Published in final edited form as:

Protein Expr Purif. 2011 November ; 80(1): 145–151. doi:10.1016/j.pep.2011.06.017.

Solubility Partner IF2 Domain I Enables High Yield Synthesis of Transducible Transcription Factors in *E. coli*

William C. Yang¹, John P. Welsh², Jieun Lee³, John P. Cooke³, and James Swartz^{1,2}

¹Department of Bioengineering, Stanford University, Stanford, CA 94305

²Department of Chemical Engineering, Stanford University, Stanford, CA 94305

³Division of Cardiovascular Medicine, Stanford University, Stanford, CA 94305

Abstract

Since the discovery that somatic cells could be reprogrammed back to a pluripotent state through the viral expression of a certain set of transcription factors, there has been great interest in reprogramming using a safer and more clinically relevant protein-based approach. However, the search for an efficient reprogramming approach utilizing the transcription factors in protein form requires a significant amount of protein material. Milligram quantities of transcription factors are challenging to obtain due to low yields and poor solubility. In this work, we describe enhanced production of the pluripotency transcription factors Oct4, Sox2, Klf4, Nanog, and Lin28 after fusing them to a solubility partner, IF2 Domain I (IF2D1). We expressed and purified milligram quantities of the fusion proteins. Though the transcription factor passenger proteins became insoluble after removal of the IF2D1, the un-cleaved Oct4, Sox2, Klf4, Nanog fusion proteins exhibited specific binding to their consensus DNA sequences. However, when we administered the un-cleaved IF2D1-Oct4-R9 and IF2D1-Sox2-R9 to fibroblasts and measured their ability to influence transcriptional activity, we found that they were not fully bioactive; IF2D1-Oct4-R9 and IF2D1-Sox2-R9 influenced only a subset of their downstream gene targets. Thus, while the IF2D1 solubility partner enabled soluble production of the fusion protein at high levels, it did not yield fully bioactive transcription factors.

Introduction

Induced pluripotent stem cells (iPSCs) were first generated from somatic cells by ectopic expression of a set of transcription factors (TFs) using viral constructs [1, 2]. These virus-generated iPSCs provided hope for realizing patient-specific cell therapies. However, because viruses integrate foreign DNA into the host cell genome, the resultant cells were not suitable for clinical applications. Given that viral integration can lead to gene silencing and tumorigenesis [3], these initial discoveries sparked a search for safer, non-integrating methods of iPSC generation. Since then, adenoviruses [4], plasmids [5], and transposons [6, 7] have been used to generate iPSCs. Though adenoviruses, plasmids, and transposons reduce the risk of insertional mutagenesis, they can still potentially leave behind undesired genetic material.

© 2011 Elsevier Inc. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Arguably, the safest method of iPSC generation is direct administration of the TFs in protein form. Protein-based generation of iPSCs has also been successfully demonstrated: mouse iPSCs were generated using purified, refolded mouse TFs with polyarginine (R9) protein transduction domains (PTDs) that enable cellular entry [8] while human iPSCs were generated using crude human cell lysate containing human TFs that also included R9 PTDs [9, 10]. Arginine-rich sequences such as the R9 give the fusion proteins a positively charged moiety for binding to the negatively charged cell surface. Bound proteins are subsequently taken into the cell through endocytosis [11]. However, the efficiency of protein-based reprogramming was less than 0.001%, orders of magnitude lower than that of virus-based reprogramming. Furthermore, human iPSCs have not yet been generated using purified recombinant transducible TFs. Thus, there is a need for more robust and efficient methods of protein-based nuclear reprogramming.

The search for optimal reprogramming conditions will likely require large amounts of purified recombinant TF reagents. The most efficient way to make recombinant protein is *in vivo* expression in bacterial hosts. *E. coli* are easy to genetically manipulate, grow rapidly, and have simple nutrient requirements. Thus, recombinant TFs such as Oct4 and Nanog [12], HIV-TAT fusions of Oct4 and Sox2 [13], and polyarginine fusions of the Yamanaka factors (Oct4, Sox2, cMyc, and Klf4) [14] have been expressed *in vivo* using *E. coli*, but yields were not reported. The mouse iPSCs were generated using refolded proteins [8], indicating that soluble expression of these TFs is problematic.

We have also had difficulty expressing soluble TFs *in vivo* using *E. coli* in our own laboratory. As a result, we have used *E. coli*-based cell-free protein synthesis (CFPS) to produce these TFs [15]. To perform CFPS, *E. coli* are lysed at high pressures to extract both protein production machinery as well as inner membrane vesicles for energy regeneration. This extract is incubated with template DNA encoding the protein of interest in a chemical environment that mimics the *E. coli* cytoplasm [16]. CFPS offers certain advantages for producing difficult-to-make proteins. Decoupling protein synthesis from maintenance of host cell health permits production of toxic proteins and its open nature permits easy perturbation of protein production conditions, such as the addition of cofactors to improve folding.

Though we made microgram-to-milligram quantities of soluble TFs using CFPS [15], we can potentially produce soluble TFs using *in vivo* expression in *E. coli* more cost-effectively. This is because *in vivo* protein production is a one-step process in which proteins are produced during a single cell culture while CFPS is a multi-step process. However, we would need expression conditions that generate high soluble yields.

One popular method for boosting soluble production of recombinant protein *in vivo* is to fuse a solubility partner to the protein of interest. Solubility partners such as glutathione s-transferase (GST) and maltose-binding protein (MBP) have been fused to transcription factors and improved soluble yields to varying degrees [17]. The effect of solubility partners is often protein specific [18]. Thus, we set out to explore the effect of using a solubility partner to enhance the soluble *in vivo* production of the transducible pluripotency TFs.

