Gene position affects protein expression in a 2A polycistronic vector system

Qian Lab, Department of Pathology
Joseph Wall, Ziqing Liu, Li Qian

ABSTRACT

Cellular reprogramming of fibroblasts to cardiomyocyte-like cells is achieved by the introduction of a combination of cardiac transcription factors. Successful reprogramming requires a stoichiometric balance of these transcription factors at the protein level. Traditional methods to introduce exogenous gene expression required multiple plasmids to be transfected into a cell for reprogramming. A problem is that cells hardly take up these reprogramming factor plasmids at optimal stoichiometric ratios, which decreases reprogramming efficiency. Polycistronic vectors are attractive thanks to the ability to introduce one plasmid with desired genes for expression into a cell, as opposed to multiple plasmids. The challenge of polycistronic vectors is that eukaryotic cells do not have the cellular machinery to process polycistronic proteins during translation. However, viral oligopeptides (called 2A peptides) have been identified that enable ribosomes to skip the formation of a certain peptide bond, allowing for “cleavage” of polycistronic peptides. Despite this advantage, widespread adoption of 2A polycistronic vectors has been limited due to the lack of systematic comparison of the different 2A sequences in a polycistronic system.

This project sought to characterize the effect of gene position on protein expression in tricistronic 2A constructs containing fluorescent protein (FP) genes in varying positions. To investigate a 2A polycistronic system, mouse embryonic fibroblasts (MEF-T) were infected with retroviruses carrying the polycistronic vector. These transduced cells were then analyzed via fluorescent microscopy and flow cytometry in order to quantify the expression of each fluorescent protein. It was determined that the gene in the first position achieved the highest level of expression, followed by the third position, with the second position yielding the lowest protein
expression. Such a finding is important because it not only shows that gene order matters, but it will potentially help further the utilization of 2A polycistronic systems for cellular reprogramming.

INTRODUCTION

Biological processes, such as growth and development, require a suite of proteins working together to accomplish a task, be it cell signaling or transcriptional regulation. A balanced expression of each protein is vital to the function and success of the cell. Protein expression is regulated by the transcription and translation of genes, therefore, an understanding of co-expression of genes is necessary to unlock the mechanisms behind numerous biological processes. Cellular reprogramming is a rapidly growing field of research that has made strides in understanding this importance.

It has been shown that differentiated cells, such as adult mouse fibroblasts, have the ability to be reverted back to a pluripotent state by introducing a defined set of transcription factors. These factors must be expressed in a manner that allows a change in the expression of genes already present in the DNA of the cell. The newly expressed genes will then alter the cellular machinery and organization of the cell to induce pluripotency.

Since the discovery of induced pluripotent stem cells (iPSCs) in 2006, the field of direct reprogramming has experienced rapid development. Through direct reprogramming, a cell can be converted to another cell type without reverting to a stem cell state. This process allows increased reprogramming efficiency, with decreased heterogeneity of the target population. Direct reprogramming has also redefined the understanding of the plasticity of somatic cells, which can be influenced by not only transcription factors, but also epigenetic factors and the cellular microenvironment.

Direct cardiac reprogramming is of particular interest to the Qian lab. Our goal is to understand the molecular basis of cardiomyocyte specification and maturation, which will then
allow the clinical application of cellular reprogramming in heart disease. Three factors, Gata4 (G), Mef2c (M), and Tbx5 (T), were found to influence the specification and development of healthy cardiomyocytes based on their temporal and dosage expression. Our lab has shown that the induction of cardiac fibroblasts (CF) to induced cardiomyocytes (iCMs) depends on a stoichiometric balance of the reprogramming factors, G, M, and T.

Traditional methods of iCM generation require transducing fibroblasts with pooled viruses encoding G, M, or T. An obstacle of this approach is uncontrollable ratios of G, M, and T expression among transduced fibroblasts due to the possibility of heterogeneous populations that may or may not be expressing all three factors. Polycistronic vectors are attractive because they allow the introduction of a single plasmid carrying all three genes. Research into polycistronic vectors in the area of cardiac reprogramming has been attempted in the form of placing multiple genes after a single promoter. However, this method resulted in reprogramming efficiency that was marginally better than traditional methods.

