Published in final edited form as:

Biotechnol J. 2018 October ; 13(10): e1800226. doi:10.1002/biot.201800226.

Single Copy Transgene Integration in a Transcriptionally Active Site for Recombinant Protein Synthesis

Sofie A. O'Brien¹, Kyoungcho Lee^{1,α}, Hsu-Yuan Fu^{1,β}, Zion Lee¹, Tung S. Le¹, Christopher S. Stach², Meghan G. McCann¹, Alicia Q. Zhang¹, Michael J. Smanski², Nikunj V. Somia³, Wei-Shou Hu^{1,*}

¹Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455-0132 USA

²Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, MN 55455-0132 USA

³Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN 55455-0132 USA

Abstract

For the biomanufacturing of protein biologics, establishing stable cell lines with high transgene transcription is critical for high productivity. Modern genome engineering tools can direct transgene insertion to a specified genomic locus and can potentially become a valuable tool for cell line generation. In this study, we surveyed transgene integration sites and their transcriptional activity to identify characteristics of desirable regions. A lentivirus containing destabilized Green Fluorescent Protein (dGFP) was used to infect Chinese hamster ovary cells at a low multiplicity of infection, and cells with high or low GFP fluorescence were isolated. RNA sequencing and Assay for Transposase Accessible Chromatin using sequencing data showed integration sites with high GFP expression are in larger regions of high transcriptional activity and accessibility, but not necessarily within highly transcribed genes. This method was used to obtain high Immunoglobulin G (IgG) expressing cell lines with a single copy of the transgene integrated into transcriptionally active and accessible genomic regions. Dual recombinase-mediated cassette exchange was then employed to swap the IgG transgene for erythropoietin or tumor necrosis factor receptor-Fc. This work thus highlights a strategy to identify desirable sites for transgene integration and to streamline the development of new product producing cell lines.

*Corresponding Author Wei-Shou Hu, Address: 421 Washington Avenue SE, Minneapolis, MN 55455-0132 USA, Phone: (612) 626-7630, Fax: (612) 626-7246, acre@umn.edu.

^αPresent address: Sanofi, Paris, France

^βPresent address: MycoMagic Co., Ltd., New Taipei City, Taiwan

CONFLICT OF INTEREST

The authors declare no commercial or financial conflict of interest.

Publisher's Disclaimer: This is the peer reviewed version of the following article: O'Brien, S. A.; Lee, K.; Fu, H.-Y.; Lee, Z.; Le, T. S.; Stach, C. S.; McCann, M. G.; Zhang, A. Q.; Smanski, M. J.; Somia, N. V.; Hu, W.-S. Single Copy Transgene Integration in a Transcriptionally Active Site for Recombinant Protein Synthesis. *Biotechnology Journal* 2018, 13 (10), 1800226. <https://doi.org/10.1002/biot.201800226>. which has been published in final form at <https://doi.org/10.1002/biot.201800226>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Keywords

Bioprocess engineering; Cell culture; CHO cells; Protein expression; Recombinant proteins

1 INTRODUCTION

Protein therapeutics have transformed the treatment of many complex diseases. In the past five years, over five dozen recombinant therapeutic proteins were approved by the FDA.^[1] These new products include monoclonal antibodies directed at PD-1 (Keytruda (Merck) and Opdivo (Bristol Meyers Squibb)) and PD-L1 (Tecentriq; Genentech) for treating melanoma and bladder cancers, Fc fusion protein of Factor VIII/IX for hemophilia (Eloctate and Alprolix; Biogen), and bispecific antibodies anti-CD19/anti-CD3 for the treatment of leukemia (Blincyto; Amgen). Many more biologics are currently under development and in late phase clinical trials. With this new generation of biologics, it is increasingly important to have efficient methods for their production.

Chinese Hamster Ovary (CHO) cells are the most commonly used host cells for the manufacture of recombinant therapeutics.^[2, 3] Traditionally, production cell lines are generated by random integration of the product transgene into the host cell followed by gene amplification and screening for high producing cell clones. Although this method of cell line generation has been in practice for over three decades and such cell lines have been used to produce the majority of the products on the market, the success of the method relies on screening a large number of clones for their productivity and stability over time.^[4-8] With the advances in genome engineering, new ways of generating cell lines, especially directing the integration of transgenes into specific loci in the host cell genome, are being explored.

There are a variety of tools available for the targeted integration of transgenes into specific loci, such as zinc finger nucleases, TALE nucleases, and CRISPR/Cas9.^[9-11] Another possible strategy called Recombinase Mediated Cassette Exchange (RMCE) allows for replacement of the sequence between two non-compatible recombination sites, previously inserted into the genome, with a new gene. A number of methods have been developed for RMCE using site specific recombinases, such as the Cre/*lox*, Flp/*FRT*, and Φ C31/*attP*-B systems.^[12] Integration of transgenes into CHO cells has been achieved at a reasonable efficiency using both targeted integration^[13-15] and cassette exchange.^[16-19] Another similar method, dual RMCE, utilizes two different recombinases (such as Flp and Cre). Dual RMCE was reported to have a higher integration efficiency than traditional RMCE by preventing excision or other internal recombination effects,^[20, 21] and has been used efficiently in CHO.^[22] The success of these methods hinges on prior knowledge of a desirable site for transgene integration. One approach is to select a site identified from a high producing cell line. However, a high producing cell line generated using past methods often has multiple transgenes integrated into the host cell genome, and identifying the site that contributes most to transgene transcription is time consuming, often involving serial knockout of the multiple transgene copies.