We chose Domain I of the *E. coli* translation initiation factor IF2 (IF2D1) as our soluble partner. IF2 is an *E. coli* translation initiation factor that binds to the 30S ribosome. However, Domain I (residues 1-157, approximately 20 kDa) of IF2 itself does not bind the ribosome, nor does it have any catalytic function [19]. IF2D1 forms a well-ordered and compact globular domain which is connected to the desired products by a flexible linker. When used as a fusion partner, IF2D1 has been shown to boost the soluble production of active streptavidin [20]. Furthermore, its 20 kDa size is smaller than that of other solubility

partners such as the 30 kDa GST and the 40 kDa MBP. For these reasons, we chose to investigate IF2D1 as a TF solubility partner.

In this work, we expressed the set of polyarginine-tagged (R9) pluripotency transcription factors (Oct4, Sox2, Nanog, Lin28, cMyc, Klf4) using the IF2D1 solubility partner. For all the TFs except cMyc, IF2D1 improved both total and soluble TF fusion protein expression. We then purified the fusion proteins using cation exchange chromatography and attempted to remove the IF2D1 using tobacco etch virus (TEV) protease cleavage. However, the passenger TFs became insoluble following cleavage in spite of exploring cleavage in a number of buffer conditions. Therefore, we proceeded to evaluate the un-cleaved proteins. First, we performed DNA binding assays with un-cleaved IF2D1-Oct4-R9, IF2D1-Sox2-R9, IF2D1-Klf4-R9, and IF2D1-Nanog-R9. These un-cleaved fusion proteins exhibited specific binding to their consensus DNA sequences, which prompted us to test their bioactivity. We administered IF2D1-Oct4-R9 and IF2D1-Sox2-R9 to human fibroblasts and assayed the ability of the un-cleaved fusion proteins to influence the expression of known downstream target genes. Our gene expression results showed that the un-cleaved Oct4 fusion and un-cleaved Sox2 fusion were not fully bioactive; specifically, the proteins did not fully recapitulate the expression induced by the positive control (the retroviral construct encoding the respective TF). Though the IF2D1 solubility partner greatly enhanced soluble production of the IF2D1-TF-R9 fusion protein, it did not yield fully bioactive protein.

Materials and Methods

Expression vectors

All genes except for IF2D1-Oct4-R9 were cloned into the pET24a vector (Novagen, San Diego, CA). IF2D1-Oct4-R9 was cloned in pY71 under control of the T7 promoter. pY71 is a 1.7 kb reduced size plasmid with a pUC19 origin of replication and a kanamycin resistance element [21]. The genes for Oct4, Sox2, Nanog, Lin28, cMyc, and Klf4 had been optimized for *E. coli* expression using DNAworks software [22] and were PCR amplified from previous constructs [15]. Each TF construct contained an artificial nuclear localization sequence (NLS) and a nona-arginine protein transduction domain (R9) at the C-terminus. TF-R9 constructs without the IF2D1 solubility partner (Fig. 1A) also contained a hexahistidine (HIS6) sequence at the C-terminus. For the IF2D1-TF-R9 constructs (Fig. 1B), the IF2D1 and TF domains were separated by a TEV protease cleavage sequence (ENLYFQS). The TEV cleavage site was flanked by a GGGGS linker sequence on either side. The HIS6 sequence was moved to the N-terminus for the IF2D1-TF-R9 constructs to facilitate removal of the IF2D1 domain via affinity immobilization following TEV protease cleavage. All constructs contained an NdeI and SalI restriction enzyme site at the N- and C-termini, respectively, for plasmid insertion. In addition, the TF genes contained BamHI and NheI restriction sites at the N- and C-termini for ease of removal and insertion of the different TF genes into either the TF-R9 or IF2D1-TF-R9 constructs.

Protein production and purification

Plasmids were transformed into *E. coli* BL21(DE3) pLysS cells and grown at 37 °C in 1 L LB media supplemented with 40 µg/mL kanamycin and 20 µg/mL chloramphenicol. At an OD₅₉₅ of 0.6, the cells were induced with 200 µM IPTG and incubated for 6-7 h at 25 °C. The cells were harvested by centrifugation and resuspended in 8 mL of 50 mM Bicine pH 8, 50 mM NaCl (load buffer) per g wet cell weight. A single pass through an Avestin EmulsiFlex C-50 high-pressure homogenizer at 20,000 psig lysed the cells. Before lysis, 1 tablet of Roche complete protease inhibitor cocktail (Roche Molecular Biochemicals, Indianapolis, IN) was added to the cell resuspension. Following homogenization, the lysates

were centrifuged at $30,000 \times g$ for 30 min and 500 U of DNaseI (Invitrogen, Carlsbad, CA) was added to the supernatant.

We initially attempted to purify the IF2D1-TF-R9s (excluding IF2D1-cMyc-R9) by adsorbing the N-terminal His₆-tag using Ni-NTA affinity resin. We were able to purify full-length and soluble IF2D1-TF-R9 fusions using the Ni-NTA resin, but the capture yield was low. Thus, we sought to exploit the largely positive charge of the proteins resulting from the R9 and nuclear localization sequences by using cation exchange chromatography.