One challenge of polycistronic vectors is that eukaryotic cells do not possess the cellular machinery necessary to cleave polypeptides from a polycistronic vector. The answer to this problem came in the form of “self-cleaving” peptides found in viruses. These 18-22 amino-acid-long viral oligopeptides, known as 2A peptides, mediate “cleavage” of polypeptides during translation in eukaryotic cells. The name “2A” refers to the region of the viral genome that was found to produce self-cleavage. The first 2A peptide was found in foot-and-mouth disease (F2A), and later E2A (equine rhinitis A virus), P2A (porcine teschovirus-1), and T2A (thosea asigna virus) were identified.

The mechanism behind 2A peptides is a ribosomal skipping action along a glycyl-prolyl bond at the C-terminus of the 2A peptide (Figure 1a). This means that when the ribosome encounters this specific sequence, the formation of the peptide bond does not occur and the ribosome moves to the next codon. This sequence is highly conserved among different 2A
peptides at the C-terminus (Figure 1b) and is necessary for the steric hindrance that causes the ribosome to skip. Due to the relatively high efficiency of 2A-mediated ribosome skipping and the short length of 2A sequences, 2A peptides lead to higher protein expression of the gene downstream of 2A as compared to other strategies for co-expression of multiple genes.

Widespread use of 2A systems is lacking due to the need for systematic comparisons between 2A sequences in polycistronic vectors. Previous studies have focused on the cleaving efficiency of 2A peptides. Some have shown that T2A is most efficient, while others suggest the better 2A peptide is P2A. The effect of 2A peptides on the ratio of protein expression is also not fully known. As a result, a 2A cloning vector system for gene expression in cellular reprogramming has not been broadly adopted.

This study was a component of a larger research project to systematically compare 2A peptides for cloning multiple genes in a bicistronic, tricistronic, or quadcistronic vector. The focus of this study was determining the effect of gene position on protein expression in a tricistronic 2A construct. Since there is a need for a balance of proteins in order to produce iCMs, we wanted to see if the gene order affected protein expression. For this experiment, we used three distinct fluorescent protein genes in place of the transcription factors. The tricistronic vector was introduced into a mouse embryonic fibroblast cell line and the protein expression was quantified through fluorescence microscopy and flow cytometry.

It was hypothesized that there would be a decrease in protein expression with gene position with the first position expressing the most, with the second less, and the third position having the least expression. This hypothesis is based on the tendency of ribosomes to drop-off as a transcript increases in length. After characterizing the effect gene position has on protein expression, the FPs can be substituted for GMT to quantify the effect on iCM generation. Overall, it is believed that an understanding of the positional effect in a polycistronic vector would help
advance the study of cellular reprogramming and would allow for advances in regenerative therapies.

METHODS AND MATERIALS

Cells
MEF-T is a cell line selected from SV40 large-T antigen-transformed MEF and was generated recently by our lab as described. MEF-T (Figure 2) was infected with the retroviruses carrying the polycistronic vectors. Cells were maintained in growth media: DMEM containing 10% Fetal Bovine Serum (FBS) and 50 μg ml⁻¹ penicillin/streptomycin (Sigma).

Plasmids
Cloning of all constructs was performed using the pGEM-T easy vector (Promega) as an intermediate. First, a 300 base pair (bp) DNA fragment containing desired restriction endonuclease sites and 2A-encoding sequences were synthesized (Genewiz) (Appendix Table 1). A 6 bp stuffer sequence was placed between each pair of endonuclease sites for gene insertion. Next, genes of interest were PCR-amplified and inserted into the cloning intermediates one by one. The templates used for PCR were: pLenti-GFP (Cell Biolabs, LTV-400) for GFP, pCSCMV-tdTomato (Addgene) for Td, and piRFP670-N1 (Addgene, #45457) for iRFP670. All tricistronic constructs were also cloned in the intermediate plasmid pGEM-T-PTE2A, and the restriction sites used were: BamHI and Nhel for the first gene, Spel and HindIII for the second gene, and Xhol and Sall for the third gene. A 6 bp kozak sequence ACCGCC was added right before the ATG start codon of the first gene in every construct and the stop codon TAA was added at the end of the last gene in every construct. Lastly, the constructs were excised from pGEM-T and inserted into the pMXs retroviral vector (Cell Biolabs) with BamHI and Sall.
Retroviral packaging

A retroviral system was utilized to package the tricistronic vector, which would then be introduced into fibroblast cells. Platinum-E (Plat-E, Cell Biolabs) cells, a potent retrovirus packaging cell line was used. Plat-E cells were maintained in growth media supplemented with 1 μg ml⁻¹ puromycin (Sigma), and 100 μg ml⁻¹ of blasticidin S (Life Technologies).

