cDNA–GFP Fusion Libraries for Analyses of Protein Localization in Mouse Stem Cells

by

Heather M.E. Murray

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Biology

Waterloo, Ontario, Canada, 2005

Author’s Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Stem cells have great potential value for treating a number of diseases and conditions, including diabetes, Parkinson's, and spinal cord injuries. Applying stem cells for therapeutic purposes will require an in-depth understanding of their biology, not only of the genes they express, but also the functions of the proteins encoded by the genes. The goal of the project presented in this thesis was to develop a method for high-throughput analyses of protein localization in mouse stem cells. Localization information can provide insight into the functions and biological roles of proteins.

One means of studying protein localization involves creating proteins with a green fluorescent protein (GFP) reporter gene and analyzing their localization using fluorescence microscopy. The research outlined in this thesis focused on developing a system to create a large number of GFP-tagged proteins by constructing a cDNA–GFP fusion library. This involved exploring methods for optimizing cDNA synthesis, designing a retroviral vector (pBES23) for the expression of cDNA–GFP fusions in mouse stem cells, and constructing a cDNA–GFP fusion library in this vector using R1 mouse embryonic stem cell mRNA. The library constructed was not successfully delivered to target cells for GFP-tagged protein expression; it was therefore not possible to characterize protein localization in mouse stem cells. Suggestions are given as to how the methods used in this thesis might be optimized further.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2DGE	two-dimensional gel electrophoresis
ATCC	American Type Culture Collection
ATP	adenosine triphosphate
CMV	cytomegalovirus
C-terminal library	a cDNA–GFP fusion library in which tagged proteins are fused to the C terminus of GFP
DEPC	diethylpyrocarbonate
DMEM	Dulbecco's modified essential media
dNTP	deoxyribonucleotide triphosphate
ds	double-stranded
DTT	dithiothreitol
EDTA	ethylene diamine tetraacetic acid
EF-1α	elongation factor 1 α
eGFP	enhanced green fluorescent protein
env	envelope
ES	embryonic stem
ExoIII	exonuclease III
FCS	fetal calf serum
FGF-1	fibroblast growth factor 1
gag	group antigens polyproteins

GFP	green fluorescent protein
GUS	β-glucuronidase
HPLC	high performance liquid chromatography
HPRT	hypoxanthine guanine phosphoribosyl transferase
ITG-α6	integrin-α 6
LTR	long terminal repeat
MBN	Mung Bean Nuclease
MMuLV	Moloney Murine Leukemia Virus
MOI	multiplicity of infection
MP	multiphoton
MS	mass spectrometry
N-terminal library	a cDNA–GFP fusion library in which tagged proteins are fused to the N terminus of GFP
NFH	neurofilament H
ORF	open reading frame
PCR	polymerase chain reaction
PGK	phosphoglycerate kinase
pol	polymerase
poly-A+ RNA	polyadenylated RNA
PON	partner of Numb
PonA	ponasterone A

RT	reverse transcriptase
RV	retroviral
SOP	sensory organ precursor
ss	single-stranded
TAE	Tris-acetate EDTA
Tet	tetracycline
TIF1 β	transcription intermediary factor 1 β
UTR	untranslated region
UV	ultraviolet

The three-letter and one-letter abbreviations for amino acids and nucleotides used in this thesis are those designated by the International Union of Biochemistry and Molecular Biology (Moss, 2005).

Chapter 1 Introduction

1.1 Overview

The term “stem cell” refers to a broad category of cells that includes members of many different origins, including embryonic, neural and hematopoietic tissues. While a precise definition of the physical and genetic features that distinguish a stem cell from other cell types remains elusive, stem cells do have distinct functional similarities: they have the capacity to self renew, and they are involved in the generation or regeneration of tissues (Blau et al., 2001). It is the potential for stem cells to give rise to mature, differentiated cells that has motivated stem cell research in the past decade. If the process of expanding and differentiating stem cell populations can be achieved in a laboratory setting, then it is not unreasonable to think that human stem cells may eventually be used for therapeutic purposes in reconstructing or regenerating human tissues (Daley et al, 2003). However, before this level of control over stem cell biology is possible, it is first necessary to have an in-depth understanding of the cellular processes that regulate their growth and differentiation. This can only be gained through extensive and exhaustive analyses of stem cell biology at a molecular level.

This research was designed to evaluate the subcellular localization of proteins expressed in stem cells. Proteins are of central importance in maintaining stem cells or driving them to differentiate (Cavaleri & Scholer, 2003; Chambers, 2004). Altering or perturbing the protein population within a stem cell can affect whether it will self-renew, commit to differentiate, or die (Unwin et al., 2003). It is my hypothesis that by analyzing the proteins present in stem cells and where they localize it will be possible to gain insight into their functions. This knowledge will help to identify which proteins are important in stem cell development and maintenance.

Whereas therapeutic stem cell application will require the use of human stem cells, this study uses mouse cells as a model system. Research on mouse development may reveal which critical conserved proteins and pathways regulate renewal in human cell development, and may help identify markers that tightly correlate with stem cell state (Rao, 2004).

1.2 Protein Localization and Function

Protein localization in stem cells is of interest because the subcellular location of a protein can offer clues to its function. For example, proteins residing on the cell surface are likely to serve as receptor or transport molecules, and proteins localized to the nucleus have a high probability of interacting with DNA. However, identifying which proteins target to a certain cellular location may be difficult to determine experimentally.

The standard method for determining protein localization in high throughput combines subcellular fractionation techniques with mass spectrometry (MS). Organelle-enriched fractions are recovered from cell lysates, and the proteins present in the recovered fractions are separated using two-dimensional gel electrophoresis (2DGE) (Huber et al., 2003). The identities of proteins of interest are determined by excising proteins from the gel and subjecting them to peptide mass fingerprinting, which involves cleaving the protein enzymatically or chemically and obtaining a "fingerprint" of the peptide fragment masses using MS (Gevaert & Vandekerckhove, 2000). Identification of the protein is achieved by comparing the peptide mass fingerprint to virtual fingerprints of protein sequences stored in databases.

Drawbacks to assigning protein localization based on their presence in a subcellular fraction are that proteins must reside in subcellular compartments that are amenable to fractionation, and must also have physical properties that allow them to be resolved using 2DGE. Homogeneous organelles are difficult to isolate because they are fragile; proteins that are not components of an organelle may be isolated in the same subcellular fraction, thereby generating false data (Brunet et al., 2003). Furthermore, fractionation-based techniques cannot be used to determine protein constituents of transient cellular structures because these structures cannot be isolated. Several classes of proteins may also be excluded from analysis because of their physical properties. Those with poor solubility, such as membrane proteins, cannot be separated using 2DGE, and proteins of low abundance, such as transcription factors, may be difficult to recover because they are rare (Harry et al., 2000). Thus, subcellular fractionation approaches are not holistic in that they do not allow localization to be assigned to all proteins a cell may potentially express.

Determining the identity of proteins based on MS spectra may also be problematic in that it may not be possible to identify a protein based on its peptide mass fingerprint. Mass spectra may not match theoretical values for a number of reasons, including missed cleavage sites, tryptophan oxidation, methylation of aspartic acid and glutamic acid-rich peptides, or modifications that arise during gel electrophoresis (Thiede et al., 2005). Furthermore, the use of a fingerprint to identify proteins is not always possible as it relies on matching peptide masses to sequence data already present in databases.

In silico methods that predict protein localization based on gene or cDNA sequences provide an alternative to determining localization experimentally. These programs, including PSORT (Nakai & Horton, 1999), Proteome Analyst (Lu et al., 2004), and TargetP (Emanuelsson et al., 2000), use input nucleic acid coding sequences to derive a virtual translation product from which predicted localizations are made. Predictions are based on the presence of documented targeting sequences, homology to known targeting sequences, physical properties of amino acids, or a combination of these features (Donnes & Hoglund, 2004). In theory, one could input sequence data from cDNAs expressed in stem cells into a program to predict the subcellular localization of each protein. In reality, such programs cannot serve as a substitute for experimentally determining protein localization. Despite advances in methodology, the error rate can be quite high for any given method, as algorithms are trained using proteins with known localizations but are less proficient at predicting localization of proteins that are target by uncharacterized targeting sequences (Schneider & Fechner, 2004).

The shortcomings of subcellular fractionation-based methods and predictive programs highlight the limitations to determining the localization of a cell’s proteome. Ideally, one would like to be able to analyze proteins regardless of their physiological abundance or solubility and without a need for subcellular fractionation. To facilitate downstream analyses, in addition to assigning a subcellular localization to a given protein, it would be useful to have knowledge of the corresponding cDNA or gene sequence. Finally, given the number of different proteins that are present in a cell at any specific time it would be preferable to be able to characterize localization in a high-throughput manner.

1.3 Determining Protein Localization Using Green Fluorescent Protein

As an alternative to subcellular fractionation methods, the localization of a protein can be determined by adding a detectable epitope or domain to a protein of interest and using the tag to track localization within cells. Although there are several types of tags available to researchers, green fluorescent protein (GFP) is used widely due to its advantages over other tags. This protein fluoresces when excited by UV light (Tsien, 1998), meaning that the localization of proteins tagged with GFP can be detected using standard fluorescence microscopy. Using this method, proteins fused to GFP do not need to be isolated in order to assess their localization, making it possible to analyze proteins that are difficult to characterize using other methods because of their physical properties or low abundance. In addition, because a GFP-tagged protein is expressed from an experimentally introduced expression vector, it is possible to relate the localization of a protein to the sequence that encodes it.

A further advantage of using GFP instead of epitope tags, such as c-myc, or reporter genes, such as β-glucuronidase (GUS) or luciferase, is that GFP-tagged proteins are intrinsically fluorescent; it is not necessary to introduce any substrates or co-factors, or fix and permeabilize cells. In live cells, the localization and often the distribution of a protein (i.e. whether it is uniformly distributed in a region of the cell or concentrated into foci) can be determined with a high level of accuracy. Because determining localization requires nothing more than visual inspection, analyzing a large number of different GFP-tagged proteins is easy to do in a high-throughput manner.

The ability to observe GFP in living cells introduces a new dimension to protein localization studies: monitoring the dynamic properties of subcellular localization. This is of particular interest in the study of stem cells, where protein translocation can occur in association with differentiation. For example, in response to retinoic-acid induced differentiation of murine F9 cells the embryogenesis-related protein transcription intermediary factor 1 (TIF-1) β relocalizes from being diffuse in the nucleoplasm to being concentrated in distinct foci on heterochromatin (Cammas et al., 2002). The homeobox protein SOX9 is located in the cytoplasm of undifferentiated gonads, but localizes to the nucleus at the time of testis differentiation in male mouse embryos (Gasca et al., 2002). Given that there are many regulatory proteins yet to be discovered in stem cells, observing proteins for changes in localization as stem cells differentiate may lead to the discovery of more proteins associated with development and differentiation, and reveal important regulatory processes.

Time-course analyses of GFP-tagged protein localization may also allow for analysis of proteins with a dynamic localization and distribution during mitosis. Although most cellular divisions are symmetric, producing two daughter cells that are identical to the parent, there are circumstances in which cells divide asymmetrically by segregating proteins into one of their two daughter cells. In fact, the proliferation of stem cell populations may arise from several rounds of symmetrical division, followed by asymmetric divisions that eventually give rise to differentiated cell types (Roegiers & Jan, 2004). Proteins that are differentially segregated in asymmetric divisions are likely to be involved in cell maintenance or differentiation.

GFP has been useful in demonstrating protein partitioning in asymmetric divisions in *Drosophila*. Bellaïche et al., (2001) introduced a GFP-tagged version of the Numb adaptor protein, Partner of Numb (PON), into *Drosophila* sense organ precursor (SOP) cells and performed fluorescence time-course imaging. GFP-PON was clearly localized to the anterior cortex of a SOP cell prior to mitotic spindle formation. Following mitosis, GFP-PON was inherited by only one daughter cell.

While the role of asymmetric divisions during vertebrate development has not yet been established, this study demonstrates the usefulness of GFP-tagged proteins in identifying asymmetric divisions. It may be possible to identify such divisions by performing time-course imaging of stem cells expressing GFP-tagged proteins during cell growth and differentiation. Given the advantages of GFP over other tags, including the ability to monitor protein dynamics, I used a GFP-tagging approach to test a system for characterizing protein localization in mouse stem cells in high throughput.

1.4 Generating GFP-Tagged Proteins on a Large Scale

Analyzing the localization of a single protein or small group of proteins using GFP-tagging is relatively straightforward so long as the cDNAs encoding the proteins of interest are known; GFP-tagged protein expression constructs can be generated using traditional cloning techniques. However, generating targeted GFP-fusions for each protein a cell expresses using this method would be prohibitively labour intensive and time consuming.

An alternative to generating such a collection of GFP-tagged proteins is to create a large number of fusions *en masse* using a cDNA–GFP fusion library. The approach involves ligating a population of cDNAs from cells of interest into an expression vector adjacent the coding sequence for GFP (Figure 1.1). This strategy allows one to generate several thousand cDNA–GFP fusions in a single ligation reaction. (In this dissertation, *cDNA–GFP* fusion specifies a nucleic acid construct, whereas *GFP-tagged protein* refers to a protein fused to GFP.)

1.5 cDNA–GFP Fusion Libraries

cDNA–GFP fusion libraries have been constructed in a variety of organisms and cell lines, and have led to the discovery of novel proteins and targeting sequences, including a human nuclear envelope protein (Rolls et al., 1999), human nuclear localization signals (Fujii et al., 1999), and plasmodesmata proteins in tobacco (Escboar et al., 2003). Such libraries have also

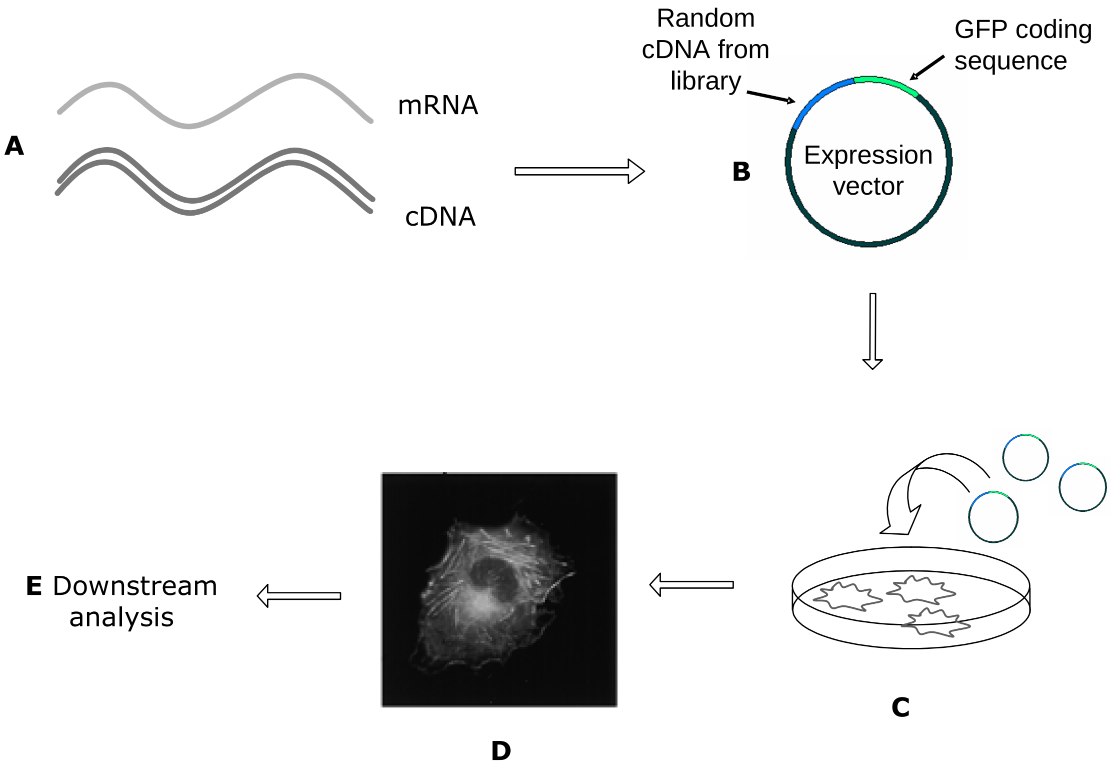

Figure 1.1 Strategy for constructing and screening a cDNA–GFP fusion library in mammalian cells.

(A) A cDNA library is made from a cell source of interest. (B) cDNAs are ligated into an expression vector adjacent to a gene for GFP. (C) cDNA–GFP fusion constructs are introduced into cells. (D) Transfected cells are imaged using fluorescence microscopy and observed for GFP-tagged protein localization. (E) Cells expressing proteins of interest are subjected to further analyses, including identifying the cDNA encoding the GFP-tagged protein.

allowed putative localizations to be assigned to previously uncharacterized proteins in *Arabidopsis thaliana* (Cutler et al., 2000). To date there has been no large-scale analysis of protein localization using a cDNA–GFP library in stem cells.

Previous studies using cDNA–GFP libraries show that the design of a library affects its overall quality and influences how cells may be screened for localization. Design considerations include the location of cDNAs relative to GFP, the method of cDNA synthesis, the library expression vector, the promoter driving expression of GFP-tagged proteins, and the method used to deliver cDNA–GFP fusions into target cells.

In a cDNA–GFP fusion library, cDNAs are ligated either downstream or upstream of the coding sequence for GFP. Consequently, the expressed proteins are fused to either the C-terminus or N-terminus of GFP (hereafter referred to as C-terminal libraries and N-terminal libraries respectively, in reference to the position of the tagged protein relative to GFP).

1.5.1 C-Terminal Libraries

As described in more detail in section 1.6, making C-terminal libraries is often easier than making N-terminal libraries because one can use standard cDNA synthesis techniques. However, in C-terminal libraries approximately 80% of GFP tagged proteins do not display a distinct subcellular localization (Cutler et al., 2000; Escobar et al., 2003). Non-localizing clones arise partly because in C-terminal libraries, cDNAs are ligated downstream of the gene for GFP; the probability that a cDNA is ligated in-frame with GFP is 1/3. However, out-of-frame fusions may be translated, resulting in non-native protein fragments tagged to GFP. Non-native proteins may contribute to the population of fluorescent clones displaying no distinct localization, or may give rise to artefactual localization. The peptides expressed from out of frame cDNAs may be able to direct protein localization due to the low information content of some targeting sequences (Manabe et al., 2002).