First, we performed small scale analytical purifications to explore load and elution conditions; we then used the information from the analytical purifications to scale up protein production. We lysed *E. coli* cells using an Avestin C50 homogenizer (Avestin, Ottawa, ON, Canada) at 20,000 psi in 50 mM Sodium Acetate, 50 mM NaCl, pH 5.0 or 50 mM Bicine, 50 mM NaCl, pH8.0 using a lysis buffer to cell mass ratio of 8 mL per 1g of cells. We then captured the IF2D1-TF-R9 lysates on 2 mL SP-Sepharose columns and washed with increasing NaCl salt concentration in 100 mM steps. From the small scale purifications, we observed that a 1:1 lysate volume to column volume ratio ensured that the load did not exceed column capacity. Loading in the Bicine buffer provided the most complete binding and the IF2D1-TF-R9s eluted between 800 mM and 1200 mM NaCl concentrations with reasonable purity (Fig. S1).

After determining the best purification conditions using the small scale purifications, we then scaled up purification using an AKTApurifier FPLC system (GE Healthcare, Waukesha, WI). Lysates were loaded at 0.5 mL/min on a 70 mL SP-Sepharose column (GE Healthcare) equilibrated with load buffer. The column was kept on ice throughout the purification to reduce product degradation. The column was washed with 2 column volumes of load buffer with 600 mM NaCl at 1.2 mL/min. Protein was eluted in 1.5 column volumes of load buffer with 1.2 M NaCl at 1.2 mL/min in 12 mL fractions. Absorbance at 280 nm was used to determine which fractions to pool. Final purified products were formulated in 50 mM Bicine pH 8, 800 mM NaCl, and 20 % sucrose by diluting the elutions with a 60% sucrose solution. Protein concentrations were measured using Qubit Quantitation (Invitrogen) and purity was estimated using gel densitometry analysis of an SDS PAGE with ImageJ software [23].

Competitive analysis of DNA binding

Single-stranded DNA nucleotides were purchased from Integrated DNA Technologies (IDT, Coralville, IA). The DNA probes were designed with three copies of the consensus sequence per probe as using concatameric sequences has been shown to improve DNA binding [24]. The Oct4 cognate DNA probe containing the Oct4 consensus sequence AGATGCAT was based on the POU-binding site from Conserved Region 4 (CR4) in POU5f1 [25]: CAGAGAGATGCATGTGAGATGCATGTGAGATGCATCG. The Sox2 cognate DNA probe containing the Sox2 consensus sequence GACAAAG was based on the HMG-binding site from CR4 in POU5F1 [25]: CAGAGGACAAAGGTGGACAAAGGTGGACAAAGCG. The Nanog cognate DNA probe containing the Nanog consensus sequence TAATGG was based on the TCF3 enhancer [26]: CAGATAATGGGTGTAATGGGTGTAATGGCG. The Klf4 cognate DNA probe containing the Klf4 motif CCCACCC was based on the Nanog enhancer [27]: CAGAGAACCCACCCATAAACCCACCCATAAACCCACCCCG. The fluorescently-labeled DNA probes were modified at the 5' end with 6-carboxyfluorescein (6-FAM). Non-fluorescently-labeled cognate competitor DNA and non-fluorescently-labeled nonsense DNA were also purchased from IDT. The forward and reverse complement oligonucleotide pairs were resuspended in 10 mM Tris-HCl, 50 mM NaCl, pH 8.0 at a

concentration of 100 μ M. Oligonucleotides were annealed by heating for 5 minutes at 95° C and allowing the tubes to cool to room temperature.

The DNA binding assay was performed using fluorescently-labeled cognate DNA and Ni-NTA agarose (QIAGEN, Valencia, CA) as previously described (ref). Briefly, 1 μ M of each IF2D1-TF-R9 protein was incubated with 50 μ L of Ni-NTA agarose and DNA binding was assessed using a variety of competitor DNA probes in Bind Buffer (BB: 20 mM HEPES-KOH pH 8.0, 50 mM KCl, 0.5 mM DTT, 0.05 mM EDTA, 1 mM MgCl₂, 5% glycerol, 0.05% Tween20). There were four test conditions: (1) fluorescent cognate, (2) cognate competitor, (3) nonsense competitor, and (4) no protein negative control. Conditions (1) and (4) contained 2 μ M of fluorescent cognate DNA, while conditions (2) and (3) contained 10 μ M of non-labeled cognate competitor DNA or non-labeled nonsense DNA, respectively in addition to 2 μ M of fluorescent cognate DNA. The no protein negative control served to confirm that DNA does not bind Ni-NTA agarose without the presence of the His₆-tagged protein.

Quantitative PCR analysis of downstream target gene expression

The human neonatal foreskin BJ fibroblast cell line (passage ~6) was cultured in DMEM with 10% FBS and 1% penicillin/streptomycin (pen-strep) antibiotics (Invitrogen) in a humidified 5% CO₂ incubator at 37° C. The respective cDNAs for Oct4, Sox2, and GFP were cloned into the retroviral pMX vector and separately transfected into 293FT cells using lipofectamine 2000 (Invitrogen). Viral supernatants were harvested 3 days later, concentrated, and used to infect human BJ fibroblasts in DMEM with 10% FBS and 1% pen/strep. At 80% confluency, BJ fibroblast cells were serum-starved using DMEM with 1% serum to cause G1 cell cycle arrest. The synchronized BJ fibroblasts were then treated with 200 nM of IF2D1-Oct4-R9 or IF2D1-Sox2-R9 fusion protein every 24 h for 120 h. Cellular RNA was extracted for real-time RT-PCR analysis at 0, 72, and 144 h.

