piggyBac-system-mediated genomic integration of large variant libraries for deep mutational scanning in mammalian cells

Yanjie Zhao [1#], Yifan Li [1#], Yi Wang [1], Weijun Su [2], Shuai Li [1*]

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; Tianjin's Clinical Research Center for Cancer, Tianjin 300060, China.

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

Contribute equally.

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

Abstract

There is a need of introducing large variant libraries into mammalian cells to facilitate deep mutational scanning (DMS). In this study, we develop an approach to integrate *in-vitro* ligation dsDNA variant library into cultured mammalian cell genome via the piggyBac transposon system. T4 DNA ligase was employed to produce dsDNA ligation products containing variant library sequence flanked by piggyBac inverted terminal repeat sequences (ITRs) at both ends. The ligation products were then cotransfected with a piggyBac transposase expression vector into mammalian cells for library integration. The clone formation assay results showed that about 29,000 resistance clones were generated from a single well of 6-well-plate transfected HEK293T cells. As a proof of principle of using the strategy for DMS, a GFP chromophore random variant library was integrated into the HEK293T cell genome. By FACS selection, we enriched GFP-positive cells. The next-generation sequencing (NGS) results showed that the GFP-chromophore amino acid sequences were restored after selection.

Keywords

Deep mutational scanning, T4 DNA ligase, piggyBac system, double-stranded DNA

Introduction

Deep mutational scanning (DMS), also known as saturated mutagenesis screen or massively parallel genetic assay, is a powerful high-throughput screening method in protein engineering that uncovers genotype-phenotype relationships (1,2). DMS has a wide range of applications, including antibody engineering (3,4), vaccine design (5-7), drug-resistant mutation identification (8), and so on. In a typical DMS experiment, thousands of mutations are introduced into the specific genomic region and then subjected to selection. After screening, next-generation sequencing (NGS) is employed to score each mutation, thereby connecting sequence with function (1,2).

Mammalian cells are the ideal platform for phenotypic screening of many types of proteins such as transmembrane proteins (9,10), proteins associated with signaling pathways (11,12), and proteins related to cellular phenotypes (7,8,13). However, mammalian DMS experiments are substantially challenged by difficulties in constructing large genome-integrated variant libraries. Currently, there are two types of approaches to introduce integrated libraries into mammalian cells, both of which have their limitations. The first type of approach is based on plasmids. The plasmid-constructs are transformed into competent cells to achieve a maximum library size of 10^9 - 10^{10} (14). Then after plasmid amplification, extraction, and cell-transfection, the plasmid-library is integrated into the mammalian genome via viral- (15), transposon- (16,17), or recombinase- (11) mediated genome integration. The library content is continuously lost in the above steps, resulting in a significant reduction of the library size (~10^4). The second type of approach is CRISPR/Cas9-mediated homology-directed repair (HDR) (3,8,13). CRISPR/Cas9 system is

employed to create double-strand breaks (DSBs) at specific genomic sites to trigger DNA damage repair mechanisms. Meanwhile, ssODNs, containing both homology arms and variant coding sequences, are transfected into cells as HDR templates, thereby incorporating the library into the DSB site. However, there are several technical hurdles in HDR, e.g., need to pre-integrate the landing pad of the library, need to optimize the CRISPR targeting efficiency, hard to distinguish integrated and non-integrated cells, and limited mutation range (<100 nt). Therefore, there is a need of developing an easy to perform, genomic-integrated, large variant library construction strategy to facilitate mammalian cell-based DMS.

In our previous study, we developed a mammalian cell-based nanobody library construction strategy named "*in-vitro* ligation" (18). The random complementary determining region (CDR) sequences are first incorporated into double-stranded DNAs (dsDNAs) through PCR amplification. Then the CDRs containing up- and down-stream dsDNAs are ligated together by T4 DNA ligase to form a full dsDNA nanobody expression cassette containing the promoter, nanobody gene body, and polyA signal sequence. The full dsDNAs are transfected into HEK293T cell for transient nanobody library expression. The NGS results show that our strategy could generate over 10^6 unique nanobody sequences. Based on this, we here develop a novel large variant library genomic integration strategy by combining *in-vitro* ligation strategy with piggyBac transposon system. To validate the feasibility of the strategy in mammalian DMS, we enriched GFP-positive cells and restored GFP-chromophore sequences from GFP-chromophore random libraries.

Materials and Methods

Cell culture

HEK293T cell line was purchased from ATCC (American Type Culture Collection) and maintained in DMEM (Dulbecco's Modified Eagle Medium, Corning, NY, USA) supplemented with 10% FBS (Fetal bovine serum, Corning) and 1% Penicillin-Streptomycin solution (10,000 U/ml, Thermo Fisher Scientific, Waltham, MA, USA) at 37 □ with 5% CO_2.

PCR amplification, restriction digestion and ligation

PCR template plasmids were constructed by VectorBuilder (VectorBuilder, Guangzhou, China, for plasmid sequences see supplementary materials). KOD Plus Neo DNA polymerase (Toyobo, Osaka, Japan) was used to perform PCR amplification. PCR products were purified (Universal DNA purification kit, TIANGEN, Beijing, China) after agarose electrophoresis, and the concentrations were determined by NanoDrop (Thermo Fisher Scientific). PCR primers were listed in Supplementary Table 1. The up- and down-stream PCR products were digested by BssSI (New England Biolabs, Ipswich, MA, USA). After agarose electrophoresis and gel purification, the up- and down-stream digested dsDNA were ligated together by T4 DNA ligase (New England Biolabs). The ligation products were electrophorized and purified for downstream transfection. To avoid UV-induced DNA damage, Blue Light Gel Imager (Sangon Biotech, Shanghai, China) was used for dsDNA visualization.