Proteins expressed in the correct reading frame will presumably localize normally if their targeting signals are available for recognition by cellular targeting machinery. However, a large number of cellular proteins contain N-terminal signal sequences that are essential for proper localization, including secreted proteins, membrane proteins, and proteins destined for endomembranes such as the ER, Golgi, and lysosomes (Ozawa et al., 2005). In a C-terminal library, the presence of GFP on the N-terminus of such proteins can block recognition of their localization sequence, and therefore these proteins will not properly localize.

The combination of expression of out-of-frame cDNAs and blocked targeting sequences means that a large number of GFP-tagged proteins in C-terminal libraries do not localize in a biologically relevant manner, and may generate significant amounts of inaccurate data. In Cutler et al. (2000), 50% of library clones observed to display a distinct subcellular localization were the result of cDNAs being expressed in the incorrect reading frame.

1.5.2 N-Terminal Libraries

While more technically challenging to make, N-terminal libraries contain a larger proportion of GFP-tagged proteins displaying a distinct subcellular localization. In such libraries, N-terminal targeting sequences are available for recognition by cellular targeting machinery. In addition, because protein synthesis initiates at the start codon of the cDNA of interest, GFP-tagged proteins are expressed in the correct reading frame. While this is advantageous in many respects, it is worth noting that a situation occurs opposite to that in C-terminal libraries: only one in three cDNA–GFP fusions will have the GFP gene in frame with the cDNA. However, because GFP-tagged proteins are expressed in frame, and N-terminal targeting sequences are available for recognition, although two out of three cDNAs will not be expressed with a fluorescent tag, the percentage of GFP-tagged proteins displaying a distinct subcellular localization will be much higher. Escobar et al. (2003) reported that more than 50% of cDNA–GFP fusions in their N-terminal library allowed for expression of a GFP-tagged protein with a distinct subcellular localization. In addition, more than 66% of localized proteins displayed native localization.

Relative to C-terminal libraries, N-terminal libraries generate more biologically relevant data, as a larger percentage of GFP-tagged proteins will localize, and the majority will target to their native localization.

1.6 Making cDNA–GFP Fusions

Constructing a fusion library requires more than ligating cDNA into an expression vector containing the coding sequence of GFP. For a cDNA–GFP fusion to direct synthesis of a GFP-tagged protein, the coding sequences of the cDNA and GFP must be in frame and contiguous; there can be no stop codons between the region encoding the N-terminal portion of the fusion protein (either cDNA or GFP) and the C-terminal protein (GFP or cDNA).

When making C-terminal libraries, in which cDNAs are ligated downstream of the GFP coding sequence, this requirement is met by modifying the coding sequence for GFP so it contains no stop codon, thus allowing translation to proceed through to the cDNA. For these libraries, cDNA may be generated using the standard method of priming polyadenylated (poly-A+) RNA with an oligo-dT primer. Ideally, this priming method initiates the synthesis of full-length cDNAs, which is advantageous because such cDNAs contain the complete open reading frame (ORF) for full-length proteins.

It is the presence of the 3´ untranslated region (UTR) and stop codon that complicates the construction of N-terminal libraries. Because cDNAs are ligated upstream of the coding sequence for GFP in the expression vector, a translational fusion requires cDNAs to have a start codon but no stop codon. Oligo-dT priming of mRNA cannot be used because of the presence of a stop codon in the full-length cDNAs. The alternative is to initiate synthesis of cDNA with random-hexamer primers.

In contrast to oligo-dT primers, which anneal to the poly-A+ tail of an mRNA, random-hexamer primers may anneal potentially anywhere in an mRNA molecule. Some will anneal near the 5´' end of an mRNA, some near the 3´ end, and some in the middle of an mRNA. Because of this, the average length of cDNA generated from priming with random hexamers is always shorter than cDNA synthesized using oligo-dT primers. However, it does allow cDNA synthesis to be initiated within an mRNA such that the resulting cDNA contains the mRNA sequence upstream of the stop codon; such cDNAs can be used to generate GFP fusion proteins in cDNA libraries.

1.7 General Considerations

It is clear that there is no ideal way to construct a cDNA–GFP fusion library, as no method allows one to generate full-length proteins fused with GFP and displaying native protein localization. However, there are advantages to using cDNA–GFP libraries to study protein localization. It is possible to generate a large number of GFP-tagged proteins in a relatively short amount of time, and one does not need any prior knowledge of transcript sequences to create fusion proteins. The ability to screen cells visually facilitates high-throughput screens, and the resolution of localization is much greater than that obtained using subcellular fractionation methods. Furthermore, localization may be determined for proteins that cannot easily be isolated and identified using other methods. As in any high-throughput approach, not all proteins will be amenable to analysis, but it is possible to gain insight that cannot be obtained using other techniques.

I used a cDNA–GFP library approach to study protein localization in mouse embryonic stem (ES) cells. To maximize the percentage of GFP-tagged proteins that localize within a library, and to increase the proportion that displays native localization, I explored ways to optimize cDNA synthesis (Chapter 2). I also constructed an expression vector appropriate for the delivery and expression of cDNA–GFP fusions to ES cells (Chapter 3). Building on the work in Chapters 2 and 3, I constructed a cDNA–GFP library from mouse ES cell poly-A+ RNA in the vector I designed, and introduced the cDNA–GFP library into cells for expression and analysis (Chapter 4).

Chapter 2 A Novel Method for removing the 3´ UTR and stop codon from cDNAs

2.1 Introduction

I chose to use an N-terminal library to study protein localization in mouse stem cells because N-terminal GFP-tagged proteins are more likely to display distinct and native subcellular localization. In previous studies, N-terminal libraries have been constructed using random-primed cDNA (Misawa et al., 2000; Escobar et al., 2003). Given the difficulties that may be associated with N-terminal libraries, most notably short average cDNA length, I developed a cDNA synthesis strategy to increase the proportion of GFP-tagged proteins expressed from full-length coding sequences. Such proteins will be more likely to display native localization, as they contain a greater proportion of the native protein and are therefore more likely to contain all intrinsic protein targeting sequences. The strategy I developed involved removing the 3´ UTR and stop codon from full-length cDNAs using the enzymes Exonuclease III (ExoIII) and Mung Bean Nuclease (MBN), and fusing the shortened cDNAs to the coding sequence of GFP.

ExoIII progressively digests one strand of double-stranded (ds) DNA in a 3´→5´ direction at 3´ ends or at nicks in ds DNA (Weiss, 1976). Following ExoIII digestion, one can remove the remaining single-stranded (ss) DNA overhang with a ss-specific nuclease, resulting in a blunt-ended molecule (Henikoff, 1987). The rate of nucleotide removal by ExoIII can be manipulated by altering the reaction temperature, salt concentration, and the relative amounts of enzyme, allowing for precise control of the amount of DNA digested (Hoheisel, 1993).

I intended to use ExoIII in combination with the ss nuclease MBN to remove the 3´ UTR and stop codon from full-length cDNAs. The length of 3´ UTRs is variable among different cDNAs, but can be estimated from sequence data. The median length of the 3´ UTR in rodents is 411 bases, with the 75th percentile being 543 bases (Makalowski et al., 1998). Therefore, deleting the terminal 500 bp from full-length ds cDNAs should remove the 3´ UTR and stop codon from the majority of cDNAs.

Since both ends of a full-length cDNA are susceptible to ExoIII degradation, the 5´ end must be protected. (Although each end of a ds cDNA has both 5´ and 3´ groups, in this thesis 5´ and 3´ refer to the ends of a ds cDNA corresponding to the 5´ and 3´ end of the template mRNA, respectively.) There are two properties of ExoIII activity that can be exploited to prevent digestion of one end of a linear DNA molecule. ExoIII cannot digest DNA with ss 3´ extensions of four or more bases (Rogers & Weiss, 1980; Guo & Wu, 1982), nor can it digest DNA containing alpha-phosphorothioate nucleotides (Putney et al., 1981). However, using these properties of ExoIII to protect one end of all cDNAs within a population is impractical, because neither can easily be used to modify only one end of linear ds DNA molecules.

I devised a strategy to protect the 5´ ends of cDNAs from digestion by physically preventing ExoIII from accessing them. The approach involves making full-length cDNA using a polymerase chain reaction (PCR)-based protocol (Zhu et al., 2001) with a biotinylated 5´ primer. If the resulting biotinylated cDNAs are bound to streptavidin-coated paramagnetic beads, the close association between the biotin and streptavidin will prevent ExoIII from accessing the susceptible 3´ OH group at the 5´ end of ds cDNAs. Following controlled degradation of the exposed end using ExoIII, the truncated cDNAs can be recovered from the reaction. Due to the strong association between biotin and streptavidin, the biotinylated strand cannot be removed from the beads. However, the non-biotinylated strand can be recovered by denaturing the ds cDNAs under alkaline conditions (Bowman & Palumbi, 1993). This truncated non-biotinylated strand can then be used as a template to direct synthesis of ds cDNA that is lacking the 3´ UTR and stop codon, yet still contains the full 5´ UTR, start codon, and the majority of the coding sequence (Figure 2.1). This truncated cDNA can direct the expression of a near to full-length protein.

This chapter outlines the experiments performed to establish protocols for applying this bead-bound truncation strategy to a population of DNA molecules. I designed a test system in which biotinylated DNA of a known length was used in a series of streptavidin-coated bead binding and elution experiments to gauge efficiency of DNA purification, bead binding, denaturation and ds DNA synthesis.

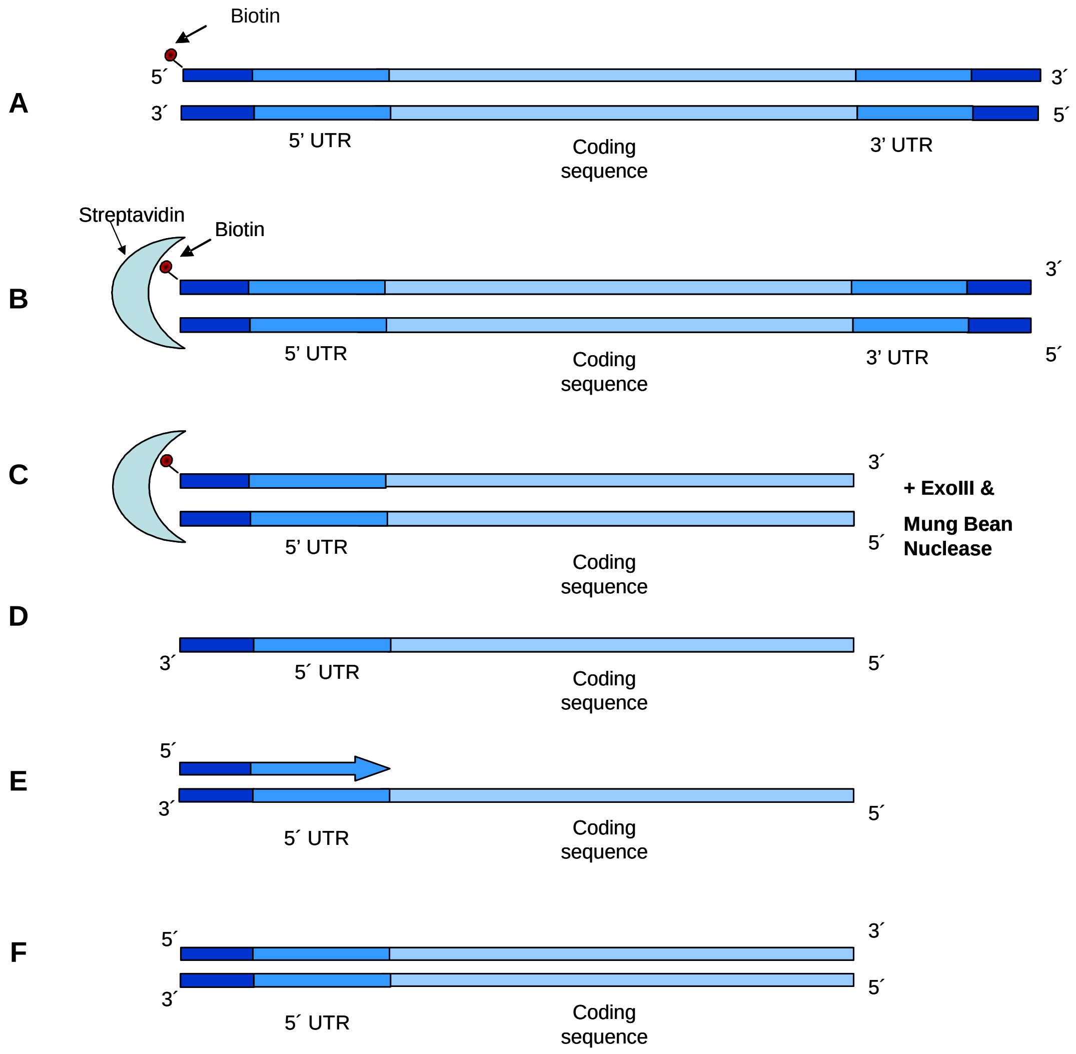

Figure 2.1 Strategy for generating cDNA lacking a 3´ UTR.

(A) cDNA is amplified using a PCR-based protocol with a biotinylated 5´ primer, resulting in full-length cDNA with a biotin group at the 5´ end and PCR priming regions flanking each end. (B) Biotinylated cDNA bound to streptavidin-coated beads. (C) ExoIII and Mung Bean Nuclease remove approximately 500 bp from the exposed 3´ end of cDNAs. (D) The non-biotinylated strand of cDNA is released from the complementary strand by denaturation under alkaline conditions. (E) ds cDNA synthesis is primed using a non-biotinylated version of the forward PCR primer. (F) Final cDNA produced lacking the 3´ UTR and stop codon.

2.2 Materials and Methods

2.2.1 Constructing a Biotinylated PCR Product of Defined Length

Primers were designed to generate a 2018 bp PCR amplification product using the plasmid pBluescript KS- (Stratagene) as a template. Biotinylated primers were high performance liquid chromatography (HPLC) purified by the manufacturer to remove residual biotin groups, which could potentially compete for binding sites on streptavidin-coated beads. Primers used in PCR reactions and DNA synthesis are shown in Table 2.1.

Table 2.1 PCR primers used to amplify a 2018 bp fragment of pBluescript

Primer	Primer Sequence
Forward	5´ ACC GTC TAT CAG GGC GAT GG
Forward-Biotinylated	5´ Biotin-ACC GTC TAT CAG GGC GAT GG
Reverse	5´ ATA AGA CTG GAT GGA

To ensure there was sufficient biotinylated DNA for analyses, multiple PCRs were performed under identical conditions. Each 50 µl reaction contained 5 µl 10 X PCR buffer –$MgCl_2$ (Fermentas), 3 µl 25 mM $MgCl_2$ (Fermentas), 1 µl 25 pmol/µl biotinylated forward primer (Sigma Aldrich), 1 µl 25 pmol/µl reverse primer (Sigma Aldrich), 1 µl 10 mM deoxyribonucleotide triphosphates (dNTPs) (Roche), 1 µl *Taq* (homemade recombinant stock), 1 µl pBluescript PCR template, and sterile MilliQ dH_20 to 50 µl. Thermal cycling was carried out in an Eppendorf Master Cycler using the parameters: 92°C for 5 min, followed by 35 cycles of 92°C for 45 sec, 65°C for 45 sec, and 72°C for 2 min 30 sec. Following this, samples were incubated at 72°C for 10 min before cooling to 10°C.

Aliquots of PCR products were electrophoresed on a 1% (w/v) agarose gel in 1 X TAE (40 mM Tris-acetate, 1 mM ethylene diamine tetraacetic acid (EDTA) pH 8.3) in the presence of ethidium bromide. A λ DNA/*Hin*dIII Marker (Fermentas) was used as a reference marker in electrophoresis. Residual primers and other PCR reactants were removed from the PCR products with a PCR Purification kit (Qiagen). One hundred microlitres of pooled PCR reactions were purified according to the manufacturer's protocol and eluted in 50 µl bead-binding buffer (0.2 M NaCl, 20 mM Tris-HCl, 1 mM EDTA, pH 7.6).

2.2.2 Binding of Biotinylated PCR Products to Streptavidin-Coated Beads

Magnesphere Paramagnetic Streptavidin-Coated Beads (Promega) were used for all binding reactions. To prepare beads for DNA binding, 100 µl of the bead suspension (1 mg/ml) were placed in a 1.5 ml microcentrifuge tube and captured in a magnetic stand (Dynal) for 1 min before removing storage liquid from the beads. All bead captures throughout the experiments were performed in the same manner. The beads were washed three times with 50 µl 7.5 mM trisodium citrate, 75 mM sodium chloride pH 7.2 followed by a fourth wash in 20 µl of this same solution. After each wash, beads were captured and the supernatant removed. Following the final wash, beads were resuspended in 100 µl of bead-binding buffer. Depending on the amount of beads required for a given reaction, a volume containing the appropriate amount of washed beads was removed to a fresh 1.5 ml microcentrifuge tube and beads were captured, the supernatant was removed, and the beads were resuspended in the experimental sample of biotinylated PCR products to a final volume of 40 µl in 1 X bead-binding buffer.

To ensure biotinylated DNA was well mixed with the streptavidin-coated beads during binding, the samples were gently mixed on a rotating rack for all incubation periods. For time course analyses of bead binding, at each sampling time point beads were captured before an aliquot of the supernatant was removed. Beads were then thoroughly resuspended and returned to the rotating rack. All bead-binding incubations were carried out at room temperature.

2.2.3 Denaturation of Bead-Bound DNA

ss DNA was eluted from the bead-bound PCR products using a denaturation protocol adapted from Sambrook & Russell (2001). Following the binding reaction, beads were captured and the supernatant containing unbound DNA was removed. Residual unbound DNA was washed from the beads by rinsing them twice with bead-binding buffer. The ss DNA was recovered by resuspending beads in 100 µl of freshly prepared 0.1 N NaOH at room temperature for 10 min with occasional agitation to keep the beads in suspension. Beads were captured and the supernatant was recovered and neutralized with 100 µl 1M Tris-Cl pH 7.5, then ethanol precipitated and stored in 70% ethanol at -20°C. For experimental trials using prolonged denaturation, samples were incubated for three times as long in the same volume and concentration of NaOH.

2.2.4 Synthesis of ds DNA from Eluted ss DNA

To generate ds DNA from ss DNA, a method was developed based on protocols for sequencing ss DNA templates using a thermophilic DNA polymerase (Sambrook & Russell, 2001). The precipitated ss DNA was resuspended in a volume of 25 µl containing 2.5 µl 10 X *Pwo* buffer (Roche), 0.5 µl 10 mM dNTPs (Roche), 0.5 µl25 pmol/µl forward primer, 0.5 µl*Pwo* polymerase (5 u/µl; Roche) and sterile MilliQ dH_20 to a volume of 25 µl. Reactants were heated to 95°C for 2 min, cooled to 65°C for 2 min and then heated to 72°C for 10 min. Following this, reactions were held at 4°C.

