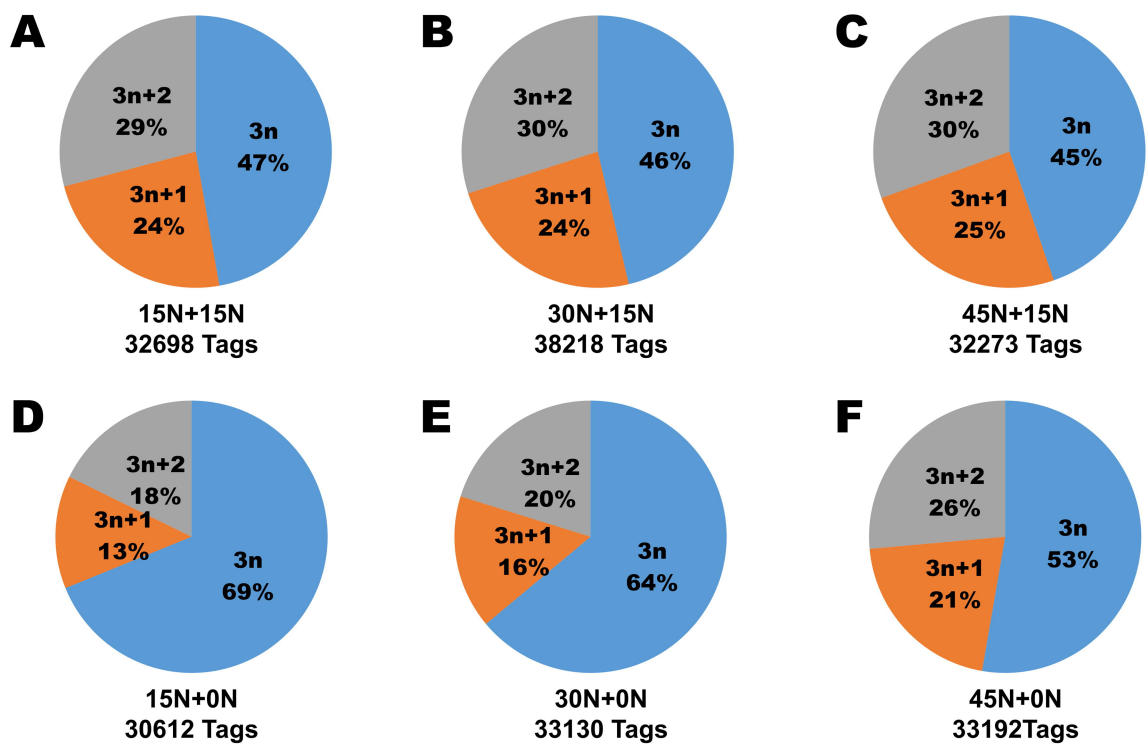
A
3n+2
29%
3n
47%
3n+1
24%
15N+15N
32698 Tags
B
3n+2
30%
3n
46%
3n+1
24%
30N+15N
38218 Tags
C
3n+2
30%
3n
45%
3n+1
25%
45N+15N
32273 Tags
D
3n+2
18%
3n+1
13%
3n
69%
15N+0N
30612 Tags
E
3n+2
20%
3n+1
16%
3n
64%
30N+0N
33130 Tags
F
3n+2
26%
3n
53%
3n+1
21%
45N+0N
33192Tags