Cell transfection, resistance selection, and clone-formation assay

HEK293T cells were plated in 6-well plates the day before transfection (70~80 % confluency). *In vitro* ligation dsDNA, PCR products, or plasmid were cotransfected with a hyperactive piggyBac

transposase expression vector (VectorBuilder) into HEK293T cells by using Lipofectamine 2000 (Thermo Fisher Scientific). Forty-eight hours after transfection, puromycin (10 μg/ml, Selleck Chemicals, Houston, TX, USA) was added into the cell culture medium to select integrated cells. For clone-formation assay, transfected cells were diluted 1,000 times and cultured in 6-well plates for ten days. Crystal violet staining was used to visualize the cell clones.

Fluorescence-activated cell sorting (FACS)

HEK293T cells were digested with trypsin. Cells were resuspended in PBS, and GFP or BFP positive cells were isolated using a BD FACSAria™ II Cell Sorter (BD Biosciences, Bedford, MA, USA). FlowJo software (FLOWJO, Ashland, OR, USA) was used to analyze the raw data.

Next-generation sequencing and data processing

Genomic DNA was extracted with a TaKaRa MiniBEST Universal Genomic DNA Extraction Kit Ver. 5.0 (TaKaRa, Kusatsu, Japan). KOD Plus Neo DNA polymerase was used to amplify the sequences containing chromophore sequence. PCR products were subjected to sequencing library construction using the TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, San Diego, CA, USA). The libraries were loaded on a NovaSeq 6000 system (Illumina) to generate PE-150 reads. PE reads were assigned to samples based on barcodes. Clean reads were merged using the FLASH (Fast Length Adjustment of SHort reads) software (v1.2.11). Inhouse Perl scripts were developed to analyze the amino acid sequences of libraries.

Results

Genomic integration of *in-vitro* ligation dsDNA via piggyBac system

In our previous research, we developed an *in-vitro* ligation strategy to introduce nanobody library into the mammalian cells (18). Briefly, both up- and down-stream dsDNAs were first generated by PCR amplification then digested by restriction enzyme BssSI to create sticky ends. T4 DNA ligase is used to link up- and down-stream dsDNAs together to form full-length dsDNA with the intact gene expression cassette. The obtained full-length dsDNAs were transfected into HEK293T cells to express the nanobody library. In the present study, we employ the piggyBac system to directly integrate *in-vitro* ligation products into the mammalian genome for DMS.

The piggyBac system could efficiently insert foreign DNA into the mammalian genome via a non-viral route (19). Here, we add piggyBac recognition sequences, inverted terminal repeat sequences (ITRs), to the ends of the up- and down-stream dsDNA products, respectively (Figure 1A). T4 DNA ligase was employed to link BssSI digested up- and down-stream dsDNAs together to form a full puromycin resistance gene expression cassette (Figure 1A). Then the *in-vitro* ligation products and a hyperactive piggyBac transposase expression vector (20) were cotransfected into HEK293T cells. Forty-eight hours after transfection, the cells were diluted 1,000 times and cultured with puromycin for ten days. The cells were stained with crystal violet for clone number counting (Figure 1B). As shown by Figure 1C, *in-vitro* ligation products could generate about 29,000 resistant clones in a single well of 6-well-plate transfected HEK293T cells. While, PCR product and plasmid, both containing puromycin resistance gene expression cassette and ITR sequences, could lead to 1.98×10^5 and 2.03×10^5 resistant clones, respectively (Figure 1C).

Screening of EGFP chromophore library

We next employed the piggyBac system to integrate the *in-vitro* ligation dsDNA variant library into the mammalian genome for downstream functional DMS. We randomized the EGFP (enhanced GFP) chromophore amino acids codons by NNK degeneracy and expected to restore the EGFP coding sequence by FACS. The full-length dsDNAs containing two NNK codons (corresponding to EGFP T65-Y66) or three NNK codons (corresponding to EGFP T65-Y66-G67) were produced by *in-vitro* ligation, respectively (Figure 2A). Then the ligation dsDNAs and piggyBac transposase expression vector were cotransfected into HEK293T cells. For each library, the transfection was performed in a single well of 6-well-plate. Forty-eight hours after transfection, puromycin was added into the cell culture medium to select for ten days. Then resistance cells were subjected to FACS isolation. GFP positive cells were isolated and cultured for the subsequent round of FACS isolation. For the 2-NNK group, the initial GFP-positive rate was 8.47% (Figure 2B). After the first round of FACS, the GFP rate was raised to 80.2%. After the second round of sorting, the GFP positive rate was further raised to 91.4%. While for the 3-NNK group, the initial GFP positive rate was 0.319% (Figure 2C). The GFP positive rate was raised to 94.3% after two rounds of FACS.

We also isolated BFP positive cells from the 2-NNK library. The initial BFP-positive rate was 0.639% (Figure 3). After the first round of FACS, the BFP rate was raised to 67.4%. After the second round of selection, the BFP positive rate was further raised to 90.8%.

Assessing chromophore sequences by NGS

The chromophore sequences of sorted GFP-positive cells were then analyzed by NGS. For the 2-NNK group, the EGFP chromophore sequence TY ranks first (Table 1) (21). The wild-type GFP chromophore SY ranks second. The third LY has also been reported as a GFP chromophore (22,23). And seven out of the top ten amino acid sequences show a conserved “XY” pattern. For the 3-NNK group, the EGFP chromophore sequence TYG ranks first (Table 1). The second GYG (24) and the third CYG (23) were also reported as GFP chromophores. Five out of the top ten amino acid sequences show a conserved “XYG” pattern. The chromophore sequences of sorted BFP-positive cells were also analyzed. As shown by Table 2, the BFP chromophore sequence SH ranks first.