2.2.5 Testing DNA Release from Streptavidin-Coated Beads

To digest bead-bound DNA with a restriction enzyme, beads were rinsed once in sterile MilliQ dH_20 and resuspended in 10 µl 10 X Buffer O+ (Fermentas), 88 µl sterile MilliQ dH_20 and 2 µl*Not*I (10 u/µl; Fermentas). Beads were well suspended, then transferred to a rotating rack in a 37°C incubator for 16 h. Following this incubation, beads were captured in a magnetic stand and the supernatant containing the digested DNA was removed and stored at -20°C.

2.2.6 DNA Analysis and Quantification

Due to small sample sizes and low concentrations of DNA, the amount of DNA present in samples could not be quantified using spectrophotometry. Instead, DNA amounts were estimated by comparing their ultraviolet (UV) fluorescence intensity in the presence of ethidium bromide relative to that of known standard DNAs. Gels were imaged with a Fluorchem 8000 Chemiluminescence And Visible Imaging System (Alpha Innotech) and spot densitometry was performed on digital images using AlphaEase^{TM} Software v. 3.1.2 (Alpha Innotech). Fluorescence intensity readings from each reference DNA fragment in the λ DNA/*Hin*dIII molecular weight marker were used to generate a plot of nanograms DNA versus fluorescence intensity. Standard curves were generated using the sum of least squares method. For samples with a very low amount of DNA, this method proved to be unreliable due to background fluorescence. For samples containing approximately 15 ng or less of biotinylated DNA, the quantity of DNA in a sample was estimated by visually comparing the fluorescence intensity of a DNA fragment to the 5.8 ng reference DNA fragment in the λ DNA /*Hin*dIII standard.

2.3 Results

Because this approach was novel and required performing many manipulations where optimal reaction conditions were not known, I devised a system to test the efficiencies of various processes using a biotinylated PCR product 2 kb in length. This biotinylated DNA was used in experiments to gauge the efficiency of biotinylated DNA purification, binding to streptavidin-coated beads, and ss DNA elution from bead-bound DNA. Eluted ss DNA was used in ds DNA synthesis reactions. Once conditions for each step were optimized, I hoped to incorporate an ExoIII/MBN digest of bead-bound DNA and eventually develop protocols for applying the entire process to a population of cDNAs.

2.3.1 DNA Purification Yield and Bead Binding Efficiency

Initial experiments were performed to assess the recovery of the 2 kb biotinylated PCR product following purification and to test the efficiency of binding biotinylated DNA to streptavidin-coated beads. Time courses of bead binding were performed in parallel: one using the amount of beads recommended by the manufacturer, and one using twice that amount in the same volume. To facilitate direct comparisons between the two binding reactions, each used an equal quantity of purified DNA from a single purification reaction. Beads were incubated as described in the methods, and samples of the supernatant were collected following 30 min, 4 h, and 20 h incubations. Samples were electrophoresed and quantified as described in 2.2.6.

Based on UV fluorescence intensities, the yield of biotinylated DNA following purification was approximately 75% (Figure 2.2). This percentage recovery varied between 75% and 85% throughout experimental trials (data not shown). The efficiency of DNA binding to beads was estimated by analyzing the binding reaction supernatants for unbound DNA. Following 30 min, 66% of the biotinylated DNA was bound to beads; increasing the bead concentration or incubation period did not affect the percentage of bound DNA. In subsequent binding trials the percentage of non-binding biotinylated DNA was shown to remain relatively constant, varying between 25% and 36% (data not shown).

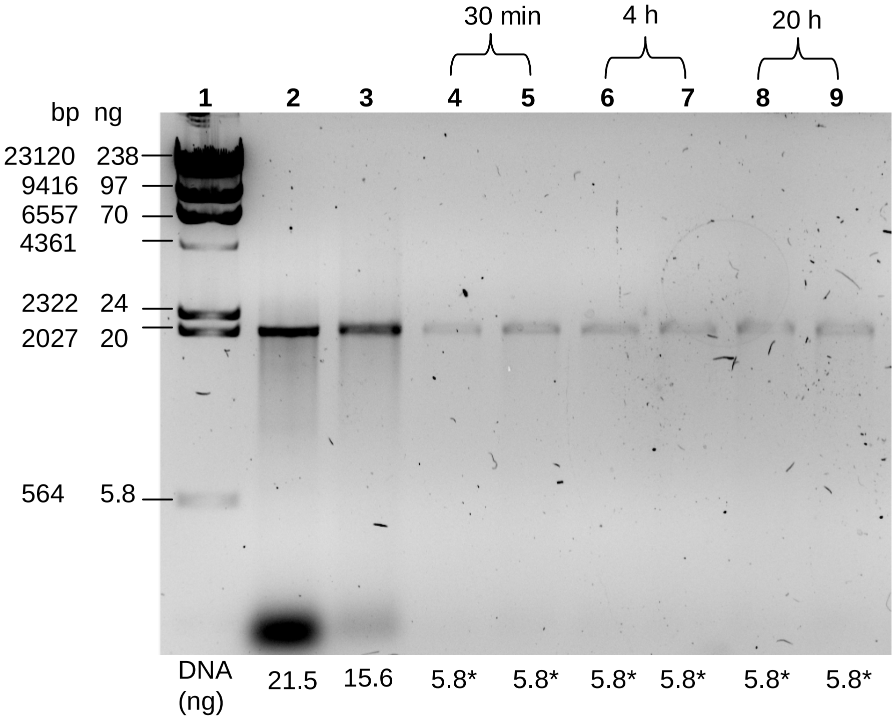

Figure 2.2 The effect of incubation time and bead concentration on the binding of biotinylated DNA to streptavidin-coated paramagnetic beads.

A time course analysis of biotinylated DNA binding to streptavidin-coated paramagnetic beads was carried out using two different bead concentrations: A (standard conditions) or B (twice the recommended amount of beads). The binding reactions were sampled at the time points shown and electrophoresed on a 1.0% (w/v) agarose gel alongside λ DNA/*Hin*dIII 500 ng (lane 1). Samples electrophoresed were unpurified biotinylated DNA (lane 2), purified biotinylated DNA (lane 3), and binding reaction supernatants collected from reactions A and B following 30 min (lanes 4 and 5, respectively), 4 h (lanes 6 and 7, respectively) and 20h incubation (lanes 8 and 9, respectively).

*Estimate of DNA amount based on visual comparison of fragment fluorescence intensity to 5.8 ng reference DNA fragment.

2.3.2 Recovery of ds DNA Following DNA Synthesis

To determine the recovery of ds DNA following DNA purification, bead binding, and ds DNA synthesis, DNA samples were obtained from each step of the entire process and electrophoresed on a 1.0% (w/v) agarose gel (Figure 2.3).

Recovery of ds DNA was poor. Although the amount of ds DNA could not be reliably measured due to low UV fluorescence intensity, it was estimated to be 5 ng, which is 8.5% of the starting DNA. Taking into account that DNA was lost during purification (15%) and bead binding (36%), the yield relative to the amount of DNA expected to be in the reaction increased to 15%.

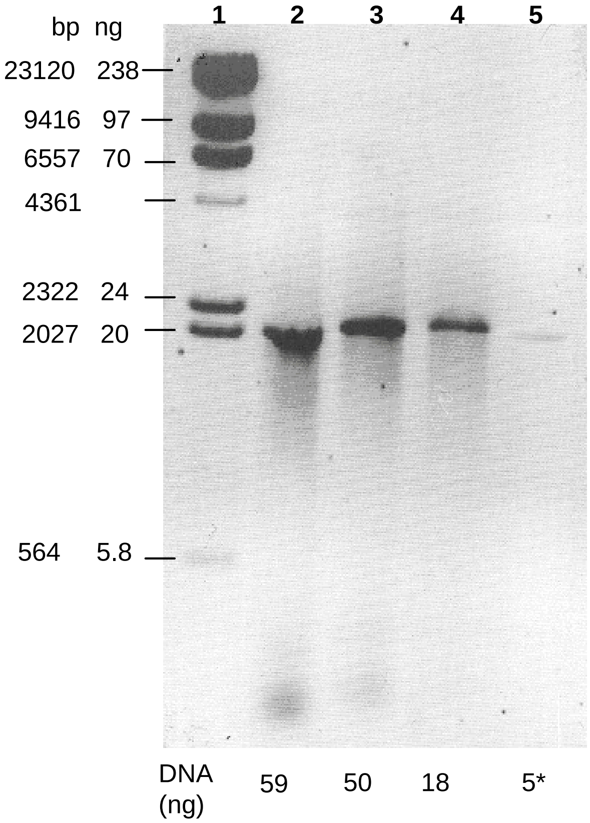

Figure 2.3 Recovery of ds DNA following biotinylated DNA purification, bead-binding, ss DNA elution, and ds DNA synthesis.

Samples were obtained from each step of a bead-binding and elution experiment and electrophoresed through a 1.0% (w/v) agarose gel. Shown are λ DNA/*Hin*dIII (500 ng; lane 1), unpurified biotinylated DNA (lane 2), purified biotinylated DNA (lane 3), supernatant containing unbound DNA following incubation with streptavidin-coated beads (lane 4), and ds DNA generated in a DNA synthesis reaction using ss DNA eluted from the beads (lane 5).

* Estimate of DNA amount based on visual comparison of fragment fluorescence intensity to 5.8 ng reference DNA fragment.

2.3.3 Effect of Prolonged Denaturation on ds DNA Yield

To test whether harsher denaturation could lead to improved yield of ds DNA, replicate samples of bead-bound DNA were used in two experiments, one employing more aggressive denaturation conditions with prolonged incubation in 0.1 N NaOH. Following denaturation, beads and supernatant were aspirated through a micropipette tip several times to promote separation of DNA strands before beads were captured and supernatant removed. Samples of ss DNA were used as a template for ds DNA synthesis.

The amount of ds DNA generated in each trial could be compared to estimate the relative recovery of ds DNA. The ds DNA yield did not increase using harsher denaturation conditions relative to standard conditions (Figure 2.4).

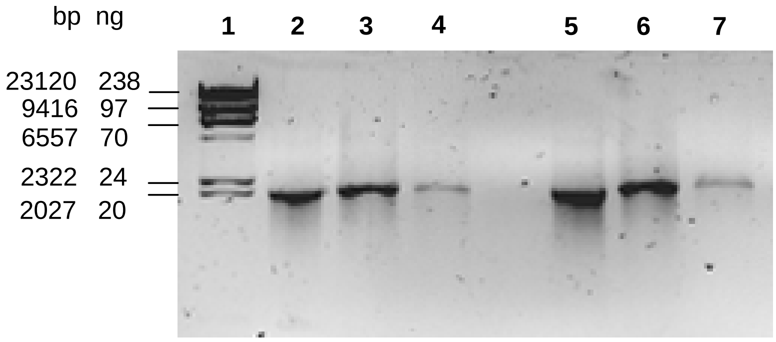

Figure 2.4 The effect of prolonged DNA denaturation on the recovery of ds DNA.

Samples of ds DNA generated from eluted ss DNA were electrophoresed on a 1.0% (w/v) agarose gel to determine yield relative to the amount of bound ds DNA. In a control experiment, DNA was denatured with standard conditions and used to make ds DNA. Samples electrophoresed were unpurified biotinylated DNA (lane 2), purified DNA (lane 3) and ds DNA (lane 4). In a second experiment, DNA was denatured under harsher conditions and similar samples were electrophoresed: unpurified biotinylated DNA (lane 5), purified DNA (lane 6) and ds DNA (lane 7). λ DNA/*Hin*dIII (500 ng) was used as a reference marker (lane1)

2.3.4 Measuring DNA Bound to Beads Following Denaturation

To determine the amount of ds DNA remaining bound to the streptavidin-coated beads following ss DNA denaturation, a *Not*I restriction enzyme digest was performed on the residual bead-bound DNA. There is a single *Not*I restriction site in the biotinylated DNA fragment, and if any biotinylated DNA was bound to the beads, a 1549 bp DNA fragment would appear in the supernatant following *Not*I digestion.

This experiment revealed that the majority of biotinylated DNA remained bead-bound DNA following denaturation (Figure 2.5).

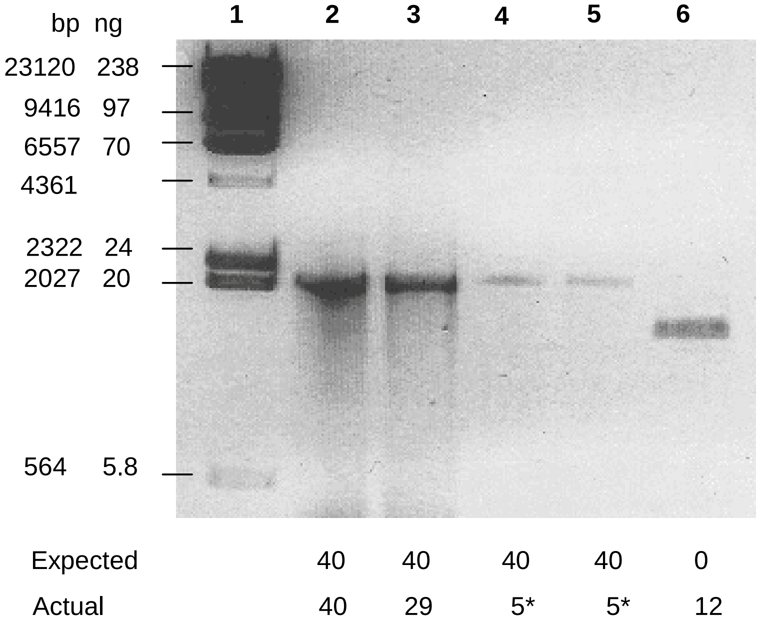

Figure 2.5 A comparison of the amount of ds DNA synthesized from eluted ss DNA and the amount of DNA remaining bound to beads.

Samples of ds DNA synthesized from eluted ss DNA were electrophoresed on a 1.0% (w/v) agarose gel alongside a sample of a *Not*I digest that was performed on beads following ss DNA elution. Samples electrophoresed were unpurified biotinylated DNA (lane 2), purified biotinylated DNA (lane 3), ds DNA synthesized from ss DNA (lanes 4 and 5), and DNA released into the bead supernatant following a *Not*I restriction digest (lane 6). The amount of expected DNA was calculated for each sample and compared to the actual yield. λ DNA/*Hin*dIII (500 ng) was used as a reference marker (lane 1).

* Estimate of DNA based on visual comparison of fragment fluorescence intensity to 5.8 ng reference DNA fragment.

2.4 Discussion

The experiments described in this chapter were designed to evaluate and optimize protocols for the recovery of near to full-length cDNAs lacking their 3´ UTRs and stop codons. The strategy involved generating 5´ biotinylated ds cDNA and binding it to streptavidin-coated beads, then using the enzymes ExoIII and MBN to shorten the exposed 3´ ends of cDNAs by 500 bp. The ds cDNAs would be recovered by eluting the non-biotinylated strands of shortened cDNAs and using them as a templates to direct ds cDNA synthesis. These shortened cDNAs would then be used in a cDNA–GFP fusion library to direct the synthesis of near to full-length proteins tagged with GFP at their N-terminus. Prior to applying this strategy to a population of cDNAs, I tested the efficiencies of the process of biotinylated DNA purification, binding to streptavidin-coated beads, ss DNA elution, and ds DNA synthesis using a biotinylated PCR product 2 kb in length.

In initial experiments, 75% of the DNA was recovered from the purification step and 66% of the biotinylated DNA bound to streptavidin-coated beads. Based on results of experimental trials, the cause of unbound DNA in a binding reaction was not immediately obvious. If incomplete binding occurred because the process was inefficient, increased binding time should lead to a decrease in the amount of unbound DNA, as would increasing the number of beads, but this was not observed (Figure 2.2).

Because the reaction conditions could not be manipulated to increase binding, the unbound DNA was determined to be a function of the DNA itself. If a portion of the PCR products did not have biotin groups, they would not be able to bind to streptavidin-coated beads. Because the primers were HPLC purified, it seems unlikely that upwards of 25% of them were not biotinylated; HPLC purification would have eliminated non-biotinylated primer molecules. A more likely explanation is that primer molecules were damaged or degraded due to storage or handling conditions.

The recovery of ds DNA from the system was 8.5% (Figure 2.3). Taking into account DNA loss during purification (15%) and bead binding (36%) the yield relative to the amount of DNA expected to be in the reaction was 15%, which is not sufficient for making a library.

Initially the basis for the low yields was not apparent. To pinpoint the cause more accurately, I performed more focused trials to determine the efficiency of individual steps in the process from DNA denaturation to ds DNA synthesis.

To see if the low yields of ds DNA could be improved by more aggressive denaturation of bead-bound DNA, I performed experiments with prolonged denaturation. This would increase the amount of ss DNA released, and therefore the amount of ds DNA that could be synthesized. However, more aggressive denaturation did not lead to improved yield, which suggested that low yields were a consequence of inefficiency in a different part of the procedure.

Before ruling out inefficient denaturation as a cause of low ds DNA yield, I wanted to have a more direct measure of the quantity of ss DNA eluted from beads. Determining this proved to be difficult, however, as the quantities of ss DNA eluted were too low to give reliable readings in a spectrophotometer. It was also not possible to quantify ss DNA by measuring fluorescence in ethidium bromide agarose gels, because ethidium bromide has a relatively low affinity for ss DNA. The fluorescence intensity of ss DNA in an ethidium bromide agarose gel cannot be compared to intensities from ds DNA reference markers.

As an alternative to measuring eluted ss DNA, I chose to measure the amount of ds DNA remaining bound to beads following denaturation. This was achieved by subjecting beads to a *Not*I restriction digest after ss DNA elution. This revealed that the majority of ds DNA was not being denatured, and was in fact remaining bound to beads following denaturation.

2.5 Conclusions

While initially this method appeared to be a promising way to create cDNA molecules for a cDNA–GFP fusion library, in retrospect, it involved several technically vague and challenging steps. DNA loss occurred at numerous stages throughout the overall process. Approximately 25% of the DNA was lost during DNA purification, and a further 25–36% of DNA did not bind to streptavidin-coated beads. Denaturing bead-bound DNA was also inefficient as evidenced by the high amounts of DNA remaining on beads following denaturation. This would not be acceptable for use in a cDNA library, because the loss of such a substantial portion of the DNA would greatly reduce the complexity, or number of independent clones. A library would likely also incur additional DNA losses in subsequent manipulations, including purification following adapter ligation and cDNA size fractionation. It is clear that several parts of the process limit the value of this strategy.