The integration site of a transgene has an effect on its expression. Insertion of a transgene into locations with known transcriptional activity has been suggested to increase its

expression.^[23, 24] Additionally, expression of reporter genes at a particular site has been shown to be affected by proximity to active genes.^[25] Using PiggyBac transposons or HIV-1 based lentivectors, transgenes have been preferentially inserted into transcriptionally active genomic regions.^[26-29] These methods have been used for product transgene integration in CHO cells to obtain high producing cell lines.^[30-32] The transcriptional activity of a gene is also affected by regional chromatin organization and accessibility.^[33-35] Recently, a decrease of transgene productivity in CHO cells was reported to correlate with increased nucleosome occupancy.^[36]

Transcriptional activation has been shown to be associated with regions of the genome lacking nucleosomes, which are often referred to as open or accessible chromatin.^[34] A number of methods are commonly used to assess the openness of chromatin structure, including MNase-seq,^[37] DNase-seq,^[38] and FAIRE-seq.^[39] Assay for Transposase Accessible Chromatin using sequencing (ATAC-seq) has recently emerged as an easily accessible method for genome wide evaluation of chromatin accessibility.^[40] In this method, a hyperactive Tn5 transposase tags native chromatin from cellular nuclei with high-throughput sequencing adapters, simultaneously fragmenting the DNA. After sequencing of the digested fragments and mapping to the genome, the frequency of the appearance of a particular sequence is indicative of the accessibility of the corresponding genomic locus.^[41] Regions low in ATAC-seq signals are inferred as being occupied by nucleosomes.^[34]

In this study, we used a lentiviral vector with a reporter gene to evaluate the relationship between transgene expression and the local context of the genome in terms of transcriptional activity and accessibility. We further demonstrate the establishment of Immunoglobulin G (IgG) producing cell lines with a single integration site in regions with high transcriptional activity and open chromatin. Finally, we utilized dual RMCE by replacing the IgG transgene with two different product genes, streamlining the creation of new product producing cell lines.

2 EXPERIMENTAL SECTION

2.1 Vector Construction

The lentivectors pLOVE-dGFP and pLOVE-IgG-dGFP were generated using pLOVE empty vector (addgene Plasmid #15948) as a backbone. pLOVE empty vector was digested with PacI and AscI to insert either destabilized GFP (dGFP), or IgG-IRES-dGFP flanked by recombination sites (Supplementary Figure S1A and B). The IgG-IRES-dGFP insert contains IgG heavy chain and light chain linked by a porcine teschovirus-1 2A (P2A) linker peptide,^[42] followed by an internal ribosome entry site (IRES) to drive translation of dGFP. The IgG is flanked by Lox511 and a minimal 34 bp flippase recognition target (FRT) site.^[43]

Donor vectors Lox511-PuroR-P2A-mCherry-P2A-TNFR-Fc-FRT and Lox511-PuroR-P2A-mCherry-P2A-EPO-FRT were constructed using Gibson assembly^[44] to combine the different elements with a minimal backbone containing a Kanamycin resistance gene for selection in bacteria and a ColE1 origin of replication. The vector is promoter-less, and contains Puromycin resistance, mCherry (a red fluorescent protein), and a product gene

(either Entanercept (a TNFR-Fc fusion protein) or Darbepoetin alfa (an engineered variant of Erythropoietin (EPO)), all linked via P2A, and flanked by Lox511 and a minimal FRT site. A schematic of the plasmid map is shown in Supplementary Figure S1C.

Recombinase vector F2AC contains flippase (Flpe) linked to cre recombinase via a *Thoseaasigna* virus 2A (T2A) linker and driven by a CMV promoter.^[22]

2.2 Cell Culture

CHO-K1 cells (ATCC CCL-61, Manassas, VA) and derived subclones were cultured in F12K Medium (Kaighn's Modification, Gibco, Waltham, MA) supplemented with 10% Fetal Bovine Serum (Gibco, Waltham, MA) and incubated at 37°C in 5% CO₂. Single cell cloning was performed via limiting dilution. Cells were diluted to 1 cell per well (100µL) and plated into tissue culture treated 96-well plates.

2.3 lentivirus production and infection

Lentiviruses were produced in 293T cells by co-transfection of pLOVE derived vectors along with helper plasmids psPAX2 and pMD2.G (addgene Plasmid #12260 and #12259). Medium was collected 24 and 48 hours after transfection and filtered through a 0.45 µm filter. Virus was then added to 10⁷ CHO-K1 (ATCC CCL-61, Manassas, VA) cells at a multiplicity of infection (MOI) of 0.01 or 0.02.

A description of all other methods, including integration site analysis, bioinformatics analysis, dual RMCE, qRT-PCR, ELISA, and statistical analysis can be found in the Supplementary Information.

3 RESULTS

3.1 Lack of correlation between transgene expression and genetic/intergenic status of integration sites

To establish a system for characterizing transgene integration sites, CHO-K1 cells were infected with a lentiviral vector containing destabilized GFP (dGFP) at a low MOI of 0.01 to reduce multiple integration events (Supplementary Figure S1A). We used dGFP with a half-life of 2 hours as a reporter^[45] to better reflect transcriptional activity. Forty-eight hours after infection, the top 1% and bottom 10% of GFP⁺ cells were sorted using FACS and subsequently cultured for weeks to expand the cell population for further characterization. The integration site of the lentivirus genome was PCR amplified using primer pairs flanking the junction of the CHO genome and the lentivector, as described in the Supplementary Methods. Amplicons from the two populations were isolated for PacBio sequencing (Supplementary Figure S2). Sequencing resulted in approximately 770K reads per sample, with an average read length of 1.6 kbp. Reads containing both genome and vector regions were isolated using BLASTn. Initial filtering removed reads with less than a 400 bp alignment to the lentiviral LTR region. The remaining reads were aligned to the Chinese hamster genome, and the integration location for each read was identified. (Supplementary Figure S3).