The day before transfection, 4-5x10⁶ cells were seeded onto 10 cm dishes in growth media without puromycin and blasticidin. The next day, pMXs-based retroviral vectors were introduced into Plat-E cells using Nanofect (ALSTEM). Generally, 20 μg of plasmid DNA was combined with 500 μL plain DMEM, then 45 μL of Nanofect reagent was combined with 500 μL plain DMEM. The Nanofect suspension was added into the DNA suspension and the mixture was vortexed for 15 seconds and incubated at room temperature for 15 minutes. The mixture was then added dropwise to the Plat-E cells. Culture media was changed to fresh media right before transfection and all reagents used were warmed to room temperature before mixing.

Transfected cells were incubated overnight at 37°C with 5% CO₂. Medium was changed the next day and the virus-containing supernatant was collected 48 hours after transfection, filtered through a 0.45 μm cellulose acetate filter (Thermo Scientific) and incubated with PEG8000 (Sigma, 4 volumes of supernatant and 1 volume of 40% PEG8000/PBS) overnight at 4°C.

MEF-T Infection with Retrovirus

Target cells were plated the same day as virus collection on 0.1% gelatin-coated 24 well-plates. The next day, the viruses were pelleted by centrifuge at 3500 rpm, 4°C for 30 minutes, re-suspended with growth media or fibroblast media supplemented with 4 μg ml⁻¹ polybrene (Life Technologies), and added to target cells immediately. Twenty-four to forty-eight hours post-infection, the virus-containing medium was replaced with growth media or fibroblast media.
Live Imaging and Flow Cytometry

At day three post-transduction, cells were washed with PBS and live cell images were taken at 20x using EVOS® FL Auto Cell Imaging System (Life Technologies). Cells were collected for flow cytometry by first dissociating with 0.05% trypsin/EDTA (Life Technologies) for 5 minutes at 37°C, then the trypsin was neutralized with growth media. The cells were pelleted, resuspended with FACS buffer (PBS supplemented with 2% FBS and 2 mM EDTA), and analyzed by BD Accuri™ C6 flow cytometer or LSRII (BD Bioscience).

Data Collection

After three days post-transduction, live-cell images of the triplicate MEF-T samples were taken to visualize the FP expression within each cell. These cells were then collected in order to run them through a flow cytometer to obtain quantitative data on the number of cells present and the median fluorescence intensity of each fluorescent protein. These quantities would allow further analysis of the positional effect of the gene order on protein expression in a polycistronic system.

Statistical Analysis

The median fluorescence intensity (dMFI) was normalized to that of GFP-Td-i670 for GFP, Td-i670-GFP for Td, and i670-Td-GFP for i670. Mean ± SEM of multiple experiments was calculated. Significant differences were determined from one-way ANOVA tests (α = 0.05) and bonferroni corrections.

RESULTS
Tricistronic constructs

In order to determine relative protein expression at different positions along a tricistronic construct, a set of vectors containing fluorophores - green fluorescent protein (GFP), Td tomato (Td), and far red fluorescent protein iRFP670 (i670) - was generated, each varying all possible fluorophore configurations with P2A and T2A sequences inserted between each gene (Figure 3). The three selected FPs represent a wide range of brightness with Td as the brightest FP available whose brightness triples GFP. i670 is a relatively dim FP with a brightness that is one-third of GFP’s. These FPs are useful because they can be tracked and sorted through flow cytometry, making them ideal for the analysis of differences in protein expression.

The tricistronic construct was then inserted into a retroviral vector backbone (pMX). Retroviruses are efficient tools for delivering genes due to their high titers for efficient temporary expression and their ability to incorporate into the genome for long-term expression. Another advantage is that retroviruses are able to transduce a variety of cell types that are proliferating.