While there is potential for this process to generate greater yields of ds DNA with some procedural optimization at the steps of DNA purification, denaturation, and ds DNA synthesis, I chose to use a more standard method of generating cDNA and did not continue experiments to obtain increased yields of ds DNA from the bead system. Because I abandoned this strategy, I did not perform experiments to characterize ExoIII digestion of bead-bound DNA fragments. This decision was made based on evidence that suggests this DNA truncation strategy would not be suitable for use with a population of DNAs with various lengths, which is the case in a cDNA library.

Using this approach with a cDNA library would introduce a size bias in the cDNA population. While there is no evidence to suggest that the cDNA amplification technique I proposed would cause a bias towards shorter cDNAs (Wellenreuther et al., 2004), the efficiency of binding to beads is inversely proportional to the length of the DNA molecules. Long cDNAs bind to streptavidin-coated beads much less efficiently than do shorter cDNAs, and long cDNAs can be difficult to capture on a streptavidin-coated beads. This bias in binding is thought to be due to charge repulsion between DNA molecules (Liu & Price, 1997). The decreased binding efficiency with increasing length is such that it is recommended to use the shortest biotinylated DNA fragments possible for bead binding (Liu & Price, 1997). Difficulties in denaturing long fragments of bead-bound DNA would further contribute to this bias towards shorter cDNAs. Longer DNA molecules are more difficult to denature and recover from beads than are shorter ones (Promega Technical Services, personal communication). Therefore, if a population of biotinylated cDNA was bound to beads, digested with ExoIII and MBN, denatured, and made ds again, there would be an overabundance of shorter cDNAs in the recovered ds cDNA population because longer cDNAs would have been selected against in both bead binding and elution. To achieve the goal of this project, which is to develop a method for determining protein localization on a large scale and in high throughput, it is essential to have a complex population of cDNA–GFP fusions to screen. Given the poor yields and the potential for a greatly reduced library complexity, I chose to pursue other avenues for creating cDNA–GFP fusions on a large scale.

Chapter 3 Vector Design

3.1 Introduction

When constructing a cDNA–GFP fusion library, the method of gene delivery and control of expression of GFP-tagged proteins are determined by the vector used for library construction. Thus, the choice of expression vector is critical to the success and utility of a library.

Standard plasmid-based mammalian expression vectors are often delivered into cells using bulk plasmid transformation techniques, such as those mediated by lipofectamine or calcium–phosphate precipitation. These methods generate DNA precipitates that can reach 2000 kb in size and contain multiple plasmid DNA molecules (Perucho et al., 1980). If a cDNA–GFP library was introduced into cells in this manner, many different cDNA–GFP fusion plasmids would be delivered to each transfected cell. Problems inherent in this approach were highlighted in a study by Rolls et al. (1999) in which a cDNA–GFP fusion library was introduced to cells using lipofectamine. Each transfected cell expressed an estimated 5 to 20 different cDNA–GFP fusion plasmids, making it extremely difficult to recover cells expressing one GFP-tagged protein. Many rounds of transfection, screening and segregation of cDNA–GFP plasmid pools were required to isolate transformants carrying individual cDNA–GFP fusions of interest.

Using electroporation to deliver plasmids to cells can allow more precise control over plasmid copy number, as the number of plasmid molecules a cell receives can be roughly controlled by adjusting the concentration of DNA in an electroporation reaction. However, the epigenetic instability of the plasmid cDNA–GFP fusions means that the expression of cDNA–GFP fusion proteins may occur only briefly, and cDNA–GFP fusions may be lost from cells before they can be identified.

To address these problems, retroviral (RV) vectors can be used to deliver cDNA–GFP fusions to cells. Viral delivery systems allow for more control over the number of vector molecules delivered to cells, and vectors stably integrate into the host cell genome. By varying the multiplicity of infection (MOI), conditions can be controlled such that on average, each cell in the transfection population receives one vector DNA molecule. Therefore, by using a RV-mediated delivery system, it should be possible to generate a population of cells, each stably expressing one cDNA–GFP fusion. This would be ideal for time-course analyses of protein expression because transgene expression is long-term. Because of their advantages over transient methods of vector delivery, I used a RV system for constructing a cDNA–GFP fusion library.

The next consideration was the promoter to be used to direct expression of the cDNA–GFP fusion sequences. While strong promoters allow for easy detection of GFP-tagged proteins, overexpression of protein products may alter cellular physiology or be toxic. To minimize this effect I wanted to use an inducible promoter to control expression. This would allow GFP-tagged protein expression to be activated only when cells are to be imaged, thus minimizing the potential toxic effects of prolonged gene overexpression.

Inducible promoters allow control over the expression of a gene based on the presence of an inducer or absence of an inhibitor; the effectors, which may be chemical, physical or developmental in nature, interact with transcription factors that regulate the activity of the associated promoter. Ideally, the level of expression can be controlled by altering the amount of the effector. Inducible systems regulated by small molecules are particularly attractive, as gene expression can be controlled simply by altering the levels of the inducing agent within a cell culture. Several inducible systems have been optimized for use in mammalian cell culture systems and are commercially available.

A popular inducible expression system is based on a prokaryotic tetracycline (Tet)-regulated inducible system, initially described by Gossen & Bujard (1992). Features of this system include lack of pleiotropic effects from the inducing agent, graded response with respect to the inducer concentration, and inducibility of up to 1000-fold from basal expression levels (Gossen et al., 1995). I explored the possibility of using this inducible system, but learned tetracycline and its derivatives are highly fluorescent in multiphoton (MP) fluorescence microscopy, and even trace amounts of Tet can make MP imaging very difficult (Majol et al., 2002). Because I was interested in imaging cells using MP fluorescence microscopy, the Tet system was not suitable for regulating gene expression in my experiments.

An alternative commercially available expression system uses an ecdysone-inducible expression cassette adapted from *Drosophila melanogaster* (No et al., 1996). Ecdysone-inducible systems have also shown low basal promoter activity and high inducibility (Wakita et al., 2001; Wyborski at el., 2001). This system was more suitable because ecdysone and its analogues, which are the inducers, do not interfere with fluorescence microscopy imaging. Furthermore, an ecdysone-inducible expression system was available in a RV format from Stratagene.

The Stratagene Complete Control ® Inducible Mammalian Expression System (Stratagene, 2002) is regulated by the activity of the *trans*-activating receptor proteins VgEcR and RXR. The gene of interest is located downstream of ecdysone-responsive *cis* elements and a minimal promoter. The *trans*-activating receptor proteins bind to the *cis* elements and induce transcription based on the concentration of the effector. The effector molecule is ecdysone or one of its analogues, such as ponasterone A (PonA). In the absence of PonA, the *trans*-activating proteins block transcription by binding to corepressors that repress cellular transcriptional machinery. When PonA is introduced, the corepressors are released and coactivators are recruited, leading to transcription of the gene of interest (Stratagene, 2002). (Figure 4.1)

The Stratagene system is carried on two *Escherichia coli* plasmids containing components of the ecdysone-inducible system in a wild-type Moloney Murine Leukemia Virus (MMuLV) backbone. The viral backbone of one vector contains coding sequences for *trans*-activating proteins, which are constitutively expressed from the cytomegalovirus (CMV) promoter in target cells; the other contains ecdysone responsive *cis* elements, a minimal promoter and a multicloning site that allows for the insertion of a gene of interest. Viral sequences in these

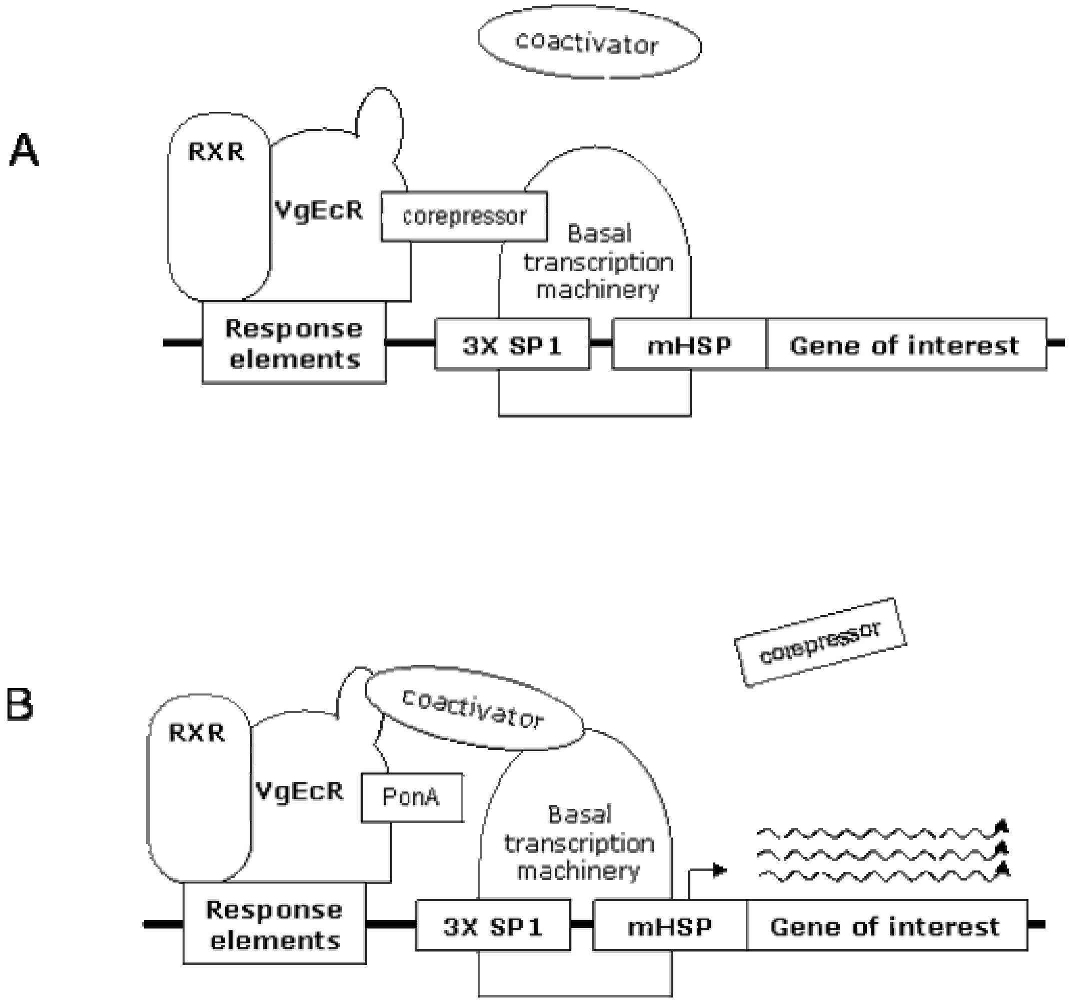

Figure 3.1 Schematic of gene regulation in the ecdysone–inducible expression system.

The *trans*-activating receptor proteins, RXR and VgEcR, bind to the ecdysone response elements upstream of a minimal promoter comprising 3X SP1 sites and a minimal heat shock promoter (mHSP). (A) In the absence of the ecdysone analogue, ponasterone A (PonA), the cellular transcription machinery is repressed. (B) When PonA is present, it binds VgEcR, which releases corepressors and recruits coactivators, thus activating transcription of the gene of interest. (Adapted from Stratagene, 2002).

plasmids are transcribed and packaged into viral particles in a packaging cell line (Stratagene, 2002).

Adapting this expression system for use in ES cells would require some modifications. Wild-type MMuLV-based vectors are not expressed in ES cell lines (Grez et al., 1990). MMuLV contains four *cis*-acting elements that act as strong silencers in mouse ES cells: the negative control region, the direct repeat element, the primer-binding site and a 100 bp region of the viral long terminal repeat (LTR) sequences (Osborne et al., 1999). Therefore, before using the Stratagene system to deliver transgenes to cells, the silencing elements would have to be removed from the viral backbone to allow for efficient expression of transgenes, as in Osborne et al. (1999).

A second consideration is that the CMV promoter has been shown to have low levels of activity in ES cells (Chung et al., 2002; Zeng et al., 2003). For efficient expression of *trans*-activating proteins, the promoter driving their expression would have to be changed to an ES cell active promoter.

The two-vector nature of the Stratagene system has additional implications for the delivery and control of inducible genes by retroviruses. All the elements necessary for inducible expression cannot be delivered in a single retroviral vector because they would exceed the 10 kb packaging limit of MMuLV (Coffin et al., 1997). Because components of this inducible system are in two different vectors, inducible expression requires the delivery of both a control vector (containing the *trans*-activating proteins) and the response vector (containing the gene of interest and *cis* elements) to cells. However, it is technically impossible to ensure delivery of precisely one copy of a cDNA–GFP fusion along with one copy of the *trans*-activating plasmid to each cell in a transfection.

Even if this level of control over delivery were possible, clones would vary markedly in inducibility due to positional effects of retrovirus integration, resulting in poor induction or high levels of uninduced expression (Stratagene, 2002). In fact, well-regulated inducibility appears to be the exception rather than the rule, and occurs in a minority of double-transfected cells (Blau & Rossi, 1999). Clearly this is not practical for use with a cDNA–GFP fusion library. Given the requirement for screening and validating each individual clone for its inducibility, this is not a high-throughput approach and cannot provide the level of control over gene expression that is desired.

Although the initial experimental design involved the use of an inducible system, I ultimately concluded this was not feasible and the plan was abandoned. Instead, I used HSC1, a MMuLV-based retroviral vector that has been optimized for delivery and expression of transgenes in mouse stem cells. It contains mutations in all four known retroviral silencing elements present in the wild-type MMuLV, and is at least 150-fold more effective at directing gene expression in ES cells than is MMuLV (Osborne et al., 1999). The vector HSC1-PGK-eGFP was provided by a Stem Cell Network collaborator, Dr. James Ellis, from the University of Toronto. In this vector the transcription of an enhanced GFP (eGFP) variant is directed by the human phosphoglycerate kinase (PGK) promoter (Yao et al., 2003). To obtain expression levels appropriate for imaging GFP-tagged protein localization, Dr Ellis recommended exchanging the promoter for the stronger human elongation factor 1 alpha (EF-1α) promoter. The use a constitutive promoter does not circumvent the problems associated with the overexpression of cDNA–GFP fusion proteins. However, the use of a single plasmid simplifies transfection and screening. To construct a vector appropriate for a cDNA–GFP fusion library, further vector sequence considerations were to include restriction sites to accept cDNAs upstream of the gene for GFP. Also, the start codon may be removed from the GFP gene to reduce the occurrence of GFP expression from empty vectors that may be introduced into cells.

This chapter outlines the process of making a HSC1 derivative vector for the expression of cDNA–GFP fusions. This RV vector can accept inserts with *Sal*I/*Not*I ends in fusion with GFP and express them under the control of the EF-1α promoter

.3.2 Materials and Methods

3.2.1 Removal of the PGK Promoter from HSC1-PGK-eGFP

Supercoiled HSC1-PGK-eGFP (10 µg) was digested with 20 u *Nco*I (Fermentas) in a volume of 100 µl in 1 X Tango+ Buffer (Fermentas) for 5 h. Following digestion, *Nco*I was heat inactivated at 65°C for 20 min and the DNA was desalted by ethanol precipitation. The DNA pellet was resuspended in 5 µl Buffer 2 (New England Biolabs), 44 µl sterile MilliQ dH_20 and 1 µl MBN (New England Biolabs). The reaction was incubated at 30°C for 30 min. Following this, DNA was extracted using phenol: chloroform and concentrated by ethanol precipitation. The DNA pellet was resuspended in 5 µl 10 X *Eco*RI Buffer (Fermentas), 44 µl sterile MilliQ dH_20 and 1 µl *Eco*RI (10 u/µl; Fermentas) and incubated at 37°C for 4 h.

Digested DNA was electrophoresed on a 1.0% (w/v) agarose gel in 1 X TAE in the presence of ethidium bromide. The 5.6 kb fragment from the digest was excised from the gel and recovered in 10 µl dH_20 using GeneClean (QBiogene) according to the manufacturer's protocol. Purified, linear vector was electrophoresed on a 1.0 % (w/v) agarose gel alongside a λ DNA/*Hin*dIII marker (Fermentas) and quantified based on relative UV fluorescence intensity.

3.2.2 Preparation of EF-1α Promoter for Ligation

The EF-1α promoter (GenBank accession number E02627) was supplied by Dr. James Ellis in plasmid KA436. PCR amplification was done using primers shown in Figure 3.2.

PCR was performed in triplicate using 5 µl 10 X Optimize EXT buffer (Finnzymes), 1 µl 25 pmol/µl forward primer, 1 µl 25 pmol/µl reverse primer, 1 µl 10 mM dNTPs (Roche), 0.5 µl DyNAzyme EXT (1 u/µl; Finnzymes), 1 µl KA436 PCR template and MilliQ dH_20 to 50 µl. Thermal cycling was carried out in an Eppendorf Master Cycler using the parameters: 92°C for 5 min, 5 cycles of 95°C for 45 sec, 53°C for 45 sec, 72°C 1 min 30 sec, 30 cycles of 92°C for 45 sec, 67°C for 45 sec, 72°C 1 min 30 sec, and upon completion, samples were held at 72°C for 10 min before cooling to 10°C.

Forward primer:

5’ GCG AAT TC*G GCC GCT CTA GAC AAT TGG*
***Eco*RI**

Reverse Primer:

5’ CTG CGG CCG CAC GTC GAC *GCT AAT TCC TCA CGA CAC C*
***Not*I** ***Sal*I**

Figure 3.2 PCR primers used in polymerase chain reaction of EF-1α promoter from plasmid KA436.

Regions of the primers complementary to the promoter are designated in italics, and restriction sites within the primers are underlined. The *Sal*I and *Not*I restriction sites were incorporated into these primers to allow for the creation of N-terminal cDNA–GFP fusions.

Aliquots of PCR products were electrophoresed on a 1.0 (w/v) % agarose gel as described in 2.2.1. Upon verifying a successful PCR, the products were pooled and purified using the Qiagen PCR Purification kit (Qiagen) according to manufacturer’s protocol, and eluted in 50 µl of sterile MilliQ dH_20.