For the top 1% population, 41.7% and 4.9% of reads mapped exclusively to lentiviral sequences and the host cell genome respectively. From the remaining 53.4% of reads, 272 integration sites on host genome were identified after low confidence regions were removed. Over half (64.3%) of the integration sites were located in intergenic regions, 34.9% were within introns, and the remaining 0.7% of sites within exons (Figure 1A). In the bottom 10% population, 254 integration sites were identified. These integration sites were distributed over intergenic, intron, and exon regions in a similar fashion as those in the top 1% population. As a measure of the abundance of these features within the genome, ten thousand randomly selected sites within the genome were examined, and the distribution of gene features was similar to the top 1% and bottom 10% pools. With respect to the gene features of the integration location, the top 1% and bottom 10% thus did not appear to be different from each other or the distribution of these features in the genome.

3.2 Isolation of integration sites for high expressing GFP clones

In a second experiment, we infected CHO-K1 cells with the same dGFP lentivirus at an MOI of 0.01~0.02, and cells in the top 1% of green fluorescence were single-cell cloned and expanded. A total of 89 high-GFP intensity clones were obtained. The junction region of the integration site was amplified by PCR, and the amplified DNA was separated by agarose gel electrophoresis to identify the number of integration sites, or the number of copies of the dGFP lentivirus in each clone. We identified 34 clones as having a single lentivirus integrated into the CHO genome. The rest of the clones were similarly distributed between having 2, 3, or 4+ copies. Gel bands from the single copy clones were excised for DNA sequencing using the Sanger method, and the integration sites were then identified by BLASTn against the Chinese hamster genome. The integration sites for these single copy clones were mostly in introns (79.4%), and the remaining 20.6% were in intergenic regions (Figure 1A). This distribution of gene features is different from integration locations of the top 1% pool, bottom 10% pool, and random locations, which were mostly in intergenic regions. It should be noted that the GFP expression in all clones is attributable to a single integration site while some cells in the pool contain multiple copies. This may account for the difference observed.

3.3 Activity of Region Surrounding Integration Site Correlates with GFP Expression

We next asked the question whether high transcriptional activity and genome accessibility in the region surrounding the locus of integration coincides with high transgene expression as reflected in GFP intensity. As an assessment of the regional transcriptional activity and accessibility, RNA-seq and ATAC-seq data from CHO-K1 cells was used to characterize the region 50 kbp upstream and downstream of the integration site. The sum of all RNA-seq and ATAC-seq reads were normalized to Reads per Kilobase per Million Reads (RPKM) and tabulated.

Density distributions of the regional transcriptional activity of each integration site for the three populations of GFP⁺ cells (top 1% single-copy clones, top 1% pool, and bottom 10% pool) were plotted in Figure 1B. As a comparison, 10,000 sites randomly selected from the genome were also plotted. The distributions of the RNA-seq data were not unimodal, as there were many integration sites with low RPKM in addition to a broad peak around

Log₂[RPKM] of zero. Hence, a Kolmogorov-Smirnov test for the equality of two different distributions was used to compare the different data sets (Table 1). In general, the top 1% pool and top 1% clone distributions were shifted towards higher activity, while the bottom 10% pool and random population were skewed towards lower activity. The transcriptional activity of the integration sites for the bottom 10% population had a similar distribution to the random population ($p = 0.093$), while that of both the top 1% pool and top 1% clones were significantly different from the bottom 10% as well as the random population ($p < 0.001$ for all comparisons).

The population density distribution of the top 1% single copy clones had a similar distribution to the top 1% pool, except skewing towards higher expression (Figure 1B). For cells in the top 1% pool, a fraction of the population is likely to have more than one copy. Thus, the GFP intensity may have resulted from combined expression from multiple loci, or from a major, high-expressing locus with accompanying inactive loci.

For integration sites located within genes, we also examined whether there was a relationship between GFP fluorescence and the RNA-seq activity of those genes. In this case, the distributions for the top 1% and the bottom 10% pools were not significantly different (Supplementary Figure S4 and Supplementary Table S1, $p = 0.129$). The data thus implies that cells isolated via sorting for high GFP fluorescence were more likely to have integrated into general regions of high transcriptional activity, while the expression of the specific gene at the integration site may not be relevant.