Cultured BJ fibroblasts were collected using TrypLE EXPRESS (Invitrogen) and treated with TRIzol® (Invitrogen). Cellular RNA was purified using an RNeasy Mini Kit (QIAGEN) according to the manufacturer's recommendations. Purified RNA was then treated by DNase I (QIAGEN) to remove genomic DNA contamination. DNase I was inactivated by heating for 5 minutes at 65° C. First-strand cDNA synthesis was performed with 2 μ g total RNA for each sample in a total volume of 20 μ L. The reverse transcription reaction was performed with random primers and incubated at 25° C for 10 min followed by 42° C for 50 min. Real-time RT-PCR analysis of mRNA was performed using Gene Expression Assays with Taqman assay primers (Applied Biosystems, Foster City, CA). Analysis of 18S mRNA served as an internal control. The TaqMan assay IDs are as follows: JARID2B: Hs01004457_m1, ZIC2: Hs00600845, B-MYB: Hs00193527_m1, UTF1: Hs03005111_g1, POU5F1: Hs00742896_s1, TCF4: Hs00162613_m1, SSBP2: Hs01044459_m1, GAP43: Hs00967138_m1, EPHA1: Hs00178313_m1, 18S: Hs99999901_s1. All PCR reactions were performed in a total volume of 20 μ L containing diluted 2 $\times$ TaqMan Universal PCR Master Mix (Applied Biosystems) and 20 $\times$ Gene Expression Assay Mix and 40 ng cDNA. All assays were performed in duplicate and run on an 7300 ABI Real time PCR System using the following conditions: 50° C for 2 min, 95° C for 10 min, and 40 cycles of 95° C for 15 sec and 60° C for 1 min. Relative quantification of the amplified products was based upon Ct values.

Results and Discussion

IF2 Domain I Improves total and soluble protein expression

We first attempted to express the four Yamanaka factors (Oct4, Sox2, cMyc, and Klf4) as well as Lin28 and Nanog under control of the T7 promoter in *E. coli* using the pET24a vector. Each gene construct contained an artificial nuclear localization sequence (NLS), nona-arginine protein transduction domain (R9), and a hexa-histidine (HIS6) sequence at the C-terminus (Fig. 1A). However, only cMyc-R9 showed appreciable accumulation, and the product was insoluble (Fig. 2). Thus, we sought to use Domain 1 of the IF2 translation initiation factor (IF2D1) as a solubility partner at the N-terminus of these TFs in an attempt to improve soluble expression of the TF-R9s. A TEV protease cleavage sequence was also included between the IF2D1 and TF sequences (Fig. 1B). With the exception of IF2D1-cMyc-R9 which remained insoluble, adding the IF2D1 solubility partner at the N-terminus greatly improved the soluble accumulation of the TF fusion protein product (Fig. 2). Thus, the IF2D1 solubility partner provides a convenient means for production of soluble fusion TF-R9s.

Cation exchange chromatography

The cell lysates (typically 4 to 5 grams of cell mass) from 1 L Luria-Bertani (LB) media *E. coli* cultures expressing the IF2D1-TF-R9s were loaded on a 2.5 cm × 15 cm (approximately 70-mL) SP-Sepharose column to capture the proteins. Using the small scale step gradient NaCl purifications as guidelines, we washed the columns with 600 mM NaCl and eluted with 1200 mM NaCl (Fig. 3A, for example). While there were significant degradation bands using the scaled-up format, we estimated approximately 30% purity for the full-length proteins using gel densitometry (Fig. 3B). Using this strategy, we purified at least 2 mg of each of the full-length IF2D1-TF-R9s in soluble form (Table 1).

The increased degradation in the large scale preparation likely resulted from the increased time that the proteins spent on the SP-Sepharose column. The lysate flowed through the 2-mL column in approximately 10 minutes, while it flowed through the 70-mL in approximately 2 hours. Thus, the immobilized IF2D1-TF-R9s were exposed to proteases in the surrounding lysate for a longer period of time. Degradation to this extent was not observed in the small scale purifications, in which the IF2D1-TF-R9s spent very little time in contact with the whole cell lysate while on the column. The purified IF2D1-TF-R9s were formulated in 50 mM Bicine pH 8.0, 800 mM NaCl, and 20 % sucrose (w/v) by diluting the elutions with a 60% sucrose solution. The purified IF2D1-TF-R9s remained soluble through multiple freeze thaws, demonstrating the stability of the purified fusion proteins.

Removal of IF2 Domain I from IF2D1-TF-R9 fusion proteins results in insolubility of TF-R9 passenger

We attempted to remove the IF2D1 domain from the fusion proteins via TEV protease cleavage. However, for all of the fusion proteins tested, the TF passenger became insoluble following removal of the IF2D1 solubility partner (Fig. S2). This was not entirely unexpected given that insolubility of a passenger protein is often, but not always, observed upon removal of the solubility partner [28-30]. We explored a number of different conditions for TEV cleavage using IF2D1-Oct4-R9 as the model protein, including cleavage temperature (4°, 18°, 30°, 37° C) and incubation time (ranging from 30 minutes to overnight), pH buffers (acetate pH 5.0; phosphate pH 7.0; HEPES pH 7.2; bicine, pH 8.0), and urea partial denaturation and refolding. In the urea partial denaturation experiments, fusion protein was cleaved in low concentrations of urea ranging from 0 to 2 M to partially unfold the protein. We did not wish to unfold the entire protein because the DNA binding studies suggest a portion of the protein is correctly folded. We hypothesized that partial

denaturation may allow the incorrectly-folded regions of the proteins to unfold and refold while preserving the more stable, correctly-folded DNA binding domain.

However, in all cases we were unable to obtain soluble Oct4-R9 passenger after removal of the IF2D1 domain. The failure of the Oct4 to remain soluble following cleavage indicates that the protein is most likely not folded correctly. Indeed, it has been suggested that a solubility partner such as IF2D1 may simply act as a passive shield to protect incorrectly folded portions of the passenger protein from aggregation [28-30].