EF-1α PCR products were kinased in a 60 µl reaction containing 6 µl 10 X Reaction Buffer A (Fermentas), 60 pmol adenosine triphosphate (ATP) (Roche) and 30 u T4 Polynucleotide Kinase (Fermentas). Following incubation at 37°C for 30 min, the kinase was heat inactivated at 70°C for 10 min and the DNA was recovered by ethanol precipitation. Kinased DNA was digested in a mixture of 5 µl 10 X *Eco*RI Buffer (Fermentas), 44 µl sterile MilliQ dH_20 and 1 µl *Eco*RI (10 u/µl) for 4 h. The DNA was purified by phenol: chloroform extraction, ethanol precipitated and resuspended in 10 µl sterile MilliQ dH_20.

3.2.3 Ligation of EF-1α Promoter into Linearized HSC1

The *Eco*RI digested and kinased PCR product containing the EF-1α promoter was inserted into the HSC1 vector. The ligation reaction contained 100 ng of *Nco*I/MBN/*Eco*RI digested HSC1, 60 ng of *Eco*RI digested and kinased EF-1α promoter, 4 µl 5 X ligase buffer (Gibco BRL), 1 µl T4 DNA ligase (10 u/µl, Fermentas) and MilliQ dH dH_2O to 20 µl. The molar ratio of insert to vector was 3:1.

Ligations were incubated at 16°C overnight. Following incubation, the products were ethanol precipitated. The DNA pellet was dried in a Speedvac until no traces of moisture were visible and the ligation products were resuspended in 10 µl sterile MilliQ dH_2O.

3.2.4 Electroporation of DH5 α with Ligation Mixtures

Ligation mixtures were electroporated into electrocompetent *E. coli* DH5 α cells. For each electroporation, a 40 µl aliquot of frozen cells was placed on ice and allowed to thaw until just barely liquid. A 5 µl sample of a the ligation mix was added to the cells, mixed gently and quickly transferred to an ice-cold, sterile 2 mm electroporation cuvette (BioRad). Cuvettes were placed in a BioRad Gene Pulser electroporator, and pulsed once with 1.25 V, 200 Ω resistance, 25 µF. Following the pulse, 1 ml SOC media (Sambrook and Russell, 2001), was added to the cells and they were transferred to a sterile 13x100 mm glass culture tube and placed in a 37°C water bath with vigorous shaking. After 1 h, a 20 µl and 200 µl sample of each electroporation culture was plated on solid LB-agar supplemented with 100 µg/ml ampicillin. Plates were incubated at 37°C overnight.

3.2.5 Isolation and Sequence Verification of Vector pBES23

Transformants containing the EF-1α insert were identified by PCR. Colonies were sampled using a toothpick that was lightly streaked onto a fresh LB agar plate supplemented with 100 µg/ml ampicillin before being placed 10 µl of sterile MilliQ dH_2O. One microlitre of this suspension was used as the source of template for PCR using conditions outlined in section 3.2.2 to amplify the promoter fragment.

Clones from which the EF-1α promoter could be amplified were prepared for sequencing using a Qiaprep DNA Plasmid Miniprep kit (Qiagen) according to the manufacturer’s protocol and the promoter and cloning site were sequenced at the University of Waterloo in-house sequencing facility. Primers were 5´ GCG AAT TCG GCC GCT CTA GAC AAT TGG and 5´ CAC GCT GAA CTT GTG GC.

3.3 Results and Discussion

The goal of work presented in this chapter was to design a vector for constructing an N-terminal cDNA–GFP fusion library. The requirements of such a vector were that it be retroviral, contain a viral backbone that would not silence in ES cells, have a promoter to drive expression of cDNA–GFP fusions, include restriction sites to allow ligation of cDNAs upstream of the coding sequence for GFP, and for the GFP to have no start codon.

I was successful in constructing the HSC1 derivative RE vector containing the EF-1α promoter and *Sal*I and *Not*I recognition sites, and named this vector pBES23 (Figure 3.3). This vector is suitable for the directional cloning of inserts with *Sal*I/*Not*I ends and can allow for the expression of cDNA fragments as N-terminal fusions with GFP, provided the insert sequences contain an initiation codon and no stop codon. In this vector, the coding sequence for GFP does not have a Met initiation codon. This was achieved by subjecting HSC1 linearized by *Nco*I to a MBN digest, thereby removing the protruding ends containing the ATG codon.

A further improvement to the current vector would be to modify the coding sequence between the *Not*I site and the GFP coding sequence by removing one base pair. This is because in the current vector, the codon created at the juncture of the cDNA and the *Not*I site is NNG, where N is any nucleotide (Figure 3.2 B). When these two nucleotides are TA, a stop codon is created, meaning the cDNA will not be expressed with a GFP tag. The frequency of this occurrence is one in sixteen in-frame cDNAs. By removing one base pair, there is no possibility of creating a stop codon at the junction of the cDNA and GFP. At the time I stopped working to write this thesis, I was designing a more suitable expression vector for cDNA GFP-fusion proteins.

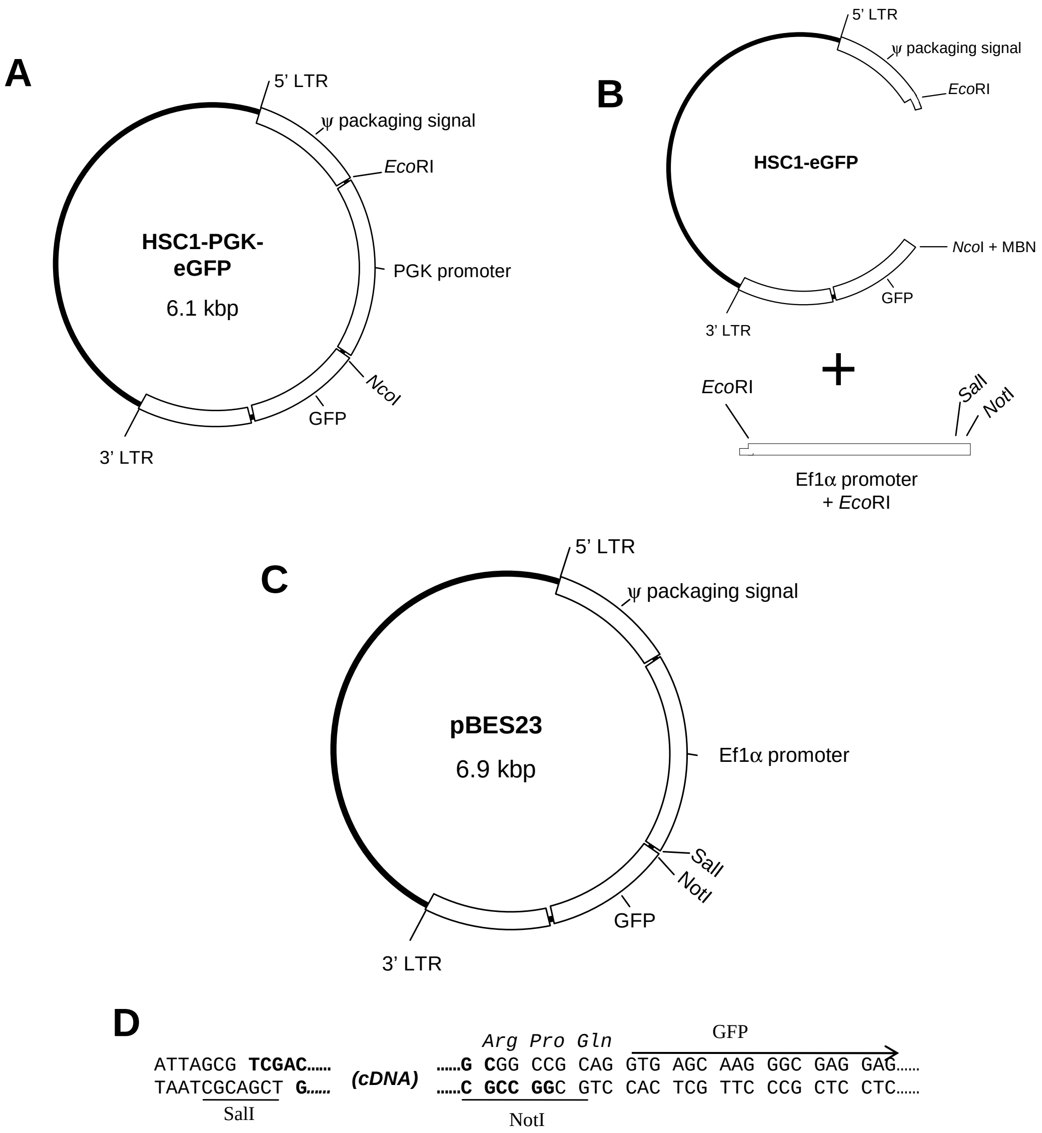

Figure 3.3 Map of retroviral expression vector pBES23.

The vector pBES23 was constructed from the vector HSC1-PGK-eGFP (A). The PGK promoter was removed by digesting HSC1-PGK-eGFP with *Nco*I, MBN and *Eco*RI, and the EF1α promoter was generated by PCR and prepared for ligation by digesting a with *Eco*RI (B). The retroviral expression vector created following ligation of the EF1α promoter (pBES23) can accept inserts with *Sal*I/*Not*I ends, and allows such inserts to be expressed as in fusion with GFP from the EF1α promoter (C). A schematic of the vector multiconing site is shown in (D). The predicted amino acid sequence of the region between a cDNA insert and GFP are shown in italics, and restriction enzyme recognition sites are underlined.

Chapter 4 Constructing a cDNA–GFP Fusion Library

4.1 Introduction

The initial design of this project involved creating a library of truncated cDNAs fused GFP at their C-termini. However, the approach I devised for truncating cDNA proved impractical for use with a cDNA library because of poor recovery of ds DNA and a bias towards shorter ds DNA molecules. Instead of using this novel approach, I generated a random-primed cDNA population using a commercial cDNA library kit. The SuperScript Plasmid System for cDNA Synthesis and Cloning (Invitrogen) is designed for making full-length cDNA libraries in a plasmid cloning vector, but could be adapted to generate an N-terminal cDNA–GFP fusion library. The two main procedural modifications required were substituting a random hexamer primer–adapter for an oligo-dT primer–adapter in first-strand cDNA synthesis, and cloning cDNAs into the vector pBES23 instead of the vector included in the kit.

A random hexamer primer–adapter allows one to generate random-primed cDNAs that can be directionally cloned; this is made possible by including the sequence for a restriction enzyme recognition site in the primer used for first strand cDNA synthesis. Following second-strand synthesis and addition of hemi-phosphorylated adapters, digestion of the cDNA population with the enzyme whose restriction site is within in the primer–adapter results in cDNAs with different protruding 5´ and 3´ ends. Using a restriction site that occurs infrequently in DNA, such as *Not*I, minimizes the chance of the restriction enzyme cleaving within the cDNA. The presence of a restriction enzyme site necessitates the addition of extra bases to the primer–adapter. Because the restriction enzyme site is palindromic, primer–adapters will have a higher affinity for each other than for mRNA. This may be offset by the addition of terminal non-complementary nucleic acids to the 5´ end of the primer, such as terminal GA repeats (Das et al., 2001). While it may not seem that a long primer with a relatively short random hexamer sequence can be used to generate random-primed cDNA, a primer 39 bases in length containing a *Not*I sequence and just six random nucleotides was successfully used to generate cDNA for an N-terminal cDNA–GFP fusion library (Escobar et al., 2003).

The use of random hexamer primer–adapters requires careful control of the reaction conditions for first-strand cDNA synthesis. If two or more primers anneal to a single mRNA, multiple *Not*I restriction sites will be incorporated into the cDNA. Digestion of the cDNA with *Not*I will generate multiple small cDNA fragments. It is therefore essential to minimize the ratio of primers to template in the first strand reaction such that, on average, one primer anneals to each mRNA. After *Not*I digestion, short cDNAs resulting from priming close to the 5´ end of an mRNA or from multiple priming events can be removed by size fractionation.

Following these modifications to the protocol, I constructed an N-terminal cDNA library with an average cDNA insert size of 1.1 kb, and transfected this library into a Phoenix retroviral packaging cell line.

4.2 Materials and Methods

4.2.1 Embryonic Stem Cell Culture

R1 mouse embryonic stem cells were supplied by Dr. James Ellis of the University of Toronto. Cells from passage 11 were grown on gelatin-coated T-75 culture flasks (FALCON) in ES media (Dulbecco's Modified Essential Medium (DMEM-Gibco) supplemented with 2 mM L-glutamine (Gibco), 100 mM non-essential amino acids (Gibco), 100 µM 2-mercaptoethanol (Gibco), 15% (v/v) ES-certified fetal calf serum (FCS) (Gibco), and 1000 u/ml leukemia inhibitory factor (Stem Cell Technologies) in 5% CO_2 at 37° C. Every day cells were supplied with fresh media, and when cultures were approximately 70% confluent (≅ 2–3 days) the media was removed and cells were trypsinized with 0.25% (w/v) trypsin-EDTA (Gibco BRL) for 5 min. Following this, fresh ES media was added to the culture flask and the cell suspension was triturated, transferred into a sterile 15 ml polystyrene tube (Falcon) and centrifuged at 300 x g for 5 min. The supernatant was removed and cells were resuspended in fresh ES media. The cells were passaged at a 1:5 dilution into fresh gelatin-coated T-75 flasks.

4.2.2 Total RNA Isolation

Total RNA was isolated from 2 T-75 flasks (≅ $4x10^7$ ES cells) using Tripure (Roche) according the manufacturer's protocol. RNA was resuspended in diethylpyrocarbonate (DEPC)-treated dH_2O and the concentration of nucleic acid was determined by absorbance spectroscopy at 260 and 280 nm. The yield of nucleic acid was 640 µg. An aliquot of total RNA was electrophoresed on a 1.2% (w/v) formaldehyde agarose gel in the presence of ethidium bromide to evaluate its integrity. RNA was stored at -70° C.

4.2.3 Poly-A+ RNA Purification

Poly-A+ RNA was isolated using an Oligotex mRNA Midi kit (Qiagen) according to the manufacturer's instructions with the following two exceptions. To maximize elution of poly-A+ RNA from the Oligotex beads, once the 70°C elution buffer OEB was added to the beads, samples were incubated at 70°C in a heating block for 1 min prior to centrifugation. The elution step was repeated and the recovered poly-A+ RNA was pooled in a single microcentrifuge tube, ethanol precipitated, and resuspended in 20 µl of DEPC-treated dH_2O. A one microlitre sample was used to assess RNA quantity and purity by spectrophotometry. The RNA was stored at -70°C.

4.2.4 RT–PCR of Poly-A+ RNA

Prior to constructing a cDNA library, a portion of the poly-A+ RNA was used for RT–PCR for three transcripts expressed in mouse stem cells. cDNA was synthesized in a reaction containing 1 µl anchored oligo-dT primer (5´T_{18}VN, where V is A, C or G), 100 ng poly-A+ RNA and DEPC-treated dH_2O to 11 µl. The solution was incubated 65°C in a heating block for 5 min then chilled on ice. To this sample, 4 µl 5 X MMuLV Reverse Transcriptase (RT) Buffer (Fermentas), 2 µl 10 mM dNTPs (Fermentas), and 1 µl RNase inhibitor 20 u/µl (Fermentas) were added and the entire reaction was incubated at 37 °C for 2 min. Following this, 2 µl MMuLV RT (20 u/µl; Fermentas) were added and the reaction was incubated 37°C for 1 h. Subsequently the RT was inactivated by heating at 95°C for 5 min.

PCR was carried out in a volume of 25 µl in the presence of 2.5 µl 10 X PCR buffer (500 mM KCl, 100 mM Tris-Cl pH 8.3, 15 mM $MgCl_2$), 1 µl 25 pmol/µl forward primer, 1 µl 25 pmol/µl reverse primer, 0.5 µl 10 mM dNTPs (Roche), 1 µl *Taq* (homemade recombinant stock), 1 µl cDNA template from the above reaction and sterile MilliQ dH_20. Primer pairs used for PCR are outlined in Table 4.1. Thermal cycling was carried out in an Eppendorf Master Cycler using the parameters: 92°C for 5 min, then 35 cycles of 92°C for 45 sec, 55°C for 45 sec, and 72°C for 1 min. Following this, samples were held at 10°C. Eight microlitres of each reaction were electrophoresed on a 1.8% (w/v) agarose gel in the presence of ethidium bromide alongside a Generuler 100 bp DNA Ladder (Fermentas).

Table 4.1 Primers used for RT–PCR reactions of murine poly-A+ RNA and the expected size of PCR products

Target sequence	Primer sequence 5´→3´	Expected size of PCR product	GenBank Accession Number
integrin α–6	forward 5´CCC AAG GAG ATT AGC AAT GG reverse 5´CCT GGA ACG AAG AAC GAG AG	300 bp	NM_008397
β–actin	forward 5´GAC AAC GGC TCC GGC ATG TG reverse 5´CAT TGT AGA AGG TGT GGT GC	247 bp	NM_007393
Hypoxanthine guanine phosphoribosyl transferase	forward 5´CTC GAA GTG TTG GAT ACA GG reverse 5´TGG CCT ATA GGC TCA TAG TG	350 bp	NM_013556

4.2.5 Preparation of pBES23 for Ligation

Approximately 5 µg of pBES23 were digested in a total volume of 40 µl containing 4 µl 10 X Buffer O+ (Fermentas) and 1 µl *Sal*I (10 u/µl; Fermentas). The reaction was incubated at 37°C overnight. Following this, the entire restriction digest was electrophoresed through a 1.0 % (w/v) agarose gel in the presence of ethidium bromide alongside a GeneRuler DNA Ladder Mix (Fermentas) and supercoiled pBES23. The *Sal*I digested vector, identified as a 6.9 kb DNA fragment, was excised from the gel and purified using GeneClean (QBiogene) according to the manufacturer's protocol. DNA was eluted in 30 µl of sterile MilliQ dH$_2$0. To ensure that all traces of Glassmilk were removed, the eluted DNA was placed in a clean microcentrifuge tube and centrifuged at 16.1 x g for 1 min. The supernatant containing DNA was carefully removed and placed in a clean microcentrifuge tube. A second digest was performed using 20 µl of this *Sal*I digested vector, 2.5 µl 10 X Buffer 0+ (Fermentas), 0.5 µl *Not*I (10 u/µl; Fermentas) and sterile MilliQ dH$_2$0 to 25 µl. The reaction was incubated at 37°C for 2 h, then the *Not*I was heat inactivated at 65°C for 20 min.