Figure 2

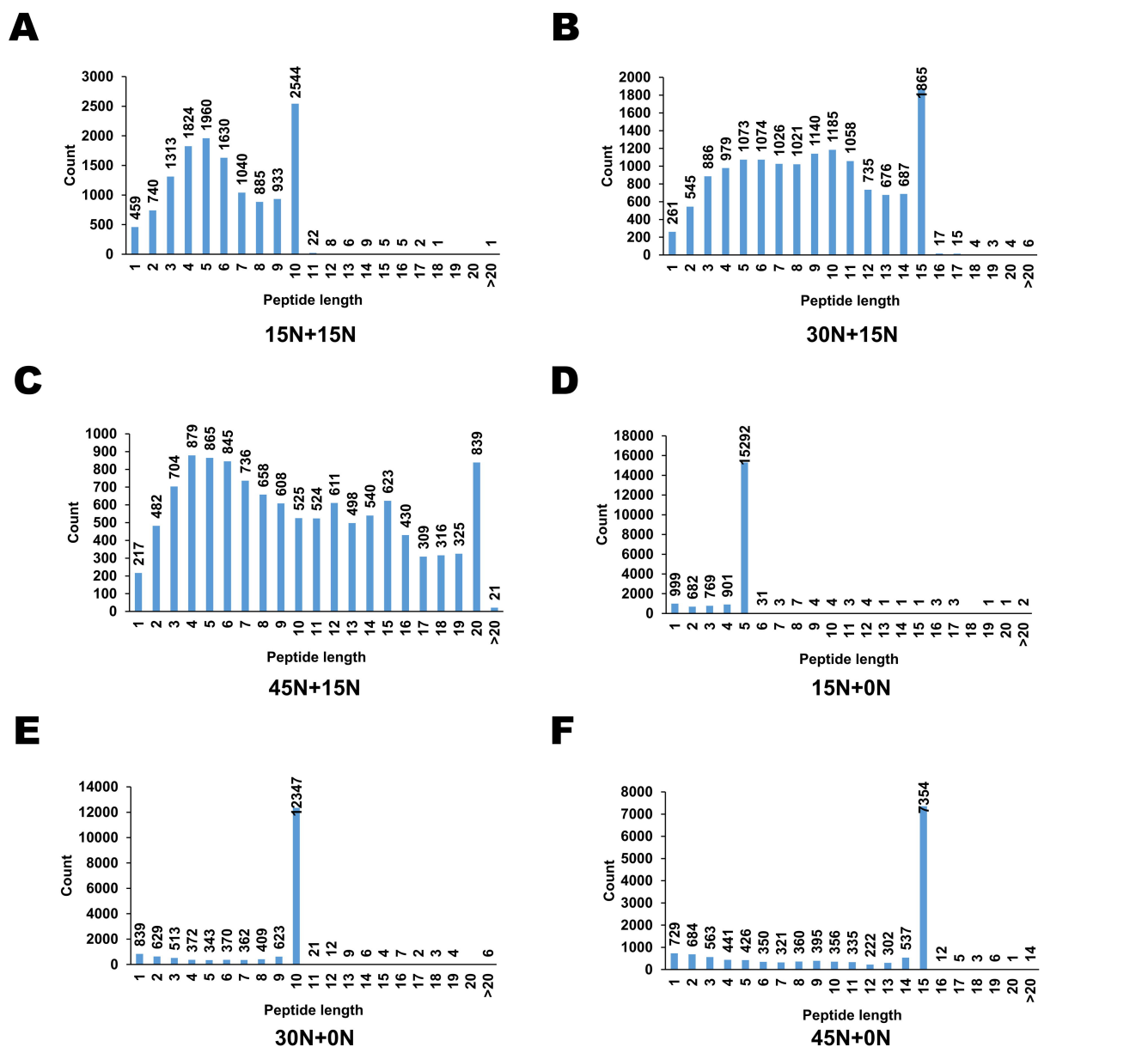
A
Count
Peptide length
15N+15N
B
Count
Peptide length
30N+15N
C
Count
Peptide length
45N+15N
D
Count
Peptide length
15N+0N
E
Count
Peptide length
30N+0N
F
Count
Peptide length
45N+0N