Discussion

To facilitate DMS in cultured mammalian cells, we developed a novel large variant library genomic integration approach by combining the *in-vitro* ligation library construction strategy with the piggyBac system. The method is easy to implement, only requires PCR, restriction digestion, ligation, and cell transfection. The feasibility of the method in mammalian DMS was confirmed by the isolation of GFP-positive cells from GFP-chromophore randomized libraries. NGS results further confirmed the overrepresentation of GFP-chromophore amino acid sequences in isolated GFP-positive cells.

We here choose BssSI (also known as BauI) restriction enzyme to digest the dsDNAs. The non-palindromic characteristics of BssSI bring both advantages and disadvantages. BssSI sticky ends block up- to up-stream or down- to down-stream dsDNA ligations. This leads to fewer bands on the gel after electrophoresis of ligation products and a higher yield of ligation products compared to the palindromic cutting sites. However, the number of commercially available non-palindromic endonucleases is far less than that of palindromic endonucleases. Thus, in some sequence contexts, there may be no appropriate non-palindromic restriction site for *in-vitro* ligation.

The piggyBac system is highly effective for integrating foreign DNA, usually plasmid, into the mammalian host genome (19). piggyBac transposase recognizes ITR sequences on both ends of the cargo sequence and inserts them into chromosomal TTAA sites. In this work, for the first time, we showed that *in-vitro* ligation dsDNAs and PCR products containing ITR sequences on both ends could be integrated into the mammalian genome by the piggyBac system (Figure 1B). As revealed by clone formation assay, the insertion efficiency of PCR products is comparable to that of plasmids (Figure 1B). The integration efficiencies of PCR products and plasmid are 11.0% and 11.3%, respectively (Figure 1C). However, the integration efficiency of *in-vitro* ligation dsDNAs is only 1.61%, lower than that of PCR products. The underlying reason requires further investigation.

In principle, other transposon systems, such as Sleeping Beauty (25) and Tol2 (26), could also integrate *in-vitro* ligation dsDNAs, containing the corresponding transposase recognition

sequences on both ends, into the mammalian genome. Moreover, instead of ssODNs or plasmids, the *in-vitro* ligation dsDNAs could serve as homologous repair templates for CRISPR/Cas9-mediated HDR. And by pre-integrating a single landing pad per cell, the *in-vitro* ligation dsDNA library could be inserted into the genome in a single-variant-per-cell manner (3,11,13).

As an initial application of this strategy, we integrated the EGFP expression cassettes containing randomized chromophore amino acid coding sequences into the HEK29T cells (cultured in 1 well of 6-well plate) and attempted to isolate GFP-positive cells (Figure 2A). The theoretical frequencies of the EGFP chromophore amino acid sequence for 2-NNK and 3-NNK libraries are 0.195% and 0.0122%, respectively. After two rounds of FACS selection, both the 2-NNK group and the 3-NNK group enriched the cell population with over 90% of GFP-positive cells. NGS was then employed to analyze the chromophore sequences of enriched GFP-positive cells. Indeed, the EGFP-chromophore amino acid sequences TY and TYG both ranked first (Table 1). This may be because the EGFP can give a brighter fluorescence signal over other GFP-chromophores (21). Moreover, NGS results also revealed that some amino acid sequences with reported GFP absorption-emission characteristics were also among the top lists. All these results demonstrated the feasibility of the strategy.

In conclusion, we developed a novel piggyBac-based large variant library integration approach. It can be applied to perform DMS in cultured mammalian cells.

Figure Legends

Figure 1. Schematic diagram and abundance analysis of piggyBac-mediated *in-vitro* ligation dsDNAs genomic integration. (A) Using *in-vitro* ligation strategy to generate dsDNA containing a Puro-resistant gene expression cassette, which is integrated into the genome by the piggyBac system to make cells resistant to puromycin. (B) Forty-eight hours after transfection, HEK293T cells were diluted 1,000-fold and cultured with puromycin for ten days. Cell clone formation was shown by crystal violet staining. PCR Product: PCR product containing Puro-resistance cassette; Plasmid: plasmid containing Puro-resistance cassette; no-transfection: HEK293T cells that have not been transfected. (C) Statistics of the number of cell clones. The total number of clones formed in each well of a 6-well plate is the clone number in the corresponding well in (B) multiplied by 1,000.

Figure 2. piggyBac system mediated genomic integration of GFP chromophore amino acid randomized library and the recovery of GFP positive cells. (A) Schematic diagram of piggyBac mediated genomic integration GFP chromophore random library (3-NNK). Puromycin resistance gene links to GFP coding region through a P2A element. FACS isolation of GFP positive cells from 2-NNK (B) and 3-NNK (C) library.

Figure 3. FACS isolation of BFP positive cells from the 2-NNK library.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [31870860 to S.L., 31971388 to W.S.].

Competing Interests

The authors have declared that no competing interest exists.

References

1. Araya, C.L. and Fowler, D.M. (2011) Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol*, **29**, 435-442.
2. Fowler, D.M. and Fields, S. (2014) Deep mutational scanning: a new style of protein science. *Nat Methods*, **11**, 801-807.
3. Mason, D.M., Weber, C.R., Parola, C., Meng, S.M., Greiff, V., Kelton, W.J. and Reddy, S.T. (2018) High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. *Nucleic Acids Res*, **46**, 7436-7449.
4. Hanning, K.R., Minot, M., Warrender, A.K., Kelton, W. and Reddy, S.T. (2021) Deep mutational scanning for therapeutic antibody engineering. *Trends Pharmacol Sci*.
5. Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S.,

Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C. *et al.* (2020) Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*, **182**, 1295-1310 e1220.