Live Cell Imaging

Once the constructs were successfully cloned, they were transfected into Plat-E, a potent retrovirus packaging cell line. Conventional NIH-3T3 based retroviral packaging systems have limited stability and produce low viral yields, which results in lower protein expression of retroviral structural proteins in packaging cells. Plat-E is based on 293T and was generated using novel packaging constructs with an EF1-α promoter to ensure longer stability and high-yield retroviral structural protein expression.

The retrovirus was collected from the Plat-E and transduced into MEF-T cells. MEF-T is an immortalized cell line that was generated in the Qian lab from MEF cells. This immortalized cell line has the same morphology as MEF (Figure 2), however it is highly proliferative (Figure 2) and easily transduced, making it ideal for transgene expression and genetic manipulation.
fluorescent images of GFP, Td, and i670 expression were captured in MEF-T cells at three days post-transduction with retroviruses encoding the different 2A constructs (Figure 4). Images were taken at 20x and the exposure for each fluorescent channel was adjusted to the positive control FP. Once the exposure was set, the same setting was used for all of the constructs. The live images show a difference in protein expression with changing position as is evident by the differences in brightness of the individual fluorescent channels. By comparing the FP in the first position versus the second or third position, it can be seen that there is a decrease in the FP expression when not in the first position. The most dramatic visual decrease can be seen when i670 is in the first position versus the other positions.

**Flow Cytometry and dMFI**

After live cell images were taken, MEF-T cells were collected and analyzed by flow cytometry for fluorescence intensity of all three FPs (Figure 5). By measuring the median fluorescent intensity, a quantitative comparison of the level of protein expression of each FP could be made. The data suggests a significant difference between gene positions in the constructs. It was determined that the first position obtained the highest protein expression, followed by the third position, with the second position yielding the lowest protein expression (Figure 6). This result was not expected, based on our hypothesis that there would be a gradual decrease in protein expression from the first position to the third position.

**DISCUSSION**

Polycistronic vectors have gained increased use due to their convenience for introducing the necessary genes into cells for reprogramming. An issue that this study addressed is the positional effect of the genes in the tricistronic vector on protein expression. A key component of this project was utilizing 2A sequences in our vector system because they increase the cleavage
efficiency of polypeptides during translation in eukaryotic cells. By transducing MEF-T with the retroviral constructs, we were able to quantify and compare the protein expression observed with each change in gene order. Our comparison of the fluorescence intensities of GFP, Td, and i670 at various positions in a tricistronic construct revealed that protein expression was highest at the first position, less at the third position, and lowest at the second position.

Theoretically, a gradual decrease of translated protein along a polycistronic vector is expected due to naturally occurring ribosome drop-off at a constant rate, which is positively correlated with transcript length. However, it is possible that the higher level of protein expression at the third gene position, as opposed to the second position, resulted from the use of T2A immediately upstream of the third gene. In a bi-cistronic construct study, our lab found T2A to be more efficient as compared to P2A bi-cistronic constructs (SF 1).

Even more interesting was the effect position had on the level of expressed protein because it seemed to be negatively correlated with the brightness of the fluorescent protein. The brightest FP, Td, did not exhibit dramatic changes in fluorescence intensity when moved to different positions in the tricistronic construct, giving a range of 73%-116% compared to that at the first position (Figure 6b). One potential reason for this lack of change in fluorescence intensity is that Td is such a bright FP that a difference in intensity is difficult to quantify through flow cytometry. A dimmer FP, such as red fluorescent protein (RFP), could be used should further testing be needed in future polycistronic systems.

GFP showed approximately 30% decrease in brightness at the third position and approximately 70% decrease at the second position when compared to that at the first position (Figure 6a). i670 resulted in nearly 90% and 95% decreases at the third and second positions, respectively (Figure 6c).

Taken together, the different levels of protein expression observed at varied positions along the tricistronic constructs likely represents the combined effects of position along the
construct, efficiency of protein expression (mediated by different 2A sequences), and brightness of the FPs. These conclusions are important for the continued study of 2A polycistronic systems in order to understand the mechanism of direct cellular reprogramming.