Figure 3

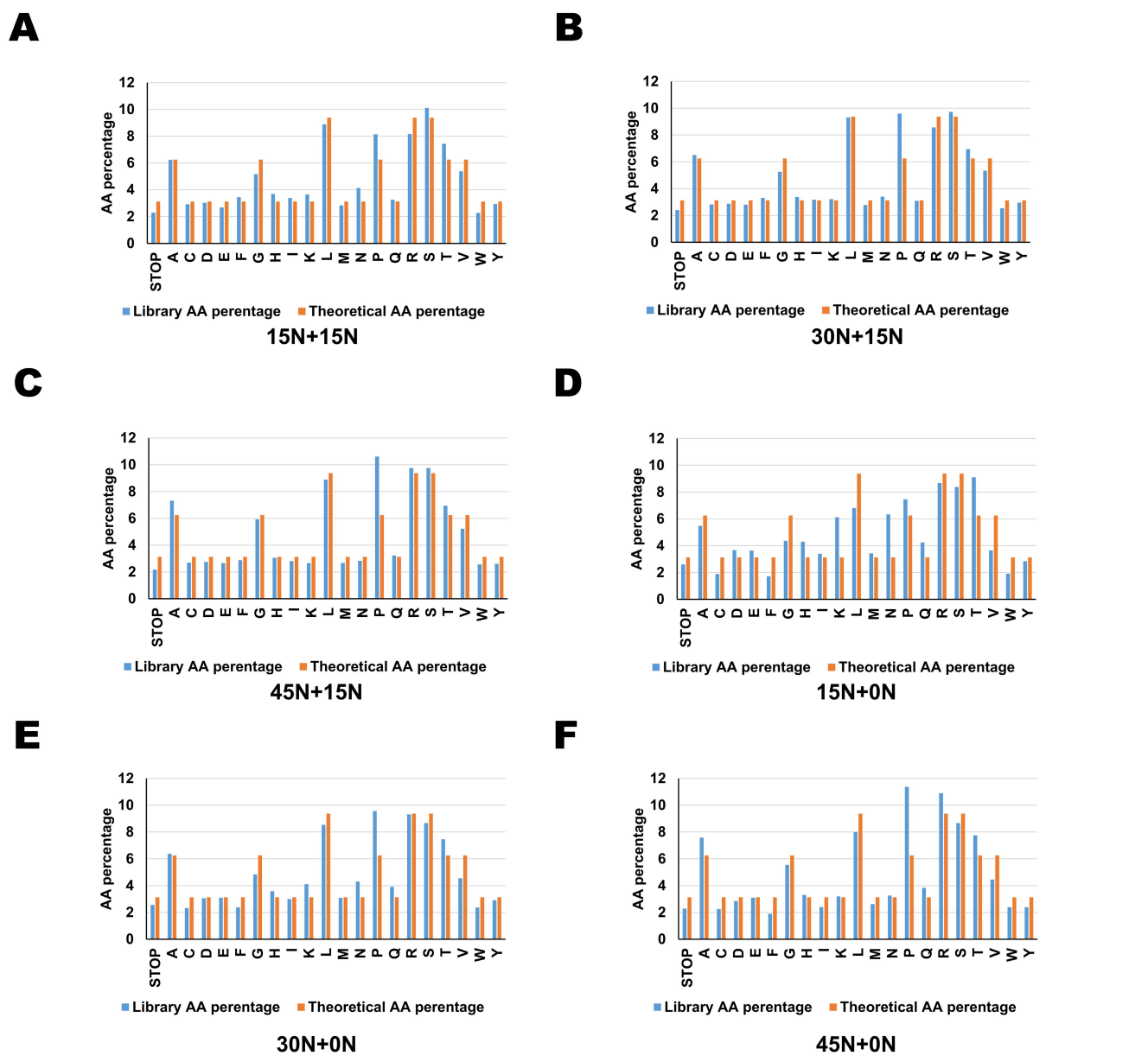
A
B
C
D
E
F
AA percentage
0
2
4
6
8
10
12
STOP A C D E F G H I K L M N P Q R S T V W Y
Library AA perentage
Theoretical AA perentage
15N+15N
30N+15N
45N+15N
15N+0N
30N+0N
45N+0N