6. Burton, T.D. and Eyre, N.S. (2021) Applications of Deep Mutational Scanning in Virology. *Viruses*, **13**.

7. Narayanan, K.K. and Procko, E. (2021) Deep Mutational Scanning of Viral Glycoproteins and Their Host Receptors. *Front Mol Biosci*, **8**, 636660.

8. Ma, L., Boucher, J.I., Paulsen, J., Matuszewski, S., Eide, C.A., Ou, J., Eickelberg, G., Press, R.D., Zhu, L.J., Druker, B.J. *et al.* (2017) CRISPR-Cas9-mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. *Proc Natl Acad Sci U S A*, **114**, 11751-11756.

9. Jones, E.M., Lubock, N.B., Venkatakrishnan, A.J., Wang, J., Tseng, A.M., Paggi, J.M., Latorraca, N.R., Cancilla, D., Satyadi, M., Davis, J.E. *et al.* (2020) Structural and functional characterization of G protein-coupled receptors with deep mutational scanning. *Elife*, **9**.

10. Heredia, J.D., Park, J., Brubaker, R.J., Szymanski, S.K., Gill, K.S. and Procko, E. (2018) Mapping Interaction Sites on Human Chemokine Receptors by Deep Mutational Scanning. *J Immunol*, **200**, 3825-3839.

11. Matreyek, K.A., Stephany, J.J. and Fowler, D.M. (2017) A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res*, **45**, e102.

12. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C. and Shendure, J. (2014) Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*, **513**,

120-123.

13. Erdogan, M., Fabritius, A., Basquin, J. and Griesbeck, O. (2020) Targeted In Situ Protein Diversification and Intra-organelle Validation in Mammalian Cells. *Cell Chem Biol*, **27**, 610-621 e615.

14. Galan, A., Comor, L., Horvatic, A., Kules, J., Guillemin, N., Mrljak, V. and Bhide, M. (2016) Library-based display technologies: where do we stand? *Mol Biosyst*, **12**, 2342-2358.

15. Beerli, R.R., Bauer, M., Buser, R.B., Gwerder, M., Muntwiler, S., Maurer, P., Saudan, P. and Bachmann, M.F. (2008) Isolation of human monoclonal antibodies by mammalian cell display. *Proc Natl Acad Sci U S A*, **105**, 14336-14341.

16. Waldmeier, L., Hellmann, I., Gutknecht, C.K., Wolter, F.I., Cook, S.C., Reddy, S.T., Grawunder, U. and Beerli, R.R. (2016) Transpo-mAb display: Transposition-mediated B cell display and functional screening of full-length IgG antibody libraries. *MAbs*, **8**, 726-740.

17. Hellmann, I., Waldmeier, L., Bannwarth-Escher, M.C., Maslova, K., Wolter, F.I., Grawunder, U. and Beerli, R.R. (2018) Novel Antibody Drug Conjugates Targeting Tumor-Associated Receptor Tyrosine Kinase ROR2 by Functional Screening of Fully Human Antibody Libraries Using Transpo-mAb Display on Progenitor B Cells. *Front Immunol*, **9**, 2490.

18. Zhao, Y., Wang, Y., Su, W. and Li, S. (2021) Construction of Synthetic Nanobody Library in Mammalian Cells by dsDNA-Based Strategies*. *Chembiochem*, **22**, 2957-2965.

19. Woodard, L.E. and Wilson, M.H. (2015) piggyBac-ing models and new therapeutic strategies. *Trends Biotechnol*, **33**, 525-533.

20. Yusa, K., Zhou, L., Li, M.A., Bradley, A. and Craig, N.L. (2011) A hyperactive piggyBac transposase for mammalian applications. *Proc Natl Acad Sci U S A*, **108**, 1531-1536.

21. Cormack, B.P., Valdivia, R.H. and Falkow, S. (1996) FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, **173**, 33-38.

22. Laitinen, O.H., Airenne, K.J., Hytonen, V.P., Peltomaa, E., Mahonen, A.J., Wirth, T., Lind, M.M., Makela, K.A., Toivanen, P.I., Schenkwein, D. *et al.* (2005) A multipurpose vector system for the screening of libraries in bacteria, insect and mammalian cells and expression in vivo. *Nucleic Acids Res*, **33**, e42.