ATAC-seq data was quantified as described in the Supplementary Methods for each integration site identified. The number of ATAC-seq reads from + 50 kbp to - 50 kbp of each integration site were normalized and reported as ATAC-seq RPKM. The population density distribution for each of the four data sets, expressed as log₂ transformed ATAC-seq RPKM, is shown in Figure 1C. Significance of differences among the median of these distributions was examined using the non-parametric Mann-Whitney Test (Table 1). All three GFP⁺ populations were significantly different from the random set ($p < 0.005$). Additionally, the distribution of both top 1% sets, the single copy clones as well as the pool, had significantly higher ATAC-seq expression over the bottom 10% pool ($p < 0.001$). No significant difference was observed between the single copy and pooled top 1% cell integration sites ($p = 0.51$).

qRT-PCR was then used to measure the GFP transcript level of the three GFP⁺ populations (Figure 1D). The pooled top 1% of GFP⁺ cells had significantly higher transcript of GFP compared to the pooled bottom 10% (2^5 (32) fold higher, T-test, $p < 0.001$). Among the 34 clones isolated from the top 1% of GFP⁺ cells, the expression of GFP falls within a narrow range. Their transcription level of GFP was similar to the top 1% pool, which is expected as the clones were isolated from a similar population of top 1% GFP positive cells. The data thus suggests a positive correlation between GFP fluorescence intensity, GFP transcript level, and integration into regions with higher transcriptional activity and accessibility as measured by RNA and ATAC sequencing.

3.4 Single Copy IgG Producing Cell Line

Having shown that the top GFP⁺ cells obtained by infection with a low MOI of lentivirus would have a higher probability of landing the transgene into a transcriptionally active and accessible region of the CHO genome, we set out to use the same method to integrate a lentiviral vector containing IgG heavy chain and light chain genes. It has been previously shown that GFP fluorescence correlated with titer of a co-expressed product gene,^[30] which further supports our method. The heavy chain and light chain genes were followed by an internal ribosome entry site (IRES) to drive translation of dGFP (Supplementary Figure S1B). The vector also contained recombination sites for flippase and cre recombinase flanking the IgG sequence, such that the product gene could be replaced at a later date.

CHO-K1 cells were infected with IgG-IRES-dGFP lentivirus at an MOI of 0.02. Forty-eight hours after infection, the top 10% of GFP⁺ cells were collected using FACS (Supplementary Figure S5), cultivated for 10 days, and then sorted again. The top 0.1% of GFP positive cells (2,982 cells) were collected into separate wells for further cell expansion. After 14 days, supernatant was collected from the 303 surviving clones and the IgG level was quantified using ELISA (Supplementary Figure S6).

The 52 clones with the highest IgG levels were further expanded in 6-well plates for 5 days before the IgG titer and transcript levels were quantified. The IgG titer in the supernatant had a large range (Figure 2A). Transcript levels for heavy chain and light chain of IgG were quantified using a one-step qRT-PCR method that, which although consistently 1-2 cycles different from the standard method we used, was better suited for screening a large number of clones, as described in the Supplementary Methods (Figure 2B). All heavy and light chain transcript levels were similar for most clones. The expression level ranged from $2^{0.7}$ to $2^{5.7}$ (1.62 to ~52) fold higher than GAPDH for light chain and from $2^{0.5}$ to $2^{6.8}$ (1.4 to ~111) fold higher for heavy chain among different clones.

The number of copies of lentivector integrated in each clone was then determined. Out of the 52 top IgG producing clones, 8 (15.4%) had a single copy, 17 (32.7%) had two copies, 10 (19.2%) had three copies, and 17 (32.7%) had more than 3 copies (Figure 2A and Figure 3A). Notably, copy number did not seem to be directly correlated with titer. Some single copy clones had the same titer or even a higher titer than clones with greater than three copies.

3.5 Three high IgG secreting clones express high level of transcripts

Three of the single copy clones, 2C10, 3A10, and 4B1, were characterized further. Both light chain and heavy chain transcripts were at levels higher than GAPDH in all three clones (Figure 3B). Specific productivity of IgG for these clones ranged from 18 – 29 pg/cell/day (Figure 3C).

The integration site of the vector in all three clones was identified. ATAC and RNA-seq of the host cell, CHO K1, were used to examine the accessibility and the transcriptional activity in the region of these three integration sites (Figure 3D-F). In the 100 kbp region surrounding the integration site, all three clones had high RNA-seq and ATAC-seq read density (Figure 1B and C, marked ∇). This implies the clones integrated into active regions

with open chromatin, which agrees with our findings from the original GFP lentivirus experiment showing high expression from active sites.

3.6 Transgene Replacement Using Dual Recombinase Mediated Cassette Exchange (RMCE)

The 2C10 cell line producing IgG was used to generate cell lines producing a new product using dual recombinase mediated cassette exchange (RMCE). For recombination, we employed a recombinase vector encoding flippase and cre recombinase. 2C10 was transfected with a donor vector containing either the sequence for Darbepoetin alfa, an engineered variant of Erythropoietin (EPO), or the sequence for Entanercept, a TNFR-Fc fusion protein, along with the recombinase vector. The donor vector contained puromycin resistance and mCherry, and the sequence for insertion was flanked by Lox511 and minimal FRT recombination sites that match those on the original lentivector. This vector was designed without a promoter for the mCherry gene, such that random integrants would not be selected during sorting (Figure 4A). The product gene was also promoter-less, meaning that expression of the product gene from random integration was highly unlikely. Transgene expression from random integration could only occur if the vector randomly integrated after an endogenous promoter.

mCherry⁺ and GFP⁺ cells were isolated using FACS five days after transfection. (Supplementary Figure S7). Clones were obtained by single cell cloning, and correct recombination was confirmed with PCR across both the lox511 and minimal FRT junctions (Supplementary Figure S8). Five EPO producing clones were obtained, and production of the new protein was verified by ELISA. Low cloning efficiency after sorting led to the isolation of only one TNFR-Fc producing clone. In both cases, the number of mCherry⁺ cells isolated was low, indicating that the efficiency of both targeted integration and random integration behind a promoter was very small. Hence, the probability that any of the isolated clones also carry a random integration is nearly nil.