4.2.6 Self-Ligation Reactions of pBES23

To gauge the efficiency of the *Not*I digest, two vector self-ligation reactions were performed: one using pBES23 digested with *Sal*I, and the other using pBES23 digested with *Sal*I and *Not*I. Reaction conditions for each ligation are shown in Table 4.2.

Table 4.2 Reaction conditions and reagent volumes for self-ligation reactions of linearized pBES23.

Ligation component	*Sal*I digested vector ligation	Sal*I*/Not*I* digested vector ligation
DNA	1 µl	1.25 µl*
5X T4 DNA ligase buffer (Gibco)	4 µl	4 µl
Sterile MilliQ dH_20	14 µl	13.75 µl
T4 DNA ligase 10 u/µl (Fermentas)	1 µl	1 µl
Total volume	20 µl	20 µl

*The volume has been increased by 25 % because the concentration of double cut DNA was reduced by 20% in the *Not*I digestion step. This ensures that both reactions have equal amounts of vector DNA.

Ligations were incubated overnight at 16°C. Following this, the DNA was precipitated, resuspended and transformed into electrocompetent *E. coli* DH5α as described in 3.2.4, with the exception that 5 µl and 10 µl samples of each transformation culture were plated onto individual LB-agar plates supplemented with 100 µg/ml ampicillin. Plates were incubated overnight at 37 °C.

4.2.7 cDNA Library Construction

Unless otherwise noted, all reaction components used in cDNA library synthesis were from a Superscript Plasmid System for cDNA Synthesis and Cloning (Invitrogen).

First Strand cDNA Synthesis

Two micrograms of mRNA were mixed with 10 pmol of a *Not*I-N_6 primer–adapter (5´$(GA)_{10}$GCGGCCGCNNNNNN where N is G, A, C or T; Sigma Aldrich) and DEPC dH_2O to 11 µl. This sample was heated at 70°C for 10 min, and then chilled on ice. To this, 4 µl 5 X First Strand Buffer, 2 µl 0.1 M dithiothreitol (DTT) and 1µl 10 mM dNTPs were added. The reaction was incubated at 42°C for 2 min, then 2 µl of Superscript II were added. The reaction was incubated at 42°C for 1 h, after which it was immediately placed on ice.

Second Strand cDNA Synthesis

To the first strand reaction, 90 µl DEPC dH_2O, 30 µl 5 X Second Strand Buffer, 3 µl 10 mM dNTPs, 1 µl *E. coli* DNA ligase, 4 µl *E. coli* DNA polymerase I, 1 µl *E. coli* RNase H, and 1 µl ^{32}P-dCTP (10 mCi/ml, 3000 mCi/mmol; Amersham Pharmacia) were added. This reaction was incubated for 2 h at 16°C. Following this, 5 µl T4 DNA polymerase were added and the 16°C incubation was continued for a further 5 min. The reaction was terminated by the addition of 10 µl 0.5 M EDTA. A 2 µl sample was removed and placed in a tube containing 43 µl 0.02 M EDTA, and 5 µl yeast tRNA. This sample was used to calculate the specific activity of ds cDNA as described in the manufacturer's protocol. The remaining cDNA was purified by adding 150 µl phenol: chloroform, mixing thoroughly and centrifuging for 5 min. The aqueous phase was removed and placed in a fresh microcentrifuge tube and mixed well with 75 µl 7.5 M ammonium acetate and 500 µl 100% ethanol. This was centrifuged for 20 min. The supernatant was removed and the cDNA pellet was washed with 500 µl 70% ethanol and centrifuged for 5 min.

*Sal*I *Adapter Ligation*

The cDNA pellet was resuspended on ice in 29 µl dH_20, 10 µl 5 X T4 DNA ligase buffer, 1 µl T4 DNA ligase (400 u/µl; New England Biolabs), and 10 µl *Sal*I adapters. This ligation reaction was mixed thoroughly and incubated at 16°C for 16 h. Following this, the reaction was stopped by extraction with phenol: chloroform and the DNA was ethanol precipitated.

*Not*I *Digestion of cDNA*

cDNA was resuspended in 5 µl 10 X Buffer 3 (New England Biolabs), supplemented with 100 µg/ml bovine serum albumin (New England Biolabs) and sterile MilliQ dH_20 to 44 µl. Six microlitres of *Not*I (10 u/µl; New England Biolabs) were added and the reaction was incubated at 37°C for 2 h, before extraction with phenol: chloroform followed by ethanol precipitation.

Column Chromatography for Size Fractionation of cDNAs

One hundred microlitres of TEN buffer (10 mM Tris-Cl, pH 7.5, 0.1 mM EDTA, 25 mM NaCl) were added to the cDNA pellet and it was allowed to rehydrate on ice. Meanwhile, a cDNA Size Fractionation Column (Invitrogen) was prepared for use according to the manufacturer’s directions. Following column equilibration, the resuspended cDNA was applied to the column and allowed to drain into the column bed. Single-drop fractions of column flow-through were collected and their volume measured according to the manufacturer's protocol. Cerenkov counts were obtained using the ^{32}P channel of a Wallac Microbeta Trilux Liquid Scintillation and Luminescence Counter 1450-023. Based on the Cerenkov counts, the fractions that contained cDNA were eluted within the first 550 µl of column flow-through were pooled and ethanol precipitated. The total amount of cDNA recovered was calculated based on knowledge of the specific activity of the cDNA as determined in second-strand synthesis.

cDNA Ligation and Library Transformation

The precipitated cDNA (≅ 15.6 ng) was resuspended in 6.1 µl sterile MilliQ dH_20, 4.0 µl 5 X T4 DNA ligase buffer (Gibco) and 8.9 µl of *Sal*I/*Not*I digested pBES23 (12.3 ng/µl). The cDNA was allowed to rehydrate on ice, then 1 µl T4 DNA ligase (10 u/µl; Fermentas) was added and the ligation reaction was incubated at 4°C overnight. Three microlitres of this reaction were used for transformations of XL-10 Gold Ultracompetent Cells (Stratagene) according to the manufacturer’s protocol. To gauge the transformation efficiency and library complexity, 5 µl of each transformation mixture were plated onto individual 60 mm LB agar plates supplemented with 100 µg/ml ampicillin. The unplated portion of each transformation sample was stored at 4°C.

4.2.8 PCR Analysis of cDNA Library Clones

To determine the approximate percentage of library clones containing cDNA inserts, PCR analysis was performed on 50 randomly selected colonies. Colonies were streaked onto a plate and PCR templates were prepared as described in 3.2.5. PCR conditions were identical to those in section 3.2.2, with the exception that the forward primer was 5´CTC AAG CCT CAG ACA GTG G and the reverse primer was 5´CTT GTA CAG CTC GTC CAT GC. Aliquots of each PCR were electrophoresed on a 1.0% (w/v) agarose gel in the presence of ethidium bromide alongside a GeneRuler Ladder Mix (Fermentas).

To estimate the average size of clones containing an insert, colonies determined to have an insert based on the results of the first round of PCR were used as a template in a second round of PCR. Reaction conditions were identical to the first round, with the exception that the reverse primer was 5´CAC GCT GAA CTT GTG GC. Aliquots of PCR were electrophoresed on a 1.0% (w/v) agarose gel in the presence of ethidium bromide alongside a GeneRuler Ladder Mix (Fermentas). Insert sizes were estimated based on the distance migrated in the gel relative to a standard curve constructed using the DNA fragments of known length in the reference marker.

4.2.9 Preparation of cDNA Library for Cell Transfection

The cDNA library was amplified by plating the remainder of each transformation on 2 X 100 mm LB-agar plates supplemented with 100 µg/ml ampicillin and incubating overnight at 37°C. Following this, the plates were flooded with 5 ml LB and the colonies were scraped off the surface of the agar using a sterile rubber policeman and transferred into a sterile 50 ml centrifuge tube (Falcon). Plasmid DNA was isolated from the bacterial cells using a Plasmid Midiprep kit (Qiagen) according to the manufacturer's protocol. Following ethanol precipitation, plasmids were resuspended in 20 µl sterile 10 mM Tris-Cl pH 7.5. One microlitre was used to assess plasmid quantity by spectrophotometry, and the remaining plasmid DNA was diluted to a concentration of approximately 1 µg/µl.

4.2.10 Cell Culture and Transfection

Cell Culture Conditions

Phoenix retroviral packaging cells (Kinsella and Nolan, 1996) were provided by Dr. James Ellis of the University of Toronto. Cells were cultured in Phoenix cell media, (DMEM supplemented with 10% (v/v) FCS (Gibco) and 4 mM L-glutamine (Gibco)) and incubated at 37°C in 5% $C0_2$. When cells reached 70-80% confluence(≅2 days) the media was removed and cells were treated with 0.25 % (w/v) trypsin-EDTA (Gibco) for 5 min, then passaged 1:5 in Phoenix cell media. Twenty-four hours prior to transfection, 70-80% confluent cells were trypsinized in the same manner as in the passaging protocol, then plated onto fresh 60 mm tissue culture treated dishes at a 1:2 dilution in Phoenix Cell media.

NIH–3T3 cells were obtained from the American Type Culture Collection (ATCC) and cultured in DMEM supplemented with 4 mM L-glutamine, 1.5 g/L sodium bicarbonate, 4.5 g/L glucose, and 10 % (v/v) bovine calf serum. Cells were trypsinized and passaged twice per week according to the ATCC protocols, and were never allowed to reach confluence.

Calcium–Phosphate Transfection of cDNA–GFP Library into Phoenix Cells

For each transfection, a 60 mm tissue culture dish of Phoenix cells was prepared as described above. Five min prior to transfection, the media was changed to fresh cell growth media containing 25 µM chloroquine diphosphate (Sigma). Meanwhile, 10 µg of the cDNA library was diluted in sterile MilliQ dH_20 to a final volume of 439 µl, then 61 µl of sterile 2 M $CaCl_2$ was added. Working quickly, 500 µl of 2 X Hepes Buffered Saline (50 mM HEPES, 10 mM KCl, 12 mM dextrose, 280 mM NaCl, 1.5 mM Na_2HPO_4, pH 7.05) was added and the solution was bubbled for exactly 10 sec using an automatic pipettor. The DNA solution was added dropwise to cells and plates were gently rocked to ensure the DNA was distributed throughout the cell culture. Following 16 h incubation, the media was exchanged for fresh cell culture media and cells were returned to the 37°C growth chamber for a further 32 h.

Retroviral Harvesting

Following 48 h of incubation, retrovirus was recovered from the transfected Phoenix cell culture supernatant by passing media removed from the cells through a sterile 45 µm filter. The filtrate was aliquoted into 1 ml fractions. Fractions to be used for NIH–3T3 transduction were stored on ice. All other aliquots were immediately frozen at -70°C in dry ice and stored at -70°C. To maintain viability for fluorescence microscopy, the Phoenix cells were supplied with fresh growth media.

NIH–3T3 Transduction

Twenty-four hours prior to viral transduction, NIH–3T3 cells were trypsinized and plated onto 60 mm tissue culture dishes at a 10% density so they would be actively dividing when the retrovirus was introduced. To perform transductions, growth media was removed from the cells and they were supplied with 2 ml fresh media along with 3 µl of 4 mg/ml polybrene (Sigma) and 1 ml retroviral supernatant. Cells were placed in a 37°C incubator for 24 h. Following this, media was replaced with 3 ml fresh media and cells were incubated for a further 24–48 h before imaging.

4.2.11 Microscopy and Cell Imaging

Microscopy was performed on a Zeiss Axiovert 200 microscope through a 10x objective lens (Carl Zeiss Inc.) For fluorescence imaging, a HBO 50 mercury vapour lamp was employed as a light source using a 500/20 nm excitation filter, 515 nm dichroic and a 535/30 nm emission filter (Chroma Technology). All digital images were captured using a Sony XCD-SX910 CCD camera (Sony) that was controlled by IEEE-1394 Digital Camera Windows Driver imaging software.

4.3 Results

The experiments outlined in this chapter describe the steps to create a random-primed cDNA–GFP fusion library in the retroviral expression vector pBES23. The library was constructed from murine ES cell poly-A+ RNA using a Superscript Plasmid System for cDNA Synthesis and Cloning, modified so cDNA synthesis was initiated from random hexamer primer–adapters. Prior to constructing this library, the quality of the the poly-A+ RNA and the vector into which cDNAs were ligated, were assessed, as these two components are critical to the success of a library. Once the library was constructed, it was delivered into Phoenix retroviral packaging cells.

4.3.1 RT–PCR of Poly-A+ RNA

To evaluate the poly-A+ RNA to be used in making a library, 100 ng were used as a template for RT–PCR of three transcripts constitutively expressed in ES cells: β-actin, hypoxanthine guanine phosphoribosyl transferase (HPRT), and integrin α-6 (ITG-α6). β-actin is a transcript constitutively expressed throughout development. The HPRT and ITG-α6 mRNAs were selected because they have been identified as being upregulated in murine ES cells (Ramhalo-Santos et al., 2002). The PCR primers were designed to target different portions of each mRNA: the 3´ end of β-actin, the central portion of HPRT, and portion of the ~5 kb ITG-α6 mRNA that would require reverse transcription of at least 2kb of the mRNA. Aliquots of reactions were electrophoresed on an agarose gel (Figure 4.1).

The PCRs were successful for all transcripts, and all PCR products were of expected size (see Table 4.1). The mRNA was determined to be suitable for making a cDNA–GFP fusion library.

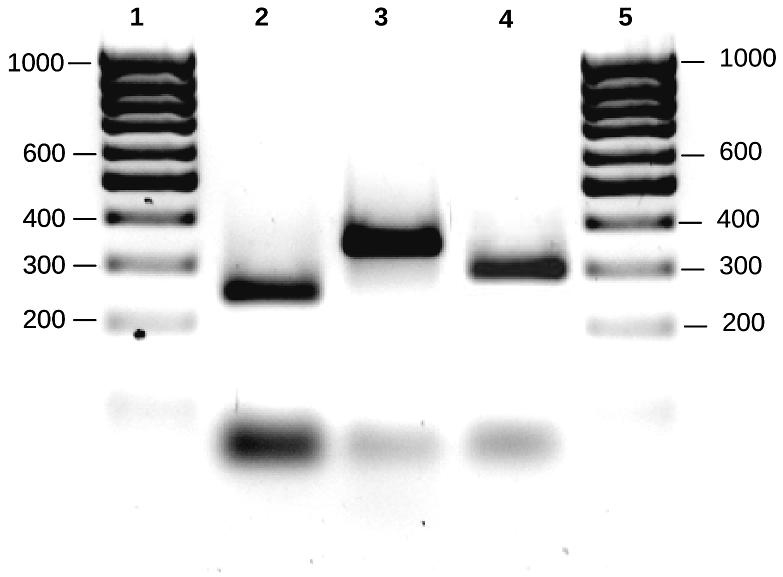

Figure 4.1 RT–PCR of mRNA used as template for constructing a cDNA library.

An aliquot of the mRNA prepared for cDNA–GFP library synthesis was used in RT–PCR for three transcripts constitutively expressed in mouse ES cells. Following PCR, 8 µl of each reaction were electrophoresed on a 1.8% (w/v) agarose gel supplemented with ethidium bromide alongside 2.5 µg of a GeneRuler 100 bp size marker (lanes 1 and 5). PCR shown were amplified from regions of β-actin (lane 2), HPRT (lane 3), and ITG-α6 (lane 4) mRNAs.

4.3.2 Assessment of the Efficiency of Vector Digestion

The cDNA was ligated into *Sal*I and *Not*I sites of the vector pBES23. Since cDNA is a limited resource, and efficient ligation depends on the vector being double cut, it was important to verify the vector's status prior to using it in a cDNA ligation reaction. The extent of *Not*I digestion was determined by comparing the number of transformants obtained from self-ligation reactions containing 100 ng of either *Sal*I or *Sal*I/*Not*I digested pBES23. These results are shown in Table 4.3.

Table 4.3 Number of colonies obtained on LB agar plates following transformation of self-ligation reactions of a single cut and a double cut vector.

Volume of transformation plated	Colonies from ligation of pBES23 *Sal*I	Colonies from ligation of pBES23 *Sal*I/*Not*I
5 µl	965	79
10 µl	TNTC	145

TNTC=too numerous to count

The large decrease in the number of self-ligations following *Not*I digestion suggests that approximately 92% of the vector DNA molecules were cut with *Not*I. While this is less preferable than 100% digestion, subjecting the *Sal*I/*Not*I digested DNA to an additional incubation with *Not*I and repeating the ligation test did not result more than 92% of vector molecules being double cut (data not shown). This 92% double-digested vector was used for constructing a cDNA–GFP fusion library.

4.3.3 cDNA Library Analysis

A cDNA library was constructed in the vector pBES23 using poly-A+ RNA isolated from exponentially growing R1 ES cells. cDNA was made using a SuperScript Plasmid System cDNA Library Synthesis Kit (Invitrogen) modified with a random hexamer primer–adapter. One third of the library ligation was transformed into ultracompetent *E. coli*. Because there is a limit to the volume of DNA one can add to a transformation reaction, two identical transformation reactions were performed to increase the number of clones obtained. The transformations had comparable efficiencies, and based on the number of coloniess obtained in a 5 µl aliquot of each 1 ml transformation reaction (334 and 263), the total number of transformants was estimated to be 1.2 x 10^5.

To estimate the percentage of clones containing a cDNA insert and the average insert size, randomly selected colonies were analyzed by PCR. Because colony PCR can be unreliable, colonies were screened in two stages. To estimate the percentage of clones containing an insert, the first round of PCR was done using a forward primer that was immediately upstream of the cDNA cloning site and a reverse primer complementary to the 3´ end of GFP. Colonies containing plasmids with no insert generated an amplification product approximately the size of the GFP coding sequence (~800 bp). Clones with a cDNA insert gave a PCR product equivalent in size to GFP plus the size of the insert. Analysis of the size of PCR products generated in this preliminary screen of 50 colonies indicated that approximately 60% of clones from the ligation reaction contained cDNA inserts (data not shown).

To estimate the molecular size of cDNA inserts more accurately, colonies found to have an insert were subjected to a second round with forward and reverse primers flanking the cDNA cloning site. The results of these PCRs are shown in Figure 4.2. The median size of cDNA inserts was 1.1 kb, with the range from 500 bp to 2.1 kb.