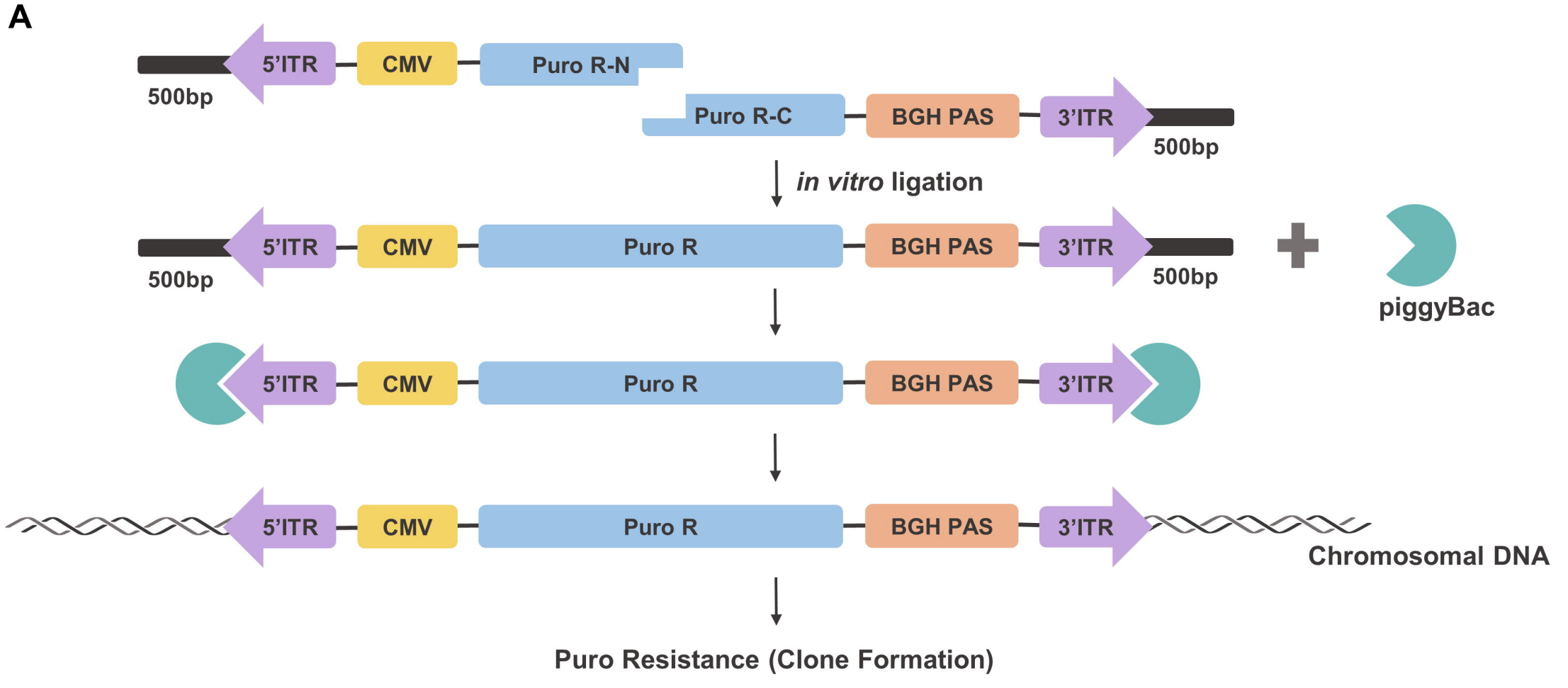
23. Cubitt, A.B., Heim, R., Adams, S.R., Boyd, A.E., Gross, L.A. and Tsien, R.Y. (1995) Understanding, improving and using green fluorescent proteins. *Trends Biochem Sci*, **20**, 448-455.

24. Jung, G., Wiehler, J. and Zumbusch, A. (2005) The photophysics of green fluorescent protein: influence of the key amino acids at positions 65, 203, and 222. *Biophys J*, **88**, 1932-1947.

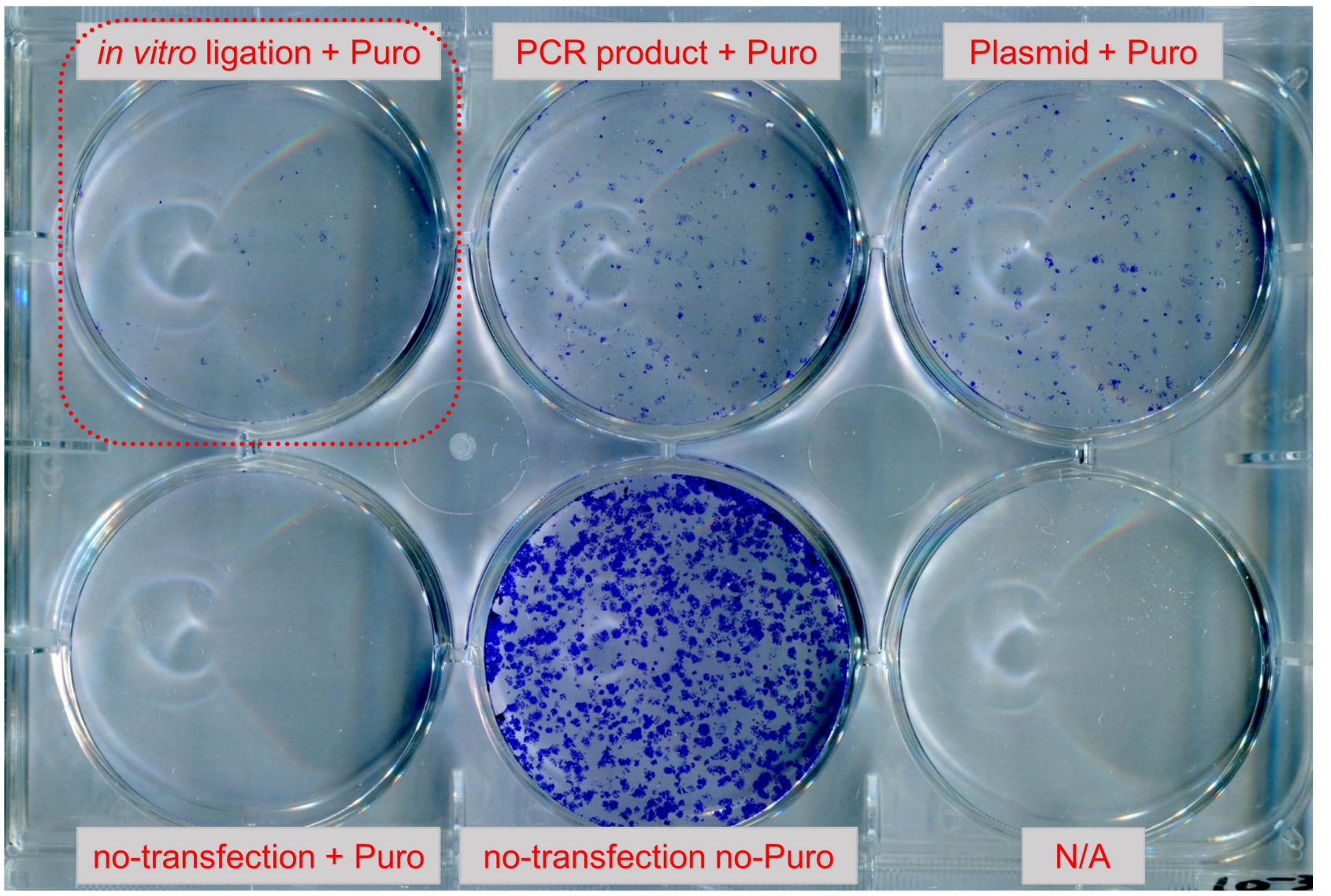
25. Ivics, Z., Hackett, P.B., Plasterk, R.H. and Izsvak, Z. (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, **91**, 501-510.