The six clones exhibited very similar growth rates, although they reached somewhat different cell densities (Figure 4B). Transcript levels of EPO varied among the different clones (Figure 4C). The transcript levels of EPO and TNFR were lower than GAPDH, ranging from $2^{0.9}$ to $2^{4.4}$ (1.9 to ~21) fold lower than GAPDH for EPO, and $2^{1.3}$ (2.5) fold lower than GAPDH for TNFR. Specific productivity of EPO varied and followed the same trend as the transcript data (Figure 4D). The TNFR-Fc producing clone had slightly higher specific productivity than the EPO clones.

4 DISCUSSION

In this work, we used a lentiviral vector to generate cells with a very low number of transgenes integrated and showed that integration into regions with increased transcriptionally activity and accessible chromatin structure have a propensity to express the transgene at higher levels. We first demonstrated this using a dGFP reporter gene, and later showed that with an IgG transgene, the isolated high dGFP expressing clones also had higher IgG productivity. We employed a lentiviral vector to avoid the heterogeneous integration events associated with random integration of a transgene. In random integration,

the vector is often truncated or ligated head-to-head or head-to-tail. Even if the copy number of the integrated transgene is determined, whether or not the gene is active or non-functional cannot be known without resorting to detailed genome sequencing. Whereas using a lentiviral vector, the integrated vector is largely intact. We also employed a dGFP that is tagged with a degradation domain to give a shorter half-life, which increases the dynamic range of its reporting of transcriptional activity.^[45] An earlier effort in our lab using wild type GFP did not result in the isolation of cells with very high transgene transcript levels.

We noted that despite the very low MOI used, a large fraction of cells still had multiple copies of the transgene integrated. This could be attributed to aggregation of virus particles, resulting in multiple viruses entering a cell during a single infection event. Hence, the high IgG expressing clones isolated using the dGFP lentivirus were still examined for a single genome integration event via identification of the integration site. Using this strategy, all of the single copy transgene producing cell lines isolated had IgG transcript levels (both heavy chain and light chain) higher than GAPDH (Figure 2B), our reference standard for a highly expressed gene.

In examining the relationship between high transcriptional activity of the transgene (as indicated by GFP intensity) and the transgene integration site, we employed PacBio sequencing technology to obtain the integration site sequence in a cell pool. The 1.5 kbp reads obtained using PacBio allowed us to identify the vector-genome junction with reasonable confidence, which would not have been possible using methods that give only 100-200 bp reads. Additionally, the depth of sequencing allowed us to identify the genomic integration site despite that nearly half of reads were non-specific (did not cover the vector-genome junction) and each integration site is present at a low level.

Many studies have looked into the mechanisms behind transcriptional activation. It has been found that active genes tend to be within regions that are sensitive to nucleases, such as DNase I, and this sensitivity can extend many kilobases on either end of the gene.^[46] Enhancers are known to recruit transcription factors and remodel regions of chromatin for increased transcriptional activity,^[47] and the majority of enhancers are found within 20-50 kbp of their target gene.^[48] More recently, it has been shown that genes associated with large enhancer regions, called super-enhancers, are expressed at even higher levels.^[49] Though we used ATAC-seq to evaluate pre-integrated sites within the genome, we can further use this chromatin accessibility data to look for enhancer or super enhancer regions in CHO to increase transgene expression in the future.

A very large fraction of the high GFP expressing and high IgG producing cells have transgenes integrated in regions with open chromatin. Heterochromatin invasion is a known mechanism of transcriptional silencing,^[50] which has led to the development of insulating elements to prevent the interference of regional chromatin upon gene integration.^[51] In production cell lines, heterochromatin silencing has been shown to be correlated with instability of recombinant gene expression.^[52] Integration of a transgene into a locus remote from a heterochromatin region may reduce the probability of heterochromatin invasion causing a reduction of the transcriptional activity of the transgene, thus enhancing the stability of the cell line. Four subclones of one of the single copy IgG producing cell lines,

2C10, were isolated and shown to have a similar specific productivity (Supplementary Figure S9). These cells were obtained after subculturing for a long duration and their specific IgG productivity remained unchanged after approximately 35-55 population doublings, suggesting that the transgene expression was indeed stable.

However, when the IgG genes were replaced with EPO by RMCE in the same cell line, the transcript levels varied about ten-fold among the EPO producing clones, as did the specific productivity. The parental cell line, 2C10, was selected for its capability to produce IgG, and the characteristics needed to produce EPO may be different. This may contribute to the variability. Other studies using RMCE to insert a product gene into a defined landing pad have also seen variations among producing clones.^[15, 19, 53, 54] Few had quantified the transgene transcript level, and each had utilized different products and integration sites. Nevertheless, the spread of the specific productivity distribution among swapped clones varied between different reports, ranging from narrow to quite wide. Single cell cloning may have contributed to the variability, which is a well-known phenomenon in CHO cells.^[55, 56] From RNA-seq data of 2C10 and the relative expression of the product genes by qRT-PCR, we estimated that the transcript levels of GAPDH, IgG-H, and IgG-L genes are at the abundance level of the top 0.41%, 0.013% and 0.013% of expressed genes respectively (Supplementary Figure S10). In comparison, the TNFR and EPO producing clones would have their product gene expressed at the level of the top 0.88% to 14.8% of genes. Even though there is some variability in their expression level, overall, the product transcript was still abundant. Therefore, RMCE was still effective in generating cell lines with a high transcript level, though the robustness of the system still requires further study. Whether those clones will also be stable, as the transgenes were integrated in an accessible region of the genome, is yet to be examined.