Un-cleaved IF2D1-TF-R9 fusion proteins bind specifically to their consensus sequences

Since removing the IF2D1 from the fusion protein rendered the TF-R9 insoluble in all the cleavage conditions we explored, we decided not to remove the IF2D1. We then determined if the un-cleaved IF2D1-TF-R9 could bind to its consensus DNA sequences. To do this, we used a competitive non-radioactive DNA binding assay that we developed [31]. It captures the His₆-tagged IF2D1-TF-R9 using Ni-NTA agarose and assesses DNA binding using fluorescently-labeled cognate DNA probes.

Using the assay, we found that the IF2D1 fusions of Oct4, Sox2, Klf4, and Nanog specifically bind to their consensus sequences (Fig. 4). The binding signal was high when the fusion proteins had been incubated with 2 μ M of their fluorescently-labeled cognate DNAs. Co-incubation with excess non-labeled cognate competitor DNA reduced the binding signal. Co-incubation of the fluorescently-labeled cognate DNA with excess non-labeled nonsense DNA had a negligible effect on the binding signal because the protein preferentially binds the cognate probe. The no protein negative control shows that the DNA does not bind to the Ni-NTA resin. To maintain consistency, we analyzed the binding specificity of the IF2D1-TF-R9s using the same binding conditions that we had previously used to assess the binding specificity of TFs without the IF2D1 or the R9 [15, 31]. Our IF2D1-TF-R9s results were comparable to those that we had previously obtained using the versions of the TFs without the IF2D1 or the R9, showing that the IF2D1 and R9 do not abrogate binding of cognate DNA [15, 31].

Though the competitive binding analysis shows that the binding is specific, the strength of the binding signal varies slightly from protein to protein. We hypothesize that the difference may be due to incorrect protein folding or to interference from the solubility partner. First, the proteins may be folded in such a way that the His₆-tag in some proteins may be more accessible than those in others. Indeed we have observed in this work that His₆-tag purification was less efficient than cation exchange. Hence, the fraction of TF fusion proteins retained by the Ni-NTA resin may differ from protein to protein. Likewise, the fusion TFs may be folded in such a way that not all the DNA binding domains may be active.

Un-cleaved IF2D1-Oct4-R9 and IF2D1-Sox2-R9 does not fully stimulate the expression of known downstream gene targets

After discovering that the un-cleaved IF2D1-TF-R9s exhibited DNA binding activity, we determined their bioactivity by measuring the effect of the fusion proteins on the expression of downstream target genes. We exposed synchronized human fibroblasts to daily treatments of 200 nM IF2D1-Oct4-R9 or IF2D1-Sox2-R9 and assessed the change in the mRNA levels of downstream target genes [32, 33] using quantitative real-time RT-PCR. We transfected fibroblasts with a retroviral construct encoding green fluorescent protein (GFP) as a negative control and a retroviral construct encoding Oct4 or Sox2 as a positive control. As expected, the viral GFP had little effect on gene expression, whereas the viral Oct4 or Sox2 did act on the target genes (Fig. 5). However, the IF2D1-Oct4-R9 only acted on SSBP2 and failed to

act on the rest (Fig. 5A-F). Similarly, the IF2D1-Sox2-R9 did not fully mimic the viral Sox2 as it induced JARID2, but not ZIC2 or b-MYB expression (Fig. 5G-I).

Our gene expression results suggest that the un-cleaved IF2D1-TF-R9s enter the cells, but do not exhibit full bioactivity. The IF2D1-Oct4-R9 downregulated SSBP2 while the IF2D1-Sox2-R9 elevated JARID2 expression compared to the negative control pMX-GFP, but not to the degree exerted by the positive control pMX-Sox2. The method of positive control retroviral induction was nearly identical to that used by the seminal retroviral iPSC generation study [1], so the retroviral vectors should have enabled ectopic expression of Oct4 and Sox2. While our partial bioactivity results suggest that IF2D1-Oct4-R9 and IF2D1-Sox2-R9 did enter the cell, the characterization of the cellular entry and bioactivity could be further strengthened by ectopically expressing Oct4 or Sox2 using exogenously delivered mRNA molecules encoding those proteins as another positive control. The technology was not available at the time these experiments were performed, but recent developments by the Rossi group at Harvard University have enabled researchers to exogenously deliver mRNA molecules to influence cellular gene expression [34].

Though the downregulation of the SSBP2 gene by IF2D1-Oct4-R9 and the upregulation of the JARID2 gene by IF2D1-Sox2-R9 suggest that IF2D1 fusion TFs entered the cells, they also suggest that either an insufficient amount of the IF2D1 fusion TFs was taken up by the cells or that the IF2D1 fusion TFs were not fully active. However, it is difficult to quantify intracellular protein delivery using current methods. Immunofluorescence provides only qualitative information. Radioactive tracing could provide quantitative information, but it is cumbersome and costly to produce radioactively labeled proteins in *E. coli*. We are currently developing a sensitive immunoassay for quantifying the levels of TFs delivered to cells. Quantifying intracellular delivery of the pluripotency TFs would not only provide further insights into the cause of the partial bioactivity, but may also offer revealing explanations into the low nuclear reprogramming efficiencies obtained when using recombinant proteins [8-10].