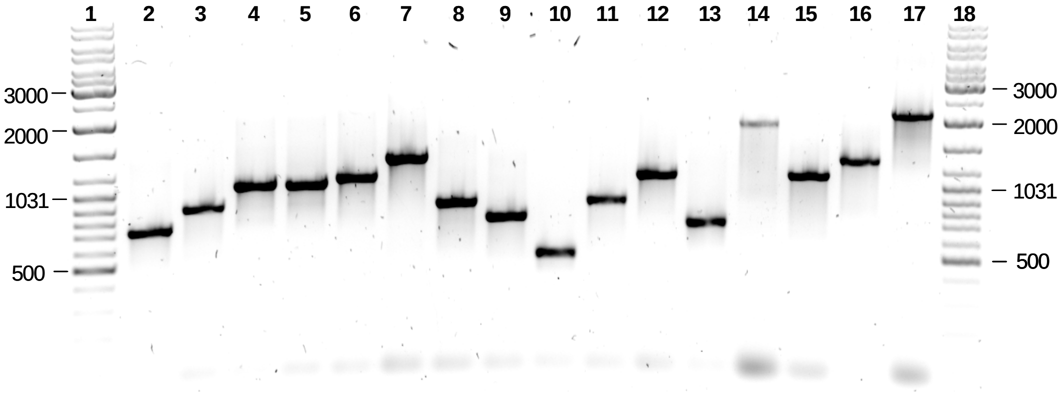

Figure 4.2 PCR analysis of the cDNA inserts in a cDNA–GFP fusion library

All randomly selected cDNA library clones that were confirmed to have an insert were used as a template in a PCR reaction to amplify the insert. An aliquot of each PCR was electrophoresed on a 1.0% (w/v) agarose gel (lanes 2–17), and 2.5 µg of GeneRuler Ladder Mix (Fermentas) were used as reference markers (lanes 1 and 18).

4.3.4 Cell Transfection and Transduction

The cDNA–GFP fusion library was transfected into Phoenix viral packaging cells. The retroviral vector KA436, which is a HSC1 derivative, was used as a positive control in transfections. This vector, which was provided by Dr. James Ellis, contains GFP under the control of the EF-1α promoter in the same RV backbone as pBES23, and was used because the GFP coding sequence in pBES23 lacks an initiation codon. It was possible to use this positive control to estimate the efficiency of Phoenix cell transfection because sequences in RV-based plasmids are translated by packaging cells. In this case, packaging cells that have been successfully transfected will express GFP, and the percentage of cells transfected can be estimated by observing cells using fluorescence microscopy.

Packaging cells transfected with the cDNA–GFP library fluoresced at a level comparable to those transfected with a control positive vector, and the efficiency was estimated to be between 30% and 40% for both transfections (Figure 4.3). In a separate round of transfections, an empty pBES23 vector was transfected and observed to direct expression of fluorescent proteins (data not shown).

Supernatants from the packaging cells were used to transduce NIH–3T3 target cells, a mouse embryonic fibroblast cell line. No cells were observed to fluoresce following any transductions, including the KA436 positive control. Using aliquots of the first packaging reaction, increasing the MOI and starting with a fresh batch of exponentially growing NIH–3T3 cells did not lead to any transduction. A second round of transfection and transduction using the same vectors as in the initial trial plus a negative control, supercoiled pBES23, produced similar results, with no viral transduction of target cells.

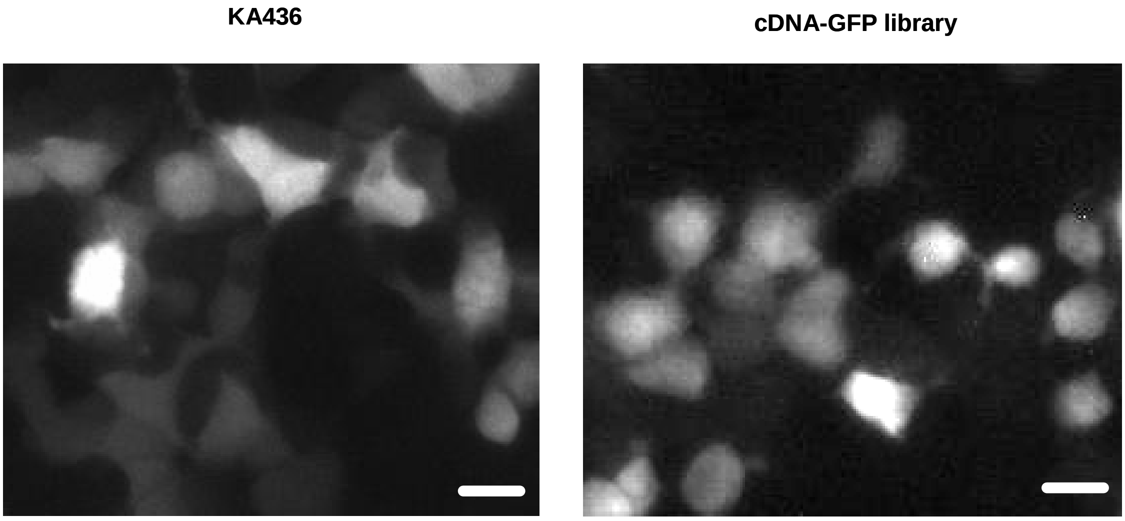

Figure 4.3 Images of Phoenix packaging cells that have been transfected with a GFP-expressing plasmid KA436 and pBES23/R1 cDNA–GFP fusion library.

Cells were transfected using a calcium–phosphate mediated protocol and imaged using fluorescence microscopy. Bar=20µm.

4.4 Discussion

In the experiments outlined in this chapter I constructed a cDNA–GFP fusion library using poly-A+ RNA isolated from exponentially growing undifferentiated R1 ES cells. This cell line was derived from a 3.5-day-old mouse embryo (Nagy et al., 1993), and was chosen because the RNA would be rich in transcripts for proteins that are important in early stem cell development. Because this RNA was a limited resource, prior to constructing a library I first verified that the RNA and expression vector pBES23 were suitable for library construction.

Using a portion of poly-A+ RNA as a template in RT–PCR reactions proved to be a useful way of checking to see if it was degraded. While this test was not comprehensive, the procedure required a minimal amount of poly-A+ RNA (~100 ng) and tested the ability of the mRNA to serve as a template for synthesis of cDNA. If the mRNA was degraded, the cDNA made from it would not contain continuous cDNA molecules of the transcripts present in the parent mRNA population, meaning it could not be used as a template in RT–PCR. Following cDNA synthesis using R1 ES cell poly-A+ RNA as a template, PCR was performed using three sets of primers specific for constitutively expressed transcripts in mouse ES cells. It was possible to amplify transcripts for both the housekeeping gene β-actin and the ES upregulated genes HPRT and ITG-α6. This was interpreted to mean the poly-A+ RNA was of reasonable quality to construct a library.

The second thing I wanted to verify was that a large percentage of the vector DNA was digested with both *Sal*I and *Not*I. Based on results of self-ligation of linearized vectors, 92% of the double-digested vector DNA was cut with both enzymes. Ideally, this number should be 100%. However, attempts to increase this percentage were not successful. A likely explanation is that the two restriction sites are too close to each other to allow efficient restriction enzyme digestion at both sites. Following cleavage with *Sal*I, the *Not*I site is just 3 bp from the end of the linear DNA molecule, and the efficiency of restriction digestion within a few base pairs of the termini of linear DNA molecules can be less than 100% for *Not*I (New England Biolabs, 2005).

For future work, a simple way to increase efficiency and ensure that a vector is double cut would be to use an expression vector with a large (~1 kb) insert between the *Sal*I and *Not*I sites. Complete digestion with both enzymes will result in a linear DNA fragment 1 kb shorter than a vector cut with one of the enzymes. This difference can easily be resolved on an agarose gel, and linear DNA of the appropriate size can be excised from the gel and used for library construction.

The fact that less than 100% of the vector used for ligation was cut with both enzymes influenced the percentage of transformants containing an insert, which was determined to be 60% based on a screen of 50 randomly selected colonies. This is lower than reports of other cDNA–GFP fusion libraries, which range from 95% (Fujii et al., 1999) to 98.6% (Bejarano & Gonzalez, 1999). The average cDNA insert size of 1.1 kb was comparable to other studies, which ranged in size from 956 bp (Escobar et al.) to 1.37 kb (Misawa et al., 2000). The insert cDNA fragments are shorter than the average length of mouse full-length cDNAs, which is estimated to between 1.97 and 2.35 kb (Okazaki et al., 2002).

The library was transfected into Phoenix packaging cells, which are a cell line that constitutively expresses MMuLV group antigens polyproteins (gag), reverse transcriptase (pol) and envelope (env) proteins. These cells can package sequences contained in RV vectors into viral particles. The initial transfection of plasmid DNA into packaging cells was successful, as both the positive control and the library transfections were highly efficient, estimated to be between 30% and 40%. Within the library transfection, GFP-tagged proteins with a distinct subcellular localization were not observed. However, based on the estimated copy number of plasmids in cells, it would be difficult to resolve localization given that a number of different GFP-tagged proteins would be expressed in one cell.

Results of a second round of transfections provided some more insight as to why it might not have been possible to resolve the localization of GFP-tagged proteins in the initial transfections. The pBES23 expression vector without an insert, in which the GFP ORF lacks a Met initiation codon, was found to direct GFP expression in packaging cells. The likely explanation for this expression of GFP is that although the initiation codon, ATG, was removed from GFP, the second codon is GTG (Val), which can be used as an initiation codon in eukaryotes (Lodish et al., 2000). Changing this codon to another codon for Val (GTC, GTA or GTT) would result in an identical amino acid sequence, but perhaps reduce the background from empty clones. It is true that if 100% of the vector was digested with both enzymes, expression of GFP from empty clones would be rare, but it might be worthwhile make this change if the vector was being redesigned, so that localization of GFP-tagged proteins would not be obscured by expression of GFP from empty vector molecules.

Following virus collection and transduction of the embryonic cell line NIH–3T3, no virally transduced cells were recovered, as no cells fluoresced following the recommended incubation time with RV particles. The basis for this failure has not yet been determined. There are several steps in the experimental procedure that may contribute to low transduction of target cells, including poor transfection of packaging cells, lack of packaging of viral vector into viral particles, and inefficient transduction of target cells with virus.

Pinpointing the reason for low transduction has proved to be difficult because, in general, there are no simple methods for gauging the efficiency of intermediate steps in viral production. Based on the high efficiency of Phoenix cell transfection, at 40%, poor initial efficiency of DNA delivery was ruled out as a possible cause. In the viral packaging step there are, however, other factors that may contribute to low viral titres. If the Phoenix cells lose their expression of any one of the gag, pol or env genes, virus cannot be packaged. It is difficult to assay cells directly for expression of these viral proteins. While it is not in the standard protocol to passage Phoenix cells in selective media, if viral packaging is inefficient it may be necessary to reselect Phoenix cells by passaging them in the presence of diphtheria toxin and/or hygromycin, which may increase expression of env and gag/pol genes, respectively (Nolan, 2005).

If viral packaging is successful, it is possible that the virus being produced is not stable. The protocols followed in this thesis, which were obtained from the lab of Dr. James Ellis, require incubating Phoenix packaging cells at 37°C following transfection. However, protocols from the lab where the cells were designed recommend incubating cells at 32°C because the viral particle half-life is much higher at 32°C than 37°C (Nolan, 2005).

A third possibility is that packaged virus cannot infect target cells. The presence of a wild-type murine retrovirus in a cell line can block transduction by a second virus through interference (Nolan, 2005), but this seems unlikely in this case, as the NIH–3T3 cells were obtained from the ATCC and are certified to be virus free.

Poor transduction can also be caused by non-optimized concentrations of polybrene, or target cells being mitotically inactive or otherwise unhealthy at the time of transduction. If target cells are healthy, the efficiency of transduction may be enhanced by using a “spin infection” protocol (Nolan, 2005). In this case, target cells are grown in 6-well microplates and overlaid with viral supernatants and polybrene as in the standard protocol. Plates are then centrifuged at 1800 rpm for up to 45 min at room temperature. This enhances the contact of virus with cells, and may increase the efficiency of transduction up to 10-fold. Given that there are still several steps in retroviral packaging and transduction that may be made more efficient, it seems likely that with some optimization it will be possible to generate a library of cDNA–GFP fusions expressed by target cells.

With knowledge of the vector sequence combined with results of our experiments, it is apparent that there are also some improvements that could be made to the vector to increase the percentage of clones in a library containing an insert and expressing a cDNA–GFP fusion. To reduce expression from empty clones, the first GTG in GFP should be replaced with another Val codon and verified to ensure it does not allow expression of GFP. As described in Chapter 3, a further improvement would be to remove one base pair from the region between the *Not*I restriction site and the GFP coding sequence in the vector to ensure that no cDNAs will create a stop codon when ligated into the vector. Finally, to facilitate the recovery of vector DNA that has been cut with both *Sal*I and *Not*I, it would be helpful to add a large filler DNA fragment between the *Sal*I and *Not*I recognition sites.

Chapter 5 Summary and conclusions

5.1 Project Summary

This project began with the goal of developing a high-throughput method for characterizing protein localization in mouse stem cells. Unfortunately, constructing a cDNA–GFP fusion library proved enormously more complicated than simply making cDNA and ligating it into an expression vector containing a gene for GFP. In fact, the process required many steps, and careful consideration had to be given to the reaction conditions for cDNA synthesis, the design of the expression vector, and the delivery of cDNA–GFP fusions to cells. As with any high-throughput approach, there are limitations to using a library to characterize protein localization. A portion of GFP-tagged proteins expressed in a cDNA–GFP fusion library will not be able to localize to their native site because of missing targeting signals, some may mislocalize and generate inaccurate data, some may be unstable or toxic and not be represented in a screen, and some may be too rare to be isolated. Within these limitations, I focused on optimizing the process of making a cDNA–GFP fusion library in order to maximize the number of cDNA–GFP fusions that could be characterized.

Drawing on information contained in previously published reports of cDNA–GFP fusion libraries, I tried to optimize methods for making N-terminal cDNA libraries. To maximize the length of the coding sequence region for each cDNA in the library, I attempted to develop a novel method to synthesize cDNAs that had their 3´ UTR and stop codon enzymatically removed. This cDNA synthesis protocol was designed to produce near to full-length coding sequences fused at their C-termini to GFP. Unfortunately, I encountered technical difficulties when developing a method for efficient directional 3´ truncation on a population of cDNAs of various lengths, and thus used random-hexamer primed cDNA.

I had planned to express the cDNA–GFP fusions from an inducible promoter. This would allow GFP-tagged protein expression to be turned on only when cells are imaged, reducing possible deleterious effects associated with GFP-tagged protein expression. However, a careful review of literature showed that due to the large variation in inducibility between transfected clones, inducible expression could not be achieved reliably in high throughput. Therefore, I used a constitutive promoter to drive expression of cDNA–GFP fusions.

The decision to use retroviral delivery was motivated by a desire to limit the number of fusion constructs expressed in each cell. This delivery system would allow a population of transduced cells to be generated, each expressing a single cDNA–GFP fusion. I achieved the goal of constructing an N-terminal cDNA–GFP fusion library in a retroviral expression vector. To date, I have not been able to transduce this library into target cells for GFP-tagged protein localization analysis, but with further optimization of retroviral packaging and delivery this should be achievable.

5.2 Critical Analysis of cDNA–GFP Fusion Library Approach

The initial stage of this project required considerable effort to study, design, and refine methods to achieve the aforementioned goals regarding the creation of a cDNA–GFP fusion library. However, the efforts spent addressing the technical aspects of making a cDNA–GFP library detracted focus from the overall project goal, which was to develop a high-throughput method for characterizing protein localization using GFP-tagged proteins. Although a cDNA–GFP fusion library is one means to generate many GFP-tagged tagged proteins, a closer analysis of the difficulties associated with the creation of these libraries revealed that there were several limitations to using a fusion library to study protein localization comprehensively in stem cells. Unfortunately, this became apparent only after I had gained hands-on experience with retroviral vectors and a more comprehensive understanding of characteristics of proteins expressed from a cDNA–GFP fusion library. Limitations of this approach include problems with clone identification and clone redundancy, the deleterious effects of overexpression of proteins with normally low abundance or potent biological activities, and the inability of protein fragments to allow for accurate monitoring of GFP-tagged protein localization.

5.2.1 Retroviral Delivery, Clone Redundancy and Clone Identification

Delivering cDNA–GFP fusions to cells using a retroviral system was initially thought to be the best way to achieve delivery of one cDNA–GFP fusion per cell. One alternative method was to isolate individual cDNA–GFP fusion plasmids and transfect them individually to target cells, which would be more labour intensive, time consuming, and costly. Using a retroviral approach, cDNA–GFP fusions are packaged and delivered to target cells in bulk reactions. Subsequently, cells expressing tagged proteins with localizations of interest are isolated and clonally expanded in culture, and cDNA fragments encoding the tagged proteins are amplified for sequencing using PCR or RT–PCR. Although this strategy allows one to go quickly from generating cDNA to obtaining GFP-tagged protein expression, a closer analysis of the population of cells generated after retroviral transduction revealed that this approach introduces numerous difficulties into the downstream analysis.

One problem with retroviral delivery is that there is a high amount of redundancy in the population of transduced cells expressing GFP-tagged proteins, meaning that the same clone may be isolated multiple times in a screen for localization. This redundancy arises because the cDNA–GFP fusion library is amplified repeatedly. Initially, cDNA–GFP fusion plasmids are maintained and amplified in bacterial cells, and plasmid DNA isolated from these cells is transfected into packaging cells in a bulk reaction. Each individual clone in the cDNA–GFP library contributes many plasmid copies to this DNA stock, meaning multiple packaging cells are transfected with the same cDNA–GFP fusion construct. Each packaging cell that receives a specific cDNA–GFP fusion creates multiple retroviral copies of it. Because of these two amplification steps, upwards of several hundred retrovirally transduced target cells may express a GFP-tagged protein derived from a single cDNA–GFP fusion construct in the initial plasmid library. The large amount of redundancy in the population of target cells means that a large amount effort will be spent repeatedly isolating and identifying the same cDNA–GFP fusion.

A second drawback to delivering cDNA–GFP fusions to cells using a retroviral approach is that one runs the risk of not being able to identify the cDNA that encodes a protein with a localization of interest. This may be due to difficulties with PCR, or poor cell viability due to culture conditions or GFP-tagged protein expression. Although the reliability of single-cell PCR to amplify a single copy of a genomic target for sequencing can be achieved with a 90% success rate, reliability of sequences from a single copy is such that more accurate data can be obtained using multiple PCRs or higher cell numbers (Gale et al., 2003). Isolating and clonally expanding single stem cells in culture is not a 100% reliable process, even if one takes the precaution of using conditioned media for cell growth. In previous research in the Jervis lab, the survival rate of single cells grown in isolation was approximately 75%. Cells may have arrested cell growth or may spontaneously undergo apoptosis, even after a few rounds of cell division. Poor cell survival may also be caused by GFP-tagged protein expression, which, as discussed in 5.2.2, may arrest cell growth or cause cell death.