This study provided evidence that genome regions with high transcriptional activities and open chromatin structure are more amenable to high level expression of integrated transgenes. They may also provide the advantage of long term stability in preserving transgene expression. With advances in genome engineering, one may insert a product transgene into a "favorable" genome region in many ways. We used a destabilized GFP vector and isolated high GFP⁺ and product producing clones. By creating a landing pad (such as a loxP site) or inserting RMCE sites into the transgene vector, one engineered clone could be used to create cell lines producing different products, as we demonstrated with the EPO and TNFR-Fc producing lines. Through RNA-seq and ATAC-seq, one may also identify regions of high preference and employ methods like CRISPR/Cas9 for targeted integration of the product gene into the genome. In a separate study, we examined genome structural variants of CHO cells and uncovered regions that incur frequent structural changes (manuscript submitted for publication). Integration of a transgene into such regions may expose the production cell line to heightened risk of losing its productivity. By targeted integration such regions may be avoided. It is important to note that the integration site of the transgene largely influences only the transcriptional activity and stability. To become a high producing cell, a cell line must be endowed with many additional hyperproductivity traits, including high secretory capacity, enhanced redox balance, and improved energy metabolism.^[57] Furthermore, it must have better growth kinetics under manufacturing conditions. Other factors, such as ribosomal occupancy, have also been shown to affect

productivity.^[58] Thus, by combining the method we presented with further cell line optimization, the process for obtaining a stable, hyperproducing cell line can be streamlined. Once such a cell line is established, the product gene can be exchanged for a new product using site-specific recombinases in order to retain both the integration locus and desirable hyperproductivity traits.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank Atsushi Inoue and Professor Scott McIvor for valuable discussion and Professor Yuri Voznyanov for the Flp-2A-Cre vector and his advice. Computational resources were provided by the Minnesota Supercomputing Institute. This work was supported in part by the Minnesota Futures Grant. SAO and MGM were supported in part by the NIGMS Biotechnology Training Program (T32GM008347- 22). TSL was supported in part by the Vietnam Education Foundation (VEF).

Abbreviations

ATAC-seq	Assay for Transposase Accessible Chromatin using sequencing
CHO	Chinese Hamster Ovary
GFP	Green fluorescent protein
MOI	Multiplicity of infection
RMCE	Recombinase-mediated cassette exchange
RPKM	Reads per kilobase per million reads mapped

5 REFERENCES

- [1]. Lagasse HA, Alexaki A, Simhadri VL, Katagiri NH, Jankowski W, Sauna ZE, Kimchi-Sarfaty C, F1000Research 2017, 6, 113. [PubMed: 28232867]
- [2]. Wurm F, Processes 2013, 1, 296.
- [3]. Bandyopadhyay AA, Khetan A, Malmberg LH, Zhou W, Hu WS, J Ind Microbiol Biotechnol 2017, 44, 785. [PubMed: 28185098]
- [4]. Kim NS, Kim SJ, Lee GM, Biotechnol Bioeng 1998, 60, 679. [PubMed: 10099478]
- [5]. Wirth D, Gama-Norton L, Riemer P, Sandhu U, Schucht R, Hauser H, Current Opinion in Biotechnology 2007, 18, 411. [PubMed: 17904350]
- [6]. Kim M, O'Callaghan PM, Droms KA, James DC, Biotechnol Bioeng 2011, 108, 2434. [PubMed: 21538334]
- [7]. Vishwanathan N, Le H, Jacob NM, Tsao YS, Ng SW, Loo B, Liu Z, Kantardjieff A, Hu WS, Biotechnol Bioeng 2014, 111, 518. [PubMed: 24108600]
- [8]. Le H, Vishwanathan N, Jacob NM, Gadgil M, Hu WS, Biotechnol Lett 2015, 37, 1553. [PubMed: 25971160]
- [9]. Joung JK, Sander JD, Nat Rev Mol Cell Biol 2013, 14, 49. [PubMed: 23169466]
- [10]. Sander JD, Joung JK, Nat Biotechnol 2014, 32, 347. [PubMed: 24584096]
- [11]. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD, Nat Rev Genet 2010, 11, 636. [PubMed: 20717154]
- [12]. Turan S, Zehe C, Kuehle J, Qiao J, Bode J, Gene 2013, 515, 1. [PubMed: 23201421]