Alternatively, the inability to remove the IF2D1 which suggests incomplete folding could again explain the difference in gene expression observed with the retroviral Oct4 and Sox2-treated cells relative to the IF2D1-Oct4-R9 and IF2D1-Sox2-R9-treated cells. TFs, especially those involved in stimulating pluripotency, participate in an intricate regulatory network and are under many levels of control; successful gene expression requires the TF to interact with not only its consensus sequence, but also other TFs, RNA polymerases, and proteins that modify epigenetics [35, 36]. Therefore, it is possible that while the DNA binding domains of the IF2D1-Oct4-R9 and IF2D1-Sox2-R9 are folded correctly (Fig. 4), the transactivation domains that interact with other proteins are not (Fig. 5). The fact that the passenger TFs precipitated after IF2D1 cleavage supports this hypothesis; passenger proteins often revert to their insoluble state following solubility partner cleavage because they were never folded correctly in the first place [28]. This can be remedied with a systematic denaturation-refolding study which takes advantage of the efficiency that the IF2D1 fusion offers in producing the protein starting material. The IF2D1 fusion technology could be combined with our filter microplate DNA binding assay [31] to create a multidimensional refolding screen that assesses both TF solubility after cleavage and TF cognate DNA binding. On-column refolding using chromatography resin in the filter microplate allows high throughput evaluations and discourages unfavorable aggregation since protein aggregation is a second order event [37]. The ability of the refolded products to bind cognate DNA can be directly assessed on the "column," and TF passenger solubility can be assessed by cleaving off the passenger while the N-terminal His₆ Tag-IF2D1 portion of the fusion protein remains attached to the stationary phase.

Another hypothesis is that the un-cleaved IF2D1 is sterically blocking the transactivation domain. One of the transactivation domains of Oct4 is located at the N-terminus [38] and the transactivation domain of Sox2 is located in the C-terminus [39]. Thus, it is possible that the N-terminal IF2D1 could interfere with key protein-protein interactions necessary for gene expression. We fused the IF2D1 at the N-terminus to take advantage of both the high expressivity (efficient translation initiation) and high solubility properties of the IF2D1 [18]. However, future work may include the exploration of a C-terminal IF2D1 fusion. Though C-terminal fusion of the IF2D1 may have a limited chance of success because the IF2D1 is not known to have any chaperone functions (unlike MBP) [29], it would still be interesting to determine if a C-terminal IF2D1 fusion can help shepherd the passenger protein to a proper conformation, rescuing inappropriate preliminary folding to result in a soluble passenger protein after IF2D1 removal.

It is very interesting that IF2D1-Oct4-R9 successfully downregulated the expression of SSBP2 (Fig. 5B). Since the IF2D1-Oct4-R9 showed DNA binding activity and downregulation may not require as many protein-protein interactions as upregulation, we suggest that the SSBP2 experiment confirms DNA binding *in vivo*. But, DNA binding does not automatically translate to full functionality. This underscores the importance of a reliable bioassay in confirming the bioactivity of an engineered transcription factor.

In conclusion, we have shown that IF2D1 enables high level expression of soluble pluripotency TF fusion proteins. However, removal of the IF2D1 solubility partner results in insolubility of the TF passenger. We then showed that un-cleaved IF2D1-Oct4-R9, IF2D1-Sox2-R9, IF2D1-Klf4-R9, and IF2D1-Nanog-R9 exhibit specific binding to their consensus DNA sequences. When we assayed for the bioactivity of the IF2D1-Oct4-R9 and IF2D1-Sox2-R9 however, we found that they do not act on all downstream gene targets. Even though the IF2D1 strategy did not yield soluble TFs or fully bioactive fusions, this approach may still be useful in producing starting material conducive for protein refolding applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Henry Zhu for his help with the quantitative PCR reactions. WCY was a recipient of NDSEG and NSF fellowships. The work was supported in part by grants from the National Heart, Lung, and Blood Institute (1U01HL100397; JPC) and the Wallace H. Coulter Foundation (JRS).

References

1. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*. 2007; 131:861–872. [PubMed: 18035408]
2. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, Slukvin II, Thomson JA. Induced pluripotent stem cell lines derived from human somatic cells. *Science*. 2007; 318:1917–1920. [PubMed: 18029452]
3. Maherali N, Hochedlinger K. Guidelines and techniques for the generation of induced pluripotent stem cells. *Cell Stem Cell*. 2008; 3:595–605. [PubMed: 19041776]
4. Stadtfeld M, Nagaya M, Utikal J, Weir G, Hochedlinger K. Induced pluripotent stem cells generated without viral integration. *Science*. 2008; 322:945–949. [PubMed: 18818365]
5. Okita K, Nakagawa M, Hyenjong H, Ichisaka T, Yamanaka S. Generation of mouse induced pluripotent stem cells without viral vectors. *Science*. 2008; 322:949–953. [PubMed: 18845712]