Figure 4

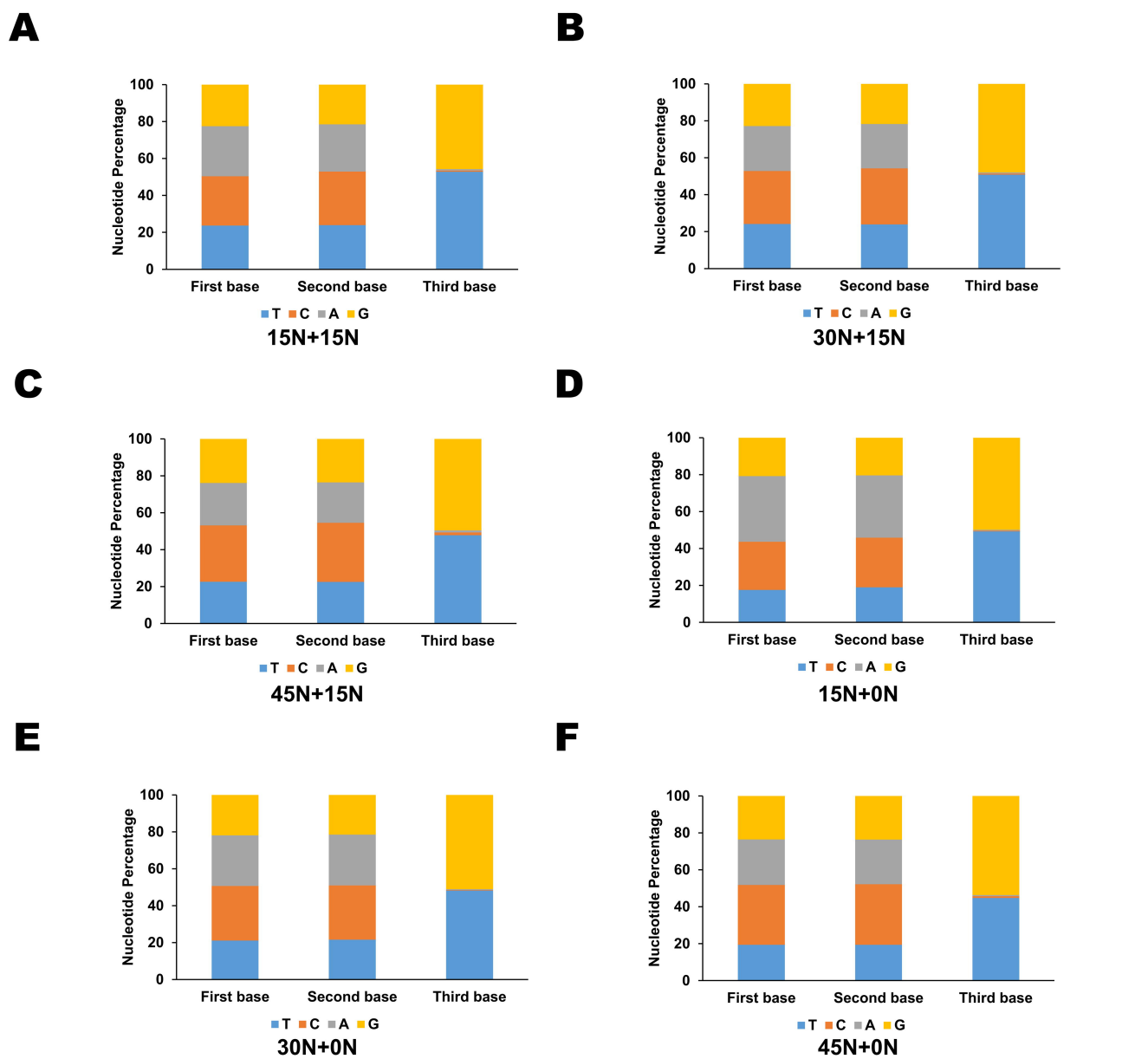
A
B
C
D
E
F
Nucleotide Percentage
100
80
60
40
20
0
First base
Second base
Third base
T C A G
15N+15N
30N+15N
45N+15N
15N+0N
30N+0N
45N+0N

Figure 5

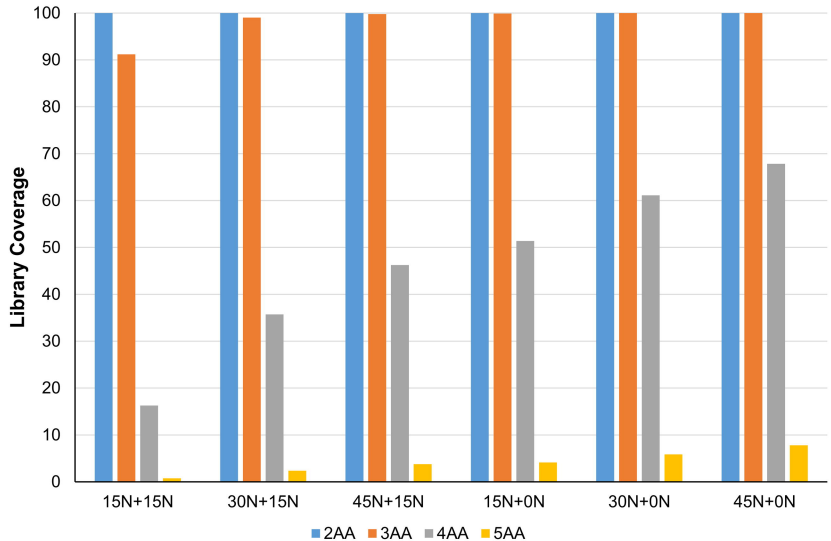
Library Coverage
100
90
80
70
60
50
40
30
20
10
0
15N+15N
30N+15N
45N+15N
15N+0N
30N+0N
45N+0N
2AA
3AA
4AA
5AA