26. Kawakami, K. (2007) Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol*, **8 Suppl 1**, S7.

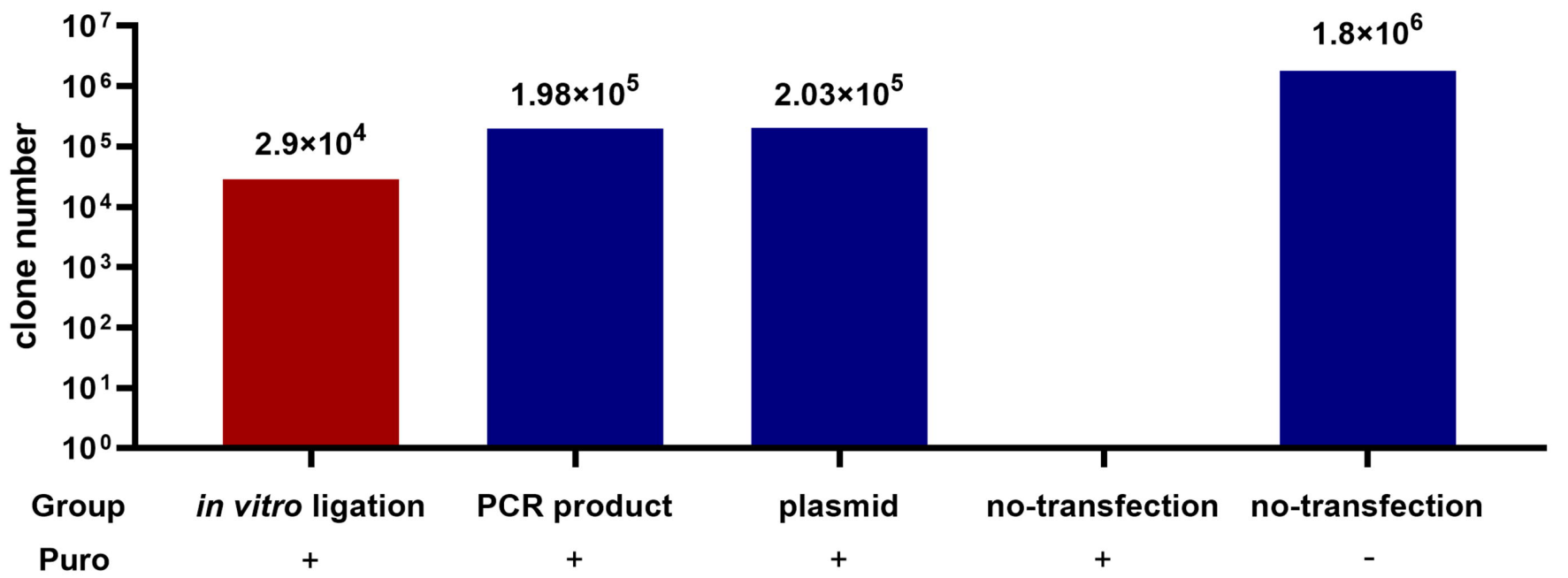
A

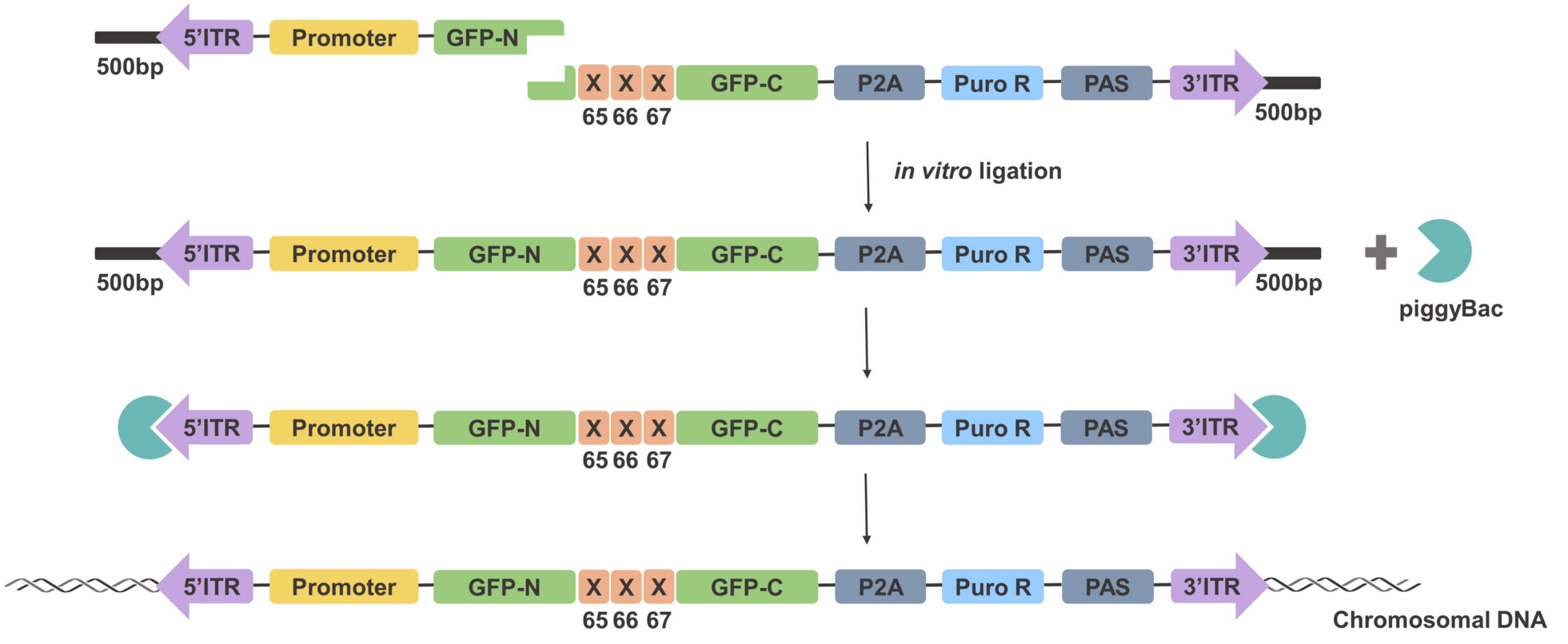

5'ITR
CMV
Puro R-N
500bp
Puro R-C
BGH PAS
3'ITR
500bp
in vitro ligation
5'ITR
CMV
Puro R
BGH PAS
3'ITR
500bp
500bp
piggyBac
5'ITR
CMV
Puro R
BGH PAS
3'ITR
5'ITR
CMV
Puro R
BGH PAS
3'ITR
Chromosomal DNA
Puro Resistance (Clone Formation)