Direct cellular reprogramming, as described earlier, is a technique that has been intensely studied in order to understand the mechanisms behind this fascinating process. It is the hope that reprogramming can be utilized in regenerative medicine in order to provide treatment for future patients, such as those who suffer from a heart attack. One of the key aspects of direct reprogramming is obtaining a proper balance in expression of factors that cause the cell to change morphology and function. Knowing how protein expression varies with gene position could enable future research into increasing the efficiency of direct cellular reprogramming.

Future directions in this project would be to utilize this polycistronic system in a reprogramming experiment to conclude if a 2A polycistronic vector is indeed more efficient at reprogramming. The FPs would be traded for the transcription factors Gata4, Mef2c, and Tbx5, which have been shown to directly reprogram fibroblasts into cardiomyocyte-like cells. With the knowledge of the positional effect on the different levels of protein expression, an optimal construct could be generated. A quad-cistronic 2A vector could also be generated in order to test if additional genes change the positional effect and protein expression in the system. The rationale behind this is that other reprogramming systems may involve more than three transcription factors in order to induce efficient and quality reprogramming.
ACKNOWLEDGEMENTS

Thank you to the members of the Qian lab who trained and helped me complete my honors thesis. It has been a pleasure conducting research alongside you all. Special thanks to Dr. Ziqing Liu who took me on as an undergraduate research assistant and trained me for the polycistronic project. I would not have been able to accomplish nearly as much or have grown as much in my scientific training without your guidance. Thank you to Dr. Amy Maddox and my writing group whose helpful discussions and critical reviews helped polish this thesis. I would also like to thank my family and friends who encouraged me during my undergraduate experience at Carolina. Your support means the world to me, and I hope to make you proud as I continue to pursue my dreams.

REFERENCES

FIGURES

Figure 1. a) 2A mechanism showing ribosomal skipping action that occurs from the lack of a peptide bond forming at the glycyl-prolyl sequence. b) 2A codon and amino acid sequences for P2A and T2A peptides that were utilized in the tricistronic vectors.

Figure 2. MEF vs. MEF-T morphology and proliferation. This immortalized cell line was generated in the Qian lab from MEF. The cell culture images (scale bar 200 µM) show that the morphology of the cell lines is the same. The proliferation rate of MEF-T is shown to be significantly faster than MEF.
**Figure 3.** The six retroviral (pMx) tricistronic constructs containing three FP genes (GFP, Td, and i670) in different positions with P2A-T2A sequences between each gene.
Figure 4. Live cell images of transduced cells with the 2A constructs show positional effect in protein expression. Left side legend shows which construct the row represents. The top legend shows which fluorescent channel the image represents. Images taken at 20x, scale bar is 200 µM.

![Image of live cell images showing positional effect in protein expression.]

Figure 5. Flow cytometry dMFI analysis of transduced MEF-T cells shows significant difference in protein expression with gene position within each tricistronic construct. GFP (green bar), Td (red bar), and i670 (purple bar). Table summarizes the significant differences from one-way ANOVA and bonferroni correction ($\alpha = 0.05$; $n = 24$).

![Image of flow cytometry dMFI analysis showing significant differences in protein expression.]

<table>
<thead>
<tr>
<th>Position</th>
<th>1 vs 2</th>
<th>1 vs 3</th>
<th>2 vs 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFP</td>
<td>*<strong>/</strong> */**/ns/ns</td>
<td>*<strong>/</strong>/ns/ns</td>
<td>*<strong>/</strong>/ns/ns</td>
</tr>
<tr>
<td>Td</td>
<td>*/**/*ns/ns</td>
<td>ns/ns/<strong>/</strong></td>
<td><em><strong>/</strong>/<strong>/</strong></em></td>
</tr>
<tr>
<td>i670</td>
<td>*<strong>/</strong>/*ns/ns</td>
<td><em><strong>/</strong>/<strong>/</strong></em></td>
<td>*<strong>/</strong>/*ns/ns</td>
</tr>
</tbody>
</table>
Figure 6. Comparison of dMFI from the varying positions of the FPs showed a significant difference for (a) GFP position 1 vs. 2 and 2 vs. 3, (c) i670 position 1 vs. 2, 1 vs. 3, and 2 vs. 3. (b) Td did not have a significant difference in fluorescence intensity. Significance was determined by one-way ANOVA ($\alpha = 0.05$; $n = 24$) and bonferroni corrections.