Figure 6

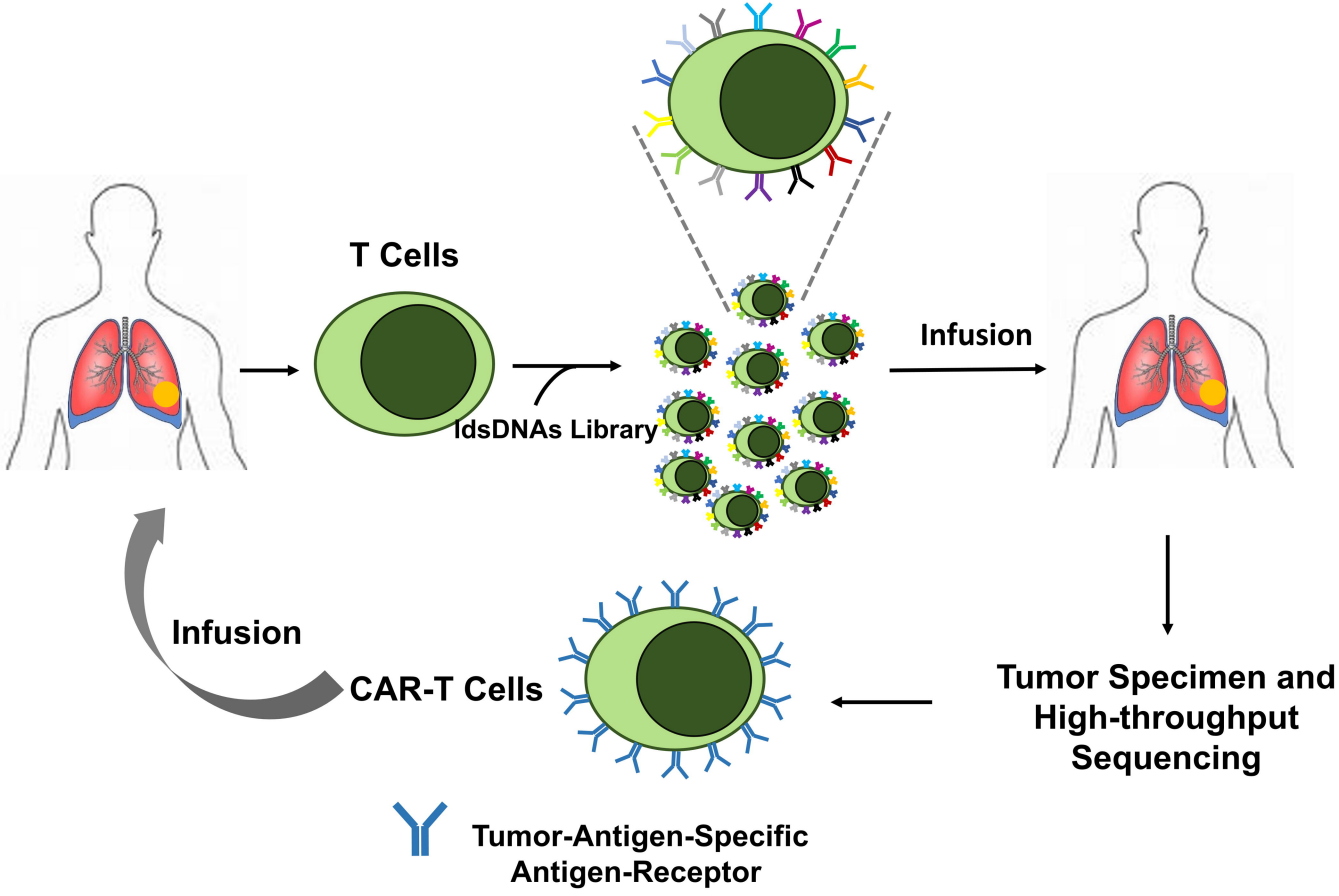
T Cells
IdsDNAs Library
Infusion
Infusion
CAR-T Cells
Tumor Specimen and
High-throughput
Sequencing
Tumor-Antigen-Specific
Antigen-Receptor

Figure 7