6. Kaji K, Norrby K, Paca A, Mileikovsky M, Mohseni P, Woltjen K. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*. 2009; 458:771–775. [PubMed: 19252477]
7. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, Cowling R, Wang W, Liu P, Gertsenstein M, Kaji K, Sung HK, Nagy A. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*. 2009; 458:766–770. [PubMed: 19252478]
8. Zhou H, Wu S, Joo JY, Zhu S, Han DW, Lin T, Trauger S, Bien G, Yao S, Zhu Y, Siuzdak G, Scholer HR, Duan L, Ding S. Generation of induced pluripotent stem cells using recombinant proteins. *Cell Stem Cell*. 2009; 4:381–384. [PubMed: 19398399]
9. Kim D, Kim CH, Moon JI, Chung YG, Chang MY, Han BS, Ko S, Yang E, Cha KY, Lanza R, Kim KS. Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell*. 2009; 4:472–476. [PubMed: 19481515]
10. Cho HJ, Lee CS, Kwon YW, Paek JS, Lee SH, Hur J, Lee EJ, Roh TY, Chu IS, Leem SH, Kim Y, Kang HJ, Park YB, Kim HS. Induction of pluripotent stem cells from adult somatic cells by protein-based reprogramming without genetic manipulation. *Blood*. 2010
11. Edenhofer F. Protein transduction revisited: novel insights into the mechanism underlying intracellular delivery of proteins. *Curr Pharm Des*. 2008; 14:3628–3636. [PubMed: 19075739]
12. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CW, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*. 2006; 38:431–440. [PubMed: 16518401]
13. Bosnali M, Edenhofer F. Generation of transducible versions of transcription factors Oct4 and Sox2. *Biol Chem*. 2008; 389:851–861. [PubMed: 18681826]
14. Chung Y, Bishop CE, Treff NR, Walker SJ, Sandler VM, Becker S, Klimanskaya I, Wun WS, Dunn R, Hall RM, Su J, Lu SJ, Maserati M, Choi YH, Scott R, Atala A, Dittman R, Lanza R. Reprogramming of human somatic cells using human and animal oocytes. *Cloning Stem Cells*. 2009; 11:213–223. [PubMed: 19186982]
15. Yang WC, Patel KG, Lee J, Ghebremariam YT, Wong HE, Cooke JP, Swartz JR. Cell-free production of transducible transcription factors for nuclear reprogramming. *Biotechnol Bioeng*. 2009; 104:1047–1058. [PubMed: 19718703]
16. Jewett MC, Swartz JR. Mimicking the Escherichia coli cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotechnol Bioeng*. 2004; 86:19–26. [PubMed: 15007837]
17. Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. *BMC Biotechnol*. 2004; 4:32. [PubMed: 15598350]
18. Sorensen HP, Mortensen KK. Advanced genetic strategies for recombinant protein expression in Escherichia coli. *J Biotechnol*. 2005; 115:113–128. [PubMed: 15607230]
19. Moreno JM, Drskjotersen L, Kristensen JE, Mortensen KK, Sperling-Petersen HU. Characterization of the domains of E. coli initiation factor IF2 responsible for recognition of the ribosome. *FEBS Lett*. 1999; 455:130–134. [PubMed: 10428486]
20. Sorensen HP, Sperling-Petersen HU, Mortensen KK. A favorable solubility partner for the recombinant expression of streptavidin. *Protein Expr Purif*. 2003; 32:252–259. [PubMed: 14965771]
21. Kuchenreuther JM, Stapleton JA, Swartz JR. Tyrosine, cysteine, and S-adenosyl methionine stimulate in vitro [FeFe] hydrogenase activation. *PLoS One*. 2009; 4:e7565. [PubMed: 19855833]
22. Hoover DM, Lubkowski J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res*. 2002; 30:e43. [PubMed: 12000848]
23. Abramoff MD, Magelhaes PJ, Ram SJ. Image Processing with ImageJ. *Biophotonics International*. 2004; 11:36–42.
24. Gadgil H, Jurado LA, Jarrett HW. DNA affinity chromatography of transcription factors. *Anal Biochem*. 2001; 290:147–178. [PubMed: 11237318]

25. Chew JL, Loh YH, Zhang W, Chen X, Tam WL, Yeap LS, Li P, Ang YS, Lim B, Robson P, Ng HH. Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. *Mol Cell Biol*. 2005; 25:6031–6046. [PubMed: 15988017]
26. Jauch R, Ng CK, Saikatendu KS, Stevens RC, Kolatkar PR. Crystal structure and DNA binding of the homeodomain of the stem cell transcription factor Nanog. *J Mol Biol*. 2008; 376:758–770. [PubMed: 18177668]
27. Jiang J, Chan YS, Loh YH, Cai J, Tong GQ, Lim CA, Robson P, Zhong S, Ng HH. A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol*. 2008; 10:353–360. [PubMed: 18264089]
28. Esposito D, Chatterjee DK. Enhancement of soluble protein expression through the use of fusion tags. *Curr Opin Biotechnol*. 2006; 17:353–358. [PubMed: 16781139]
29. Kapust RB, Waugh DS. Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci*. 1999; 8:1668–1674. [PubMed: 10452611]
30. Nallamsetty S, Waugh DS. Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expr Purif*. 2006; 45:175–182. [PubMed: 16168669]
31. Yang WC, Swartz JR. A filter microplate assay for quantitative analysis of DNA-binding proteins using fluorescent DNA. *Anal Biochem*. 2011
32. Chavez L, Bais AS, Vingron M, Lehrach H, Adjaye J, Herwig R. In silico identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach. *BMC Genomics*. 2009; 10:314. [PubMed: 19604364]
33. Sharov AA, Masui S, Sharova LV, Piao Y, Aiba K, Matoba R, Xin L, Niwa H, Ko MS. Identification of Pou5f1, Sox2, and Nanog downstream target genes with statistical confidence by applying a novel algorithm to time course microarray and genome-wide chromatin immunoprecipitation data. *BMC Genomics*. 2008; 9:269. [PubMed: 18522731]
34. Warren L, Manos PD, Ahfeldt T, Loh YH, Li H, Lau F, Ebina W, Mandal PK, Smith ZD, Meissner A, Daley GQ, Brack AS, Collins JJ, Cowan C, Schlaeger TM, Rossi DJ. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell*. 2010; 7:618–630. [PubMed: 20888316]
35. Latchman DS. Transcription factors: an overview. *Int J Biochem Cell Biol*. 1997; 29:1305–1312. [PubMed: 9570129]
36. Stein GS, Stein JL, Van Wijnen AJ, Lian JB, Montecino M, Croce CM, Choi JY, Ali SA, Pande S, Hassan MQ, Zaidi SK, Young DW. Transcription factor-mediated epigenetic regulation of cell growth and phenotype for biological control and cancer. *Adv Enzyme Regul*. 2010; 50:160–167. [PubMed: 19896493]
37. Kiefhaber T, Rudolph R, Kohler HH, Buchner J. Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation. *Biotechnology (N Y)*. 1991; 9:825–829. [PubMed: 1367356]
38. Imagawa M, Miyamoto A, Shirakawa M, Hamada H, Muramatsu M. Stringent integrity requirements for both trans-activation and DNA-binding in a trans-activator, Oct3. *Nucleic Acids Res*. 1991; 19:4503–4508. [PubMed: 1886774]
39. Nowling TK, Johnson LR, Wiebe MS, Rizzino A. Identification of the transactivation domain of the transcription factor Sox-2 and an associated co-activator. *J Biol Chem*. 2000; 275:3810–3818. [PubMed: 10660531]