- [13]. Cristea S, Freyvert Y, Santiago Y, Holmes MC, Urnov FD, Gregory PD, Cost GJ, *Biotechnol Bioeng* 2013, 110, 871. [PubMed: 23042119]
- [14]. Lee JS, Kallehauge TB, Pedersen LE, Kildegaard HF, *Sci Rep* 2015, 5, 8572. [PubMed: 25712033]
- [15]. Inniss MC, Bandara K, Jusiak B, Lu TK, Weiss R, Wroblewska L, Zhang L, *Biotechnol Bioeng* 2017, 114, 1837. [PubMed: 28186334]
- [16]. Huang Y, Li Y, Wang YG, Gu X, Wang Y, Shen BF, *J Immunol Methods* 2007, 322, 28. [PubMed: 17350648]
- [17]. Crawford Y, Zhou M, Hu Z, Joly J, Snedecor B, Shen A, Gao A, *Biotechnol Prog* 2013, 29, 1307. [PubMed: 23926028]
- [18]. Inao T, Kawabe Y, Yamashiro T, Kameyama Y, Wang X, Ito A, Kamihira M, *Journal of Bioscience and Bioengineering* 2015, 120, 99. [PubMed: 25617966]
- [19]. Zhang L, Inniss MC, Han S, Moffat M, Jones H, Zhang B, Cox WL, Rance JR, Young RJ, *Biotechnol Prog* 2015, 31, 1645. [PubMed: 26399954]
- [20]. Lauth M, *Nucleic Acids Research* 2002, 30, e115. [PubMed: 12409474]
- [21]. Lauth M, Moerl K, Barski JJ, Meyer M, *Genesis* 2000, 27, 153. [PubMed: 10992325]
- [22]. Anderson RP, Voziyanova E, Voziyanov Y, *Nucleic Acids Res* 2012, 40, e62. [PubMed: 22270085]
- [23]. Koduri RK, Miller JT, Thammana P, *Gene* 2001, 280, 87. [PubMed: 11738821]
- [24]. Mielke C, Maass K, Tummler M, Bode J, *Biochemistry* 1996, 35, 2239. [PubMed: 8652565]
- [25]. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, Berns A, Wessels LF, van Lohuizen M, van Steensel B, *Cell* 2013, 154, 914. [PubMed: 23953119]
- [26]. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T, *Cell* 2005, 122, 473. [PubMed: 16096065]
- [27]. Hackett CS, Geurts AM, Hackett PB, *Genome Biol* 2007, 8 Suppl 1, S12. [PubMed: 18047689]
- [28]. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F, *Cell* 2002, 110, 521. [PubMed: 12202041]
- [29]. Wu X, Burgess SM, *Cell Mol Life Sci* 2004, 61, 2588. [PubMed: 15526164]
- [30]. Oberbek A, Matasci M, Hacker DL, Wurm FM, *Biotechnol Bioeng* 2011, 108, 600. [PubMed: 20967750]
- [31]. Matasci M, Baldi L, Hacker DL, Wurm FM, *Biotechnol Bioeng* 2011, 108, 2141. [PubMed: 21495018]
- [32]. Balasubramanian S, Rajendra Y, Baldi L, Hacker DL, Wurm FM, *Biotechnol Bioeng* 2016, 113, 1234. [PubMed: 26616356]
- [33]. Li B, Carey M, Workman JL, *Cell* 2007, 128, 707. [PubMed: 17320508]
- [34]. Tsompana M, Buck MJ, *Epigenetics & Chromatin* 2014, 7.
- [35]. Wallrath LL, Elgin SCR, *Genes & Development* 1995, 9, 1263. [PubMed: 7758950]
- [36]. Veith N, Ziehr H, MacLeod RA, Reamon-Buettner SM, *BMC Biotechnol* 2016, 16, 6. [PubMed: 26800878]
- [37]. Cui K, Zhao K, in *Chromatin Remodeling: Methods and Protocols*, (Ed: Morse RH), Humana Press, Totowa, NJ 2012, Ch. 24.
- [38]. Song L, Crawford GE, *Cold Spring Harb Protoc* 2010, 2010, pdb prot5384.
- [39]. Giresi PG, Kim J, McDaniel RM, Iyer VR, Lieb JD, *Genome Res* 2007, 17, 877. [PubMed: 17179217]
- [40]. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ, *Nat Methods* 2013, 10, 1213. [PubMed: 24097267]
- [41]. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ, *Curr Protoc Mol Biol* 2015, 109, 21.29.1.
- [42]. Szymczak AL, Workman CJ, Wang Y, Vignali KM, Dilioglou S, Vanin EF, Vignali DA, *Nat Biotechnol* 2004, 22, 589. [PubMed: 15064769]
- [43]. Turan S, Galla M, Ernst E, Qiao J, Voelkel C, Schiedlmeier B, Zehe C, Bode J, *J Mol Biol* 2011, 407, 193. [PubMed: 21241707]
- [44]. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA 3rd, Smith HO, *Nat Methods* 2009, 6, 343. [PubMed: 19363495]

- [45]. Li X, Zhao X, Fang Y, Jiang X, Duong T, Fan C, Huang C, Kain SR, Journal of Biological Chemistry 1998, 273, 34970. [PubMed: 9857028]
- [46]. Cockerill PN, FEBS J 2011, 278, 2182. [PubMed: 21501387]
- [47]. Shlyueva D, Stampfel G, Stark A, Nat Rev Genet 2014, 15, 272. [PubMed: 24614317]
- [48]. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K, Cell Res 2012, 22, 490. [PubMed: 22270183]
- [49]. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA, Cell 2013, 153, 307. [PubMed: 23582322]
- [50]. Fahrner JA, Baylin SB, Genes Dev 2003, 17, 1805. [PubMed: 12897049]
- [51]. Ghirlando R, Giles K, Gowher H, Xiao T, Xu Z, Yao H, Felsenfeld G, Biochim Biophys Acta 2012, 1819, 644. [PubMed: 22326678]
- [52]. Dickson AJ, in Cell Line Development, (Ed: Al-Rubeai M), Springer Netherlands, Dordrecht 2009, Ch. 4.
- [53]. Wang X, Kawabe Y, Hada T, Ito A, Kamihira M, Biotechnol J 2018, e1800063. [PubMed: 29701326]
- [54]. Mayrhofer P, Kratzer B, Sommeregger W, Steinfellner W, Reinhart D, Mader A, Turan S, Qiao J, Bode J, Kunert R, Appl Microbiol Biotechnol 2014, 98, 9723. [PubMed: 25158835]
- [55]. Barnes LM, Moy N, Dickson AJ, Biotechnol Bioeng 2006, 94, 530. [PubMed: 16496399]
- [56]. Davies SL, Lovelady CS, Grainger RK, Racher AJ, Young RJ, James DC, Biotechnol Bioeng 2013, 110, 260. [PubMed: 22833427]
- [57]. Seth G, Charaniya S, Wlaschin KF, Hu WS, Curr Opin Biotechnol 2007, 18, 557. [PubMed: 18082393]
- [58]. Kallehauge TB, Li S, Pedersen LE, Ha TK, Ley D, Andersen MR, Kildegaard HF, Lee GM, Lewis NE, Scientific Reports 2017, 7, 40388. [PubMed: 28091612]