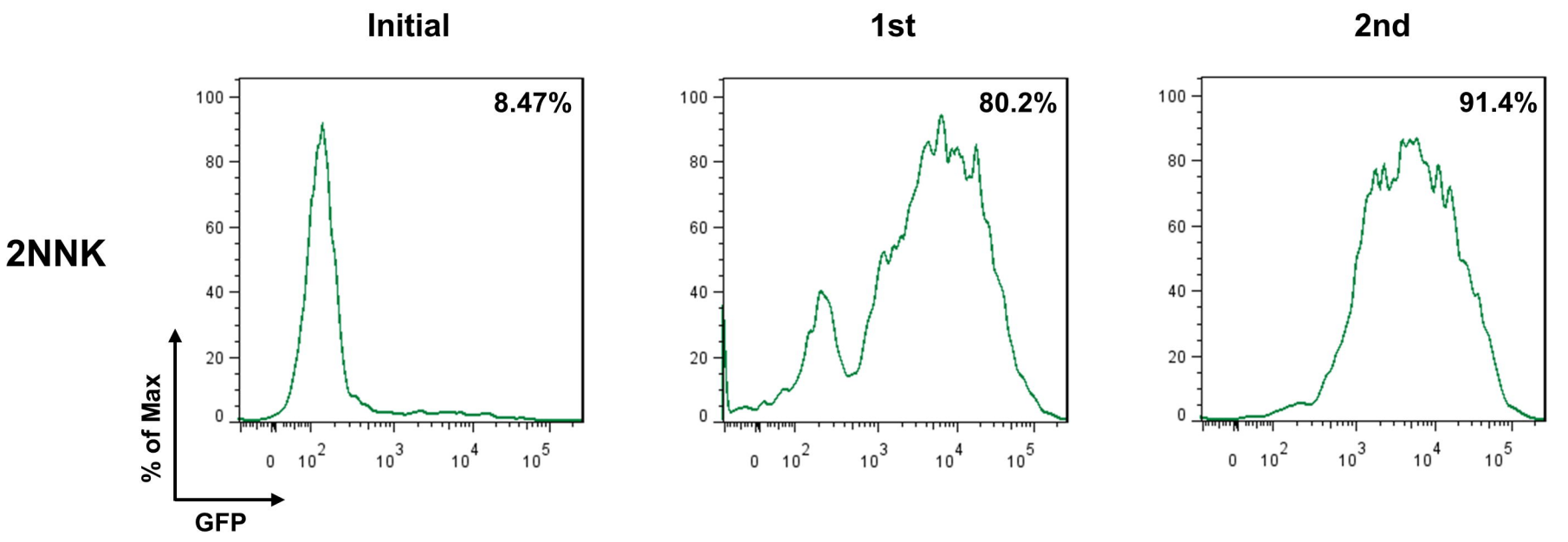
B

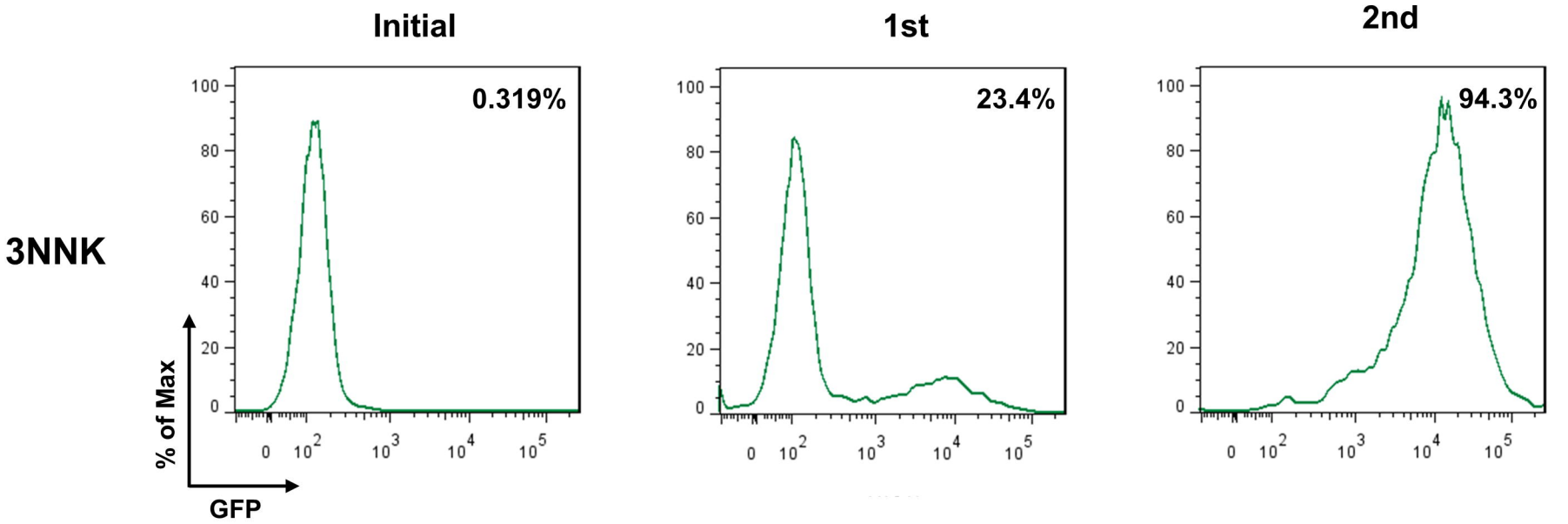

1000 X dilution
in vitro ligation + Puro
PCR product + Puro
Plasmid + Puro
no-transfection + Puro
no-transfection no-Puro
N/A

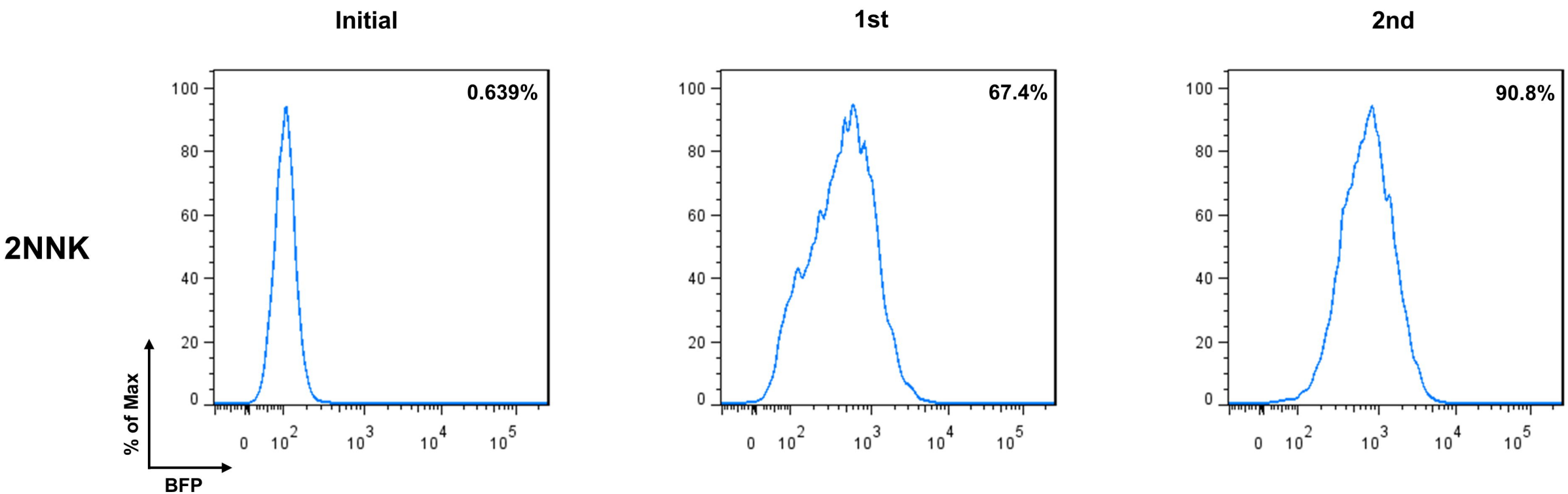
C


clone number
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
2.9×10^4
1.98×10^5
2.03×10^5
1.8×10^6
Group
in vitro ligation
PCR product
plasmid
no-transfection
no-transfection
Puro
+
+
+
+
-


A
5'ITR
Promoter
GFP-N
500bp
X X X
65 66 67
GFP-C
P2A
Puro R
PAS
3'ITR
500bp
in vitro ligation
piggyBac
Chromosomal DNA


B
Initial
1st
2nd
2NNK
8.47%
80.2%
91.4%
% of Max
GFP


C
Initial
1st
2nd
3NNK
0.319%
23.4%
94.3%
% of Max
GFP


Initial
1st
2nd
2NNK
0.639%
67.4%
90.8%
% of Max
BFP
100
80
60
40
20
0
0
10^2
10^3
10^4
10^5