PREP-11-219 Highlights

- The first domain of IF2 is a fusion partner that improves TF protein solubility
- Cleavage of IF2 from fusion protein results in insolubility of the passenger TFs
- Un-cleaved IF2D1-TF-R9 fusion proteins bind cognate DNA sequences
- But un-cleaved TF fusion proteins do not fully activate target downstream genes
- Un-cleaved IF2 may hinder transactivation; passenger TFs may be incorrectly folded

Figure 1. Fusion Protein Schematic

A. TF-R9 design and B. IF2D1-TF-R9 design. All constructs were cloned into the pET24a vector (Novagen). HIS: hexahistidine purification tag, IF2D1: IF2 Domain I, LNK: GGGGS linker, TEV: tobacco etch virus cleavage site to enable removal of the IF2D1 if necessary, Transcription Factor Coding Sequence: coding sequence for human transcription factors Oct4, Sox2, Nanog, Lin28, cMyc, and Klf4, NLS: nuclear localization sequence, KKRRKV, R9: nona-arginine protein transduction domain. Schematic depiction not drawn to scale.

Figure 2. IF2 Domain I improves total and soluble fusion protein production

Very little soluble expression of TF-R9 is seen without the use of the IF2D1 solubility partner. Cultures were induced at 0.6 OD595 and protein expression was carried out for 6-7h at 25° C. Cells were lysed at time of harvest using a Bugbuster-to-cell ratio of approximately 50 uL/mg. Total (T) and soluble (S) proteins were separated using a 15 minute, 20,000g centrifugation. Approximate 100µg of cell mass was loaded into each lane. Molecular weight ladder (L) used was the Invitrogen SeeBlue Plus2 ladder. Molecular weights: Oct4-R9 (43 kDa), Sox2-R9 (39 kDa), Nanog-R9 (39 kDa), Lin28-R9 (27 kDa), cMyc-R9 (53 kDa), Klf4-R9 (55 kDa). The IF2D1 adds approximately 20 kDa onto the molecular weight of the resulting fusion protein.

Figure 3. Purification of IF2D1-Oct4-R9

A. Chromatogram of the wash and elution steps of cation exchange chromatography purification of IF2D1-Oct4-R9 using a 70-mL SP-Sepharose column. B. Gel electrophoresis of IF2D1-Oct4-R9 elution fractions. Samples were taken from the beginning, middle, and end of the elution peak. Full-length IF2D1-Oct4-R9 is approximately 60 kDa. The other bands appear to be degradation products of the IF2D1-Oct4-R9 as we did not observe these bands in the small scale purifications. Similar purification patterns were observed for IF2D1-Sox2-R9, IF2D1-Nanog-R9, IF2D1-Klf4-R9, and IF2D1-Lin28-R9.

Figure 4. IF2D1-Oct4-R9 and IF2D1-Sox2-R9 bind to their respective cognate DNA targets
A competitive EMSA-like assay was conducted using fluorescent cognate DNA in a 96-well filter plate format. The IF2D1 fusion proteins exhibited specific DNA-binding activity (n=2).

Figure 5. IF2D1-Oct4-R9 and IF2D1-Sox2-R9 do not exhibit full bioactivity

BJ fibroblasts were treated with 200 nM of IF2D1-Oct4-R9 or IF2D1-Sox2-R9 fusion protein in DMEM with 10% FBS and 1% penicillin/streptomycin (pen-strep) antibiotics (Invitrogen) every 24 h for 120 h. Cellular RNA was extracted for real-time RT-PCR analysis at 0, 72, and 144 h. The effect of IF2D1-Oct4-R9 and IF2D1-Sox2-R9 on the gene expression of several known downstream targets was measured using quantitative real-time RT-PCR. Cells infected with retroviruses encoding GFP on day 0 served as the negative control while cells infected with retroviruses encoding Oct4 or Sox2 on day 0 served as the positive control. A-F: Genes regulated by Oct4, performed in duplicate. G-I: Genes regulated by Sox2, performed in triplicate.

Table 1
Approximated purified yields of un-cleaved IF2D1-TF-R9 fusion proteins

IF2D1-TF-R9	Purified Yield of Uncleaved Fusion Protein (mg per L culture)
IF2D1-Oct4-R9	16
IF2D1-Sox2-R9	2
IF2D1-cMyc-R9	n/a
IF2D1-Klf4-R9	16
IF2D1-Nanog-R9	2
IF2D1-Lin28-R9	10