Figure 1.

Transgene integration into active regions with open chromatin. **(A)** Gene features at integration sites found for the different populations of GFP⁺ cells (top 1% pool, bottom 10% pool, and top 1% single copy clones) and random locations. **(B) & (C)** Probability density of log₂[RPKM] for RNA-Seq **(B)** or ATAC-Seq **(C)** data across a 100 kbp window centered on integration sites found for different cell populations. (Blue: top 1% single copy clones, N = 34 integration sites; Orange: top 1% pool, N = 272 integration sites; Green: bottom 10% pool, N = 254 integration sites; Purple: 10,000 random locations). Points (▽) on the baseline denote Log₂ [RPKM] values for three single copy IgG clones. **(D)** qRT-PCR data for transcript level of GFP among top 1% single copy GFP producing clones, top 1% GFP pool, and bottom 10% GFP pool, as measured by log₂ fold change relative to GAPDH (***: p < 0.001).

Figure 2.

Single Copy IgG Cell Lines. **(A)** IgG titer from of top 52 IgG producing clones as measured by ELISA. Supernatant was collected from adherent culture of cells in 6 well plates after five days. Clones were selected based on ELISA data at the 96-well stage of single cell cloning. Color of bar corresponds to number of copies of lentivector integrated as determined by integration site analysis. **(B)** Transcript level of IgG heavy chain and light chain for top 52 clones measured using 1-step qRT-PCR. Results reported in terms of log₂ fold change relative to GAPDH.

Figure 3.

Characterization of single copy IgG producing cell lines. **(A)** Gel-electrophoresis of secondary PCR products from integration site analysis. Number of bands corresponds to number of integration sites. **(B)** qRT-PCR data measuring IgG light chain (LC) and IgG heavy chain (HC) transcript level in three single copy clones with different integration sites. Measured in terms of log₂ fold change relative to GAPDH. **(C)** Specific productivity of single copy clones assayed in adherent culture as measured by ELISA. Supernatant was collected on eighth day of culture, two days after media replacement. N = 2 wells per cell line. **(D-F)** ATAC-Seq and RNA-seq pile up from CHO-K1 for 100kb window surrounding integration sites of three single copy clones (**D**: 2C10, **E**: 3A10, **F**: 4B1) as visualized in IGV. ATAC-seq data is normalized to Reads Per Hundred Million (RPHM) and RNA-seq data is normalized to Reads Per Million (RPM). Vertical black bar denotes integration site.

Figure 4.

Replacement of the IgG Product Gene using Dual RMCE. **(A)** Schematic of dual RMCE process. Donor vector is promoter-less, and contains $Lox511$ and minimal-FRT sites that match those on the genome. F2AC is the recombinase expression vector containing flippase and cre recombinase. **(B)** Growth curve for EPO and TNFR producing clones obtained from dual RMCE. Cells were grown in adherent culture for seven days, and media was replaced on day four. **(C)** qRT-PCR data measuring transcript level of EPO and TNFR, as denoted by $\log_2$ fold change relative to GAPDH. **(D)** EPO and TNFR specific productivity as measured by ELISA. Supernatant was collected on day seven of the growth curve from **(B)**, three days after media replacement. $N = 3$ wells per cell line.

Table 1.

Kolmogorov-Smirnov test for comparison between RNA-seq RPKM distributions and Mann Whitney Test for the comparison of ATAC-Seq RPKM distributions in the 100 kbp window surrounding the integration sites.

RNA-seq RPKM Distribution Comparison	K-S Test P-Value
Top 1% Pool > Bottom 10% Pool ***	1.90×10^{-12}
Top 1% Clones > Bottom 10% Pool ***	7.03×10^{-11}
Top 1% Clones Top 1% Pool **	4.13×10^{-03}
Top 1% Pool > Random Locations ***	2.42×10^{-18}
Bottom 10% Pool Random Locations	0.093
Top 1% Clones > Random Locations ***	3.65×10^{-10}
ATAC-seq RPKM Distribution Comparison	Mann-Whitney Test P-Value
Top 1% Pool > Bottom 10% Pool ***	2.57×10^{-08}
Top 1% Clones > Bottom 10% Pool ***	1.76×10^{-04}
Top 1% Clones Top 1% Pool	0.509
Top 1% Pool > Random Locations ***	8.88×10^{-08}
Bottom 10% Pool Random Locations **	7.75×10^{-03}
Top 1% Clones > Random Locations **	2.38×10^{-03}

p < 0.05 denotes a significant difference between the distributions.

*
: p < 0.05

**
: p < 0.01

: p < 0.001