Thus, by using a retroviral delivery method it is possible that one may identify an interesting GFP expression pattern but not be able to identify the cDNA encoding it because the transgenic cells cannot be expanded in culture, or because PCR of the cDNA cannot be achieved. Given that the processes of isolating individual target cells expressing GFP-tagged proteins and expanding them in culture are laborious and time-consuming, it may take more time and effort to identify cDNAs expressed in a retroviral library than it would have taken to prepare individual plasmids from a cDNA–GFP fusion library and transfect them in a 96-well format using non-retroviral methods.

After more careful consideration, delivering cDNA–GFP fusions to cells using a retroviral approach generates many problems in downstream analysis. To eliminate the requirement of isolating and expanding cells to identify cDNAs, and to reduce the possibility of isolating the same clone many times over, it would be preferable to individually isolate and expand plasmids containing cDNA–GFP fusions and transfect them into target cells. Although the upfront work is greater, this method has the advantage that the cDNAs encoding proteins of interest have already been isolated in a plasmid stock; thus isolating and clonally expanding target cells expressing GFP-tagged proteins is not necessary. Using this approach, cDNA normalization protocols would be an effective way to reduce the representation of highly abundant or redundant cDNAs in a library.

In a normalized cDNA library, the frequency of each clone is within a narrow range so that each cDNA is represented with approximately uniform abundance (Soares et al., 1994). Normalization protocols are based on reassociation kinetics: in a population of denatured ds cDNAs, rarer species anneal less rapidly than abundant ones and the single-stranded fraction of cDNA becomes normalized during the course of the hybridization (Patanjali et al., 1991). Because the lengths of cDNAs affect their reassociation kinetics, many normalization protocols require digesting cDNAs into small fragments to limit the effect of DNA length on reassociation (Diatchenko, et al., 1996). If one wishes to normalize a library with full-length or near-to full-length cDNAs, there are fewer options for achieving this. One possibility is to use cDNA immobilized on a substrate to normalize a population of mRNAs, and then use the normalized mRNA as a template for cDNA library synthesis (Tanaka et al., 1996; Sasaki et al., 1994). An alternate method developed by Carninci et al. (2000) allows for normalization of full-length cDNAs by allowing them to hybridize to mRNA in solution. This method has the added advantage that it also enriches for cDNAs that are reverse transcribed all the way to the end of the 5´ UTR.

When exploring ways to make cDNA for a library, I performed preliminary trials using Carninci et al.'s method. However, the procedure was not designed for use with random-primed cDNA, and changing the protocol to normalize random-primed cDNA would require substantial modifications to the procedure, including altering first-strand synthesis primer sequences and concentrations, and modifying conditions for second-strand cDNA synthesis to allow for cDNAs to be used in fusion proteins. Performing these extensive trials was not within the time frame of this project.

5.2.2 The Effect of Transgene Expression on Protein Localization and Cell Viabilty

When a collection of random cDNA–GFP fusions is expressed in cells, the use of a strong constitutive promoter excludes some GFP-tagged proteins from analysis. It is true that in any high-throughput method not all molecules will be amenable to identification and characterization. However, in a cDNA–GFP fusion library approach, many of proteins that are excluded from analysis are those of high scientific interest. Proteins that are endogenously expressed at low levels or those with potent cellular activity, such as cell fate determinants or cell-specific transcription factors, can cause deleterious effects in cells when they are overexpressed. Excess levels of these proteins may lead to inaccurate localization data or pathological effects and cell death. This has been demonstrated with many biologically active and cell-type-specific proteins, including glucocortocoid-receptor–GFP fusions, where overexpression was shown to alter cellular physiology and arrest cell growth in murine cell lines (Walker et al, 1999). Native localization, translocation and cell viability could only be observed once expression levels were reduced by using an inducible expression system. Similar problems occurred in a study of neurofilament H (NFH)–GFP fusions, where overexpression quickly caused deleterious and undesirable effects, including lack of proper NFH–GFP localization into filaments, greatly increased cell size and multinucleation, and after hours in culture, cell death (Szebenyi et al., 2002). Again, an inducible expression system was required to obtain meaningful localization data. Thus, using an inducible promoter seems to be the most reliable way of obtaining biologically relevant expression and localization of GFP-tagged proteins.

5.2.3 Limits of Using Protein Fragments to Study Localization

cDNA–GFP fusion libraries may also be used to monitor proteins that translocate or asymmetrically partition during development. Proteins that target to multiple subcellular locations may contain multiple distinct targeting domains within the full-length protein (Karniely & Pines, 2005). For example, in Drosophila, the first 76 amino acids are necessary to localize the cell fate determinant Numb to the cell membrane, but the first 277 amino acids are required for partitioning during an asymmetric division (Knoblich et al., 1997). The related protein PON also has targeting domains in distinct regions of the full-length protein, with a Numb-interacting domain at the N-terminus and an asymmetric localization domain at the C-terminus (Lu et al., 1999). Efficient protein translocation may similarly require both ends of a full-length protein, as in the case of fibroblast growth factor (FGF)-1, where proper translocation and nuclear import of activated FGF-1 requires both the N- and C-termini to be present (Wesche et al., 2005).

Therefore, the use of a cDNA–GFP library has limited utility if one wishes to identify proteins with dynamic localizations because non-full length proteins may be compromised in their ability to direct localization and translocation. The use of partial cDNAs to direct expression of GFP-tagged proteins fusion reduces the likelihood of their dynamic properties being revealed in a screen because essential regions may be missing.

Finally, it is worth noting that another class of potentially biologically relevant proteins appear to be underrepresented or excluded from analysis in cDNA–GFP fusion libraries. Proteins that localize to the cell surface, which may serve as receptors for growth factors or serve as cell-specific markers, were not observed in either of the mammalian cell cDNA–GFP fusion libraries in which all observed localizations of GFP-tagged proteins are reported (Rolls et al., 1999; Manabe et al., 2002). This suggests that the use of protein fragments may not be sufficient to confer proper localization for proteins that localize to the cell surface.

5.3 Conclusions

A close examination of the cDNA–GFP library screening process and the properties of proteins expressed suggests that the range of proteins that can be identified in a library may not be comprehensive. However, this may not be a limitation in some circumstances. For example, this approach was successfully utilized to identify the protein components of tobacco plasmodesmata (Escobar et al. 2003). Because plasmodesmata cannot be isolated from cells for proteomic analysis, until Escobar et al., the protein components of the plasmodesmata remained elusive (Cilia & Jackson, 2004). The use of a fusion library allowed for the identification of 10 proteins that are potentially localized in plasmodesmata from an initial screen of 20,000 cDNA–GFP fusions (Escobar et al. 2003). Libraries have also allowed for the identification of a protein that regulates microtubule organization and cytokinesis in human cells (Manabe et al., 2002), and nuclear localization domains in human cells (Fujii et al., 1999).

The utility of cDNA–GFP fusion libraries in the context of mouse stem cell biology may be somewhat limited. The mouse transcriptome is well characterized, with an abundance of transcript information available to researchers, such as the Mouse Full-Length cDNA Encyclopedia, which contains the complete sequences of more than 60,000 full-length cDNAs (Carninci et al., 2003). Proteins expressed in mice are also relatively well characterized, with close to 10,000 characterized with respect to localization, function, post-translational modifications, structure and other characteristics (ExPASy, 2005).

Given that many of the characterized proteins are those that are most abundantly expressed, elucidating the localization and function of the more rare proteins ought to be the focus of any protein characterization strategy. However, in a cDNA–GFP fusion library, the majority of GFP-tagged proteins created will be those abundantly expressed and well characterized. Using a library approach, there is no easy way to exclude these proteins from analysis, making it difficult to isolate and identify the more rare proteins, which are of higher interest. Because of these problems and additional limitations associated with overexpression and the use of protein fragments, a cDNA–GFP fusion library may not be the best method for identifying and characterizing rare proteins in stem cells.

A further indication that cDNA–GFP fusion libraries have limited utility in protein identification and characterization is that since initial publications using this approach from 1999 to 2002, cDNA–GFP libraries have not been adopted in general use as a tool for probing protein localization. This was forecast in 2001 when Pepperkok et al. asserted that the time spent characterizing redundant, inaccurate or previously characterized cDNAs would limit the utility of large-scale random GFP fusion approaches as a tool to characterize protein localization.

5.4 Recommendations for future work

While a cDNA–GFP fusion strategy may not be an effective method for studying protein localization in live cells, high-throughput protein localization studies using GFP-tagged proteins are quite possible. However, instead of generating fusions *en masse* using a cDNA library approach, it is possible to tailor localization studies to focused sets of proteins. Creating fusion proteins in high-throughput is also possible without a need for restriction enzyme mediated cloning. The first report of this, far in advance of more recent widespread application, was in 2000, when Simpson et al. used Gateway^{TM}-mediated recombination technology to clone more than 500 full-length human ORFs as GFP fusions and characterized protein localization in live cells. At the time that I began this project, this approach was viewed as being too labour intensive, but given the limitations of cDNA–GFP fusion libraries outlined above it may be a better way to create large numbers of GFP-tagged proteins. A significant advantage of Simpson et al.'s approach is that once the amplified ORFs are recombined into an entry vector, the same entry vector can be used to generate a variety of fusion proteins with either N- terminal or C- terminal fluorescent tags, and cyan or yellow variants of GFP depending on the destination vector .

This method has only recently been adopted for generating collections of GFP-tagged proteins in other organisms, including *Aspergillius niger* (Toews et al., 2004), *Plasmodium falciparum* (Tonkin et al., 2004), *Arabidopsis thaliana* (Koroleva et al., 2005) and *Escherichia coli* (Suzuki et al., 2005). These studies have proven to be enormously useful, allowing researchers to construct hundreds of full-length GFP-tagged proteins. This method has not yet been adapted for use in mouse stem cells, but would be highly useful. After purchasing the Gateway^{TM} vector system and cloning the ORFs of interest into a Gateway^{TM} donor vector, the cost of constructing a GFP-tagged expression construct is approximately $5 US (Koroleva et al., 2004).

If one wished to adapt this approach to perform in-depth analyses of protein localization and dynamics in mouse stem cells, transcripts of potential interest may easily be identified from any one of a number of microarray studies of transcript abundance over the course of mouse development (e.g. Wigle et al., 2001; Ramalho-Santos et al., 2002; Ivanova et al, 2002; Rao & Stice., 2004; Perez-Iratxeta et al, 2005). Full-length sequences of cDNAs expressed in early development have also been identified in large-scale stem-cell-specific cDNA projects, such as the National Institute on Aging's cDNA project in mouse stem cells and early embryos (Carter et al., 2003). Many of these transcripts would be ideal candidates for creating GFP-tagged proteins, as they may encode proteins that play important roles in stem cell biology.

Once cDNAs of interest are identified, it is necessary to obtain a copy of their cDNA sequence in order to create a fusion protein. Although it is possible to attempt to construct these using RT–PCR or obtain them individually from the ATCC (Carter et al, 2003), the RIKEN Mouse Genome Encyclopedia DNA Book would be a valuable resource for obtaining many cDNA clones. Its pages contain more than 60,000 full-length cDNA clones in plasmids that can be recovered using PCR (Kawai & Hayashizaki, 2003). GFP-tagged versions of the proteins within these cDNAs can be made as N- and or C- terminal fusions (or both), and the level of transgene expression can be controlled by selecting an appropriate promoter in the destination vector. Because vectors need not be retroviral, larger capacity vectors can be used, allowing for the use of single vector inducible expression systems (e.g. Mizuguchi et al., 2003). If one selected a few hundred candidate genes and created GFP-ORF fusions, although the upfront work is greater than making a library, the actual utility is much greater. In an N-terminal cDNA–GFP fusion library with 10^5 clones, it is reasonable to assume that 1/10 clones will not contain an insert, 1/3 of cDNAs will not be in frame with GFP, 1/2 of these clones will not display a subcellular localization, 1/3 of these will display non-native localization, an estimated 9/10 of clones isolated will be previously characterized or redundant, and an unknown proportion will have further problems because of the effects of overexpression (conservatively estimated at 30%).

From the initial 10^5 clones, fewer than 1% of the cDNA–GFP fusions in the library may be of interest and amenable to analysis, and they will be difficult to identify and analyze from the initial pool of 100,000. Thus, although the upfront work is greater using a full-length coding sequence Gateway$^{\text{TM}}$-mediated cloning approach, the actual utility of the clone set is much higher than those generated using a cDNA–GFP fusion approach. Proteins are all full-length, the set of tagged proteins is focused to those that are uncharacterized or of interest, the location of GFP relative to the protein can be controlled for each clone, the level of expression can be tightly controlled, there is no clone redundancy, and the identity of each clone is known without having to isolate and expand transfected cells. Perhaps most usefully, it opens the door for many different analyses using experimental replicates or different experimental conditions.

High-throughput imaging of protein localization can be achieved by spotting GFP-fusion construct DNAs onto transfected cell microarrays (Ziauddin & Sabatini 2001). Cells growing on these arrays express the protein encoded by the plasmid in the spot on which they grow. Defined arrays of GFP-tagged proteins would be enormously useful to study the behaviours of proteins under different conditions because hundreds or thousands of different proteins can be analyzed in parallel. In the context of stem cell biology, these sets of GFP-tagged proteins can be used to study changes in localization or dynamics in response to growth factors or environmental conditions by coupling high-throughput imaging with time-course analyses. Collections of GFP-tagged proteins would also be useful in pharmaceutical applications, where protein changes in response to drugs can be imaged in high throughput. The use of transfected cell microarrays in the context of pharmaceutical applications has already been recognized (Bailey et al., 2002), but coupling it with GFP-tagging technology would allow for the effects of drugs on protein behaviour and dynamics to be monitored to a level not currently possible.

Bibliography

Bailey, S.N., Wu, R.Z. and Sabatini, D.M. (2002). Applications of transfected cell microarrays in high-throughput drug discovery. *Drug Discov Today*. **15**, S113-8.

Bejarano, L.A. and Gonzalez, C. (1999). Motif trap: a rapid method to clone motifs that can target proteins to defined subcellular localisations. *J Cell Sci*. **112,** 4207-11.

Bellaïche, Y., Gho, M., Kaltschmidt, J.A., Brand, A.H. and Schweisguth, F. (2001). Frizzled regulates localization of cell-fate determinants and mitotic spindle rotation during asymmetric cell division**.** *Nat Cell Biol*. **3**, 50-57.

Blau, H.M. and Rossi, F.M. (1999). Tet B or not tet B: advances in tetracycline-inducible gene expression. *Proc Natl Acad Sci USA*. **96,** 797-9.

Blau, H.M., Brazelton, T.R. and Weimann, J.M. (2001). The evolving concept of a stem cell: entity or function? *Cell.* **105**, 829-41.

Bowman, B.H. and Palumbi, S.R. (1993). Rapid production of single-stranded sequencing template from amplified DNA using magnetic beads. *Methods Enzymol.* **224**, 399-406.

Brunet, S., Thibault, P., Gagnon, E., Kearney, P., Bergeron, J.J. and Desjardins, M. (2003). Organelle proteomics: looking at less to see more. *Trends Cell Biol*. **3,** 629-38.

Cammas, F., Oulad-Abdelghani, M., Vonesch, J.L., Huss-Garcia, Y., Chambon, P. and Losson, R. (2002). Cell differentiation induces TIF1beta association with centromeric heterochromatin via an HP1 interaction. *J Cell Sci*. **115**, 3439-48.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M. and Hayashizaki, Y. (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res*. **10**, 1617-30.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., et al. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res*. **13**, 1273-89.

Carter, M.G., Piao, Y., Dudekula, DB., Qian, Y., VanBuren, V., Sharov, A.A., Tanaka, T.S., Martin, P.R., Bassey, U.C., Stagg, C.A., et al. (2003). The NIA cDNA project in mouse stem cells and early embryos. *C R Biol*. **326**, 931-40.

Cavaleri, F. and Scholer, H.R. (2003). Nanog: a new recruit to the embryonic stem cell orchestra. *Cell.* **113**, 551-2.

Chambers, I. (2004). The molecular basis of pluripotency in mouse embryonic stem cells. *Cloning Stem Cells*. **6**, 386-91.

Chung, S., Andersson, T., Sonntag, K.C., Bjorklund, L., Isacson, O. and Kim, K.S. (2002). Analysis of different promoter systems for efficient transgene expression in mouse embryonic stem cell lines. *Stem Cells*. **20**, 139-45.

Cilia, M.L. and Jackson, D. (2004). Plasmodesmata form and function. *Curr Opin Cell Biol*. **16,** 500-6.

Coffin, J., Hughes, S. and Varmus, H. (1997). *Retroviruses*. Cold Spring Harbor Laboratory Press, N.Y.

Cutler, S.R., Ehrhardt, D.W., Griffitts, J.S. and Somerville, C.R. (2000). Random GFP::cDNA fusions enable visualization of subcellular structures in cells of Arabidopsis at a high frequency. *Proc Natl Acad Sci USA*. **97**, 3718-23.

Daley, G.Q., Goodell, M.A. and Snyder, E.Y. (2003). Realistic prospects for stem cell therapeutics. *Hematology*. **1**, 398-418.

Das, M., Harvey, I., Chu, L.L., Sinha, M. and Pelletier, J. (2001). Full-length cDNAs: more than just reaching the ends. *Physiol Genomics*. **6**, 57-80.

Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D. and Siebert, P.D. (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA*. **93**, 6025-30.

Donnes, P. and Hoglund, A. (2004). Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics*. **2**, 209-15.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*. **300**, 1005-16.

Escobar, N.M., Haupt, S., Thow, G., Boevink, P., Chapman, S., and Oparka, K. (2003). High-throughput viral expression of cDNA-green fluorescent protein fusions reveals novel subcellular addresses and identifies unique proteins that interact with plasmodesmata. *Plant Cell*. **15**, 1507-23.

ExPASy. (2005). Swiss-Prot Protein Knowledgebase Release 47.6 Statistics. http://ca.expasy.org/sprot/relnotes/relstat.html. Accessed August 5, 2005.

Fujii, G., Tsuchiya, R., Ezoe, E. and Hirohashi, S. (1999). Analysis of nuclear localization signals using a green fluorescent protein-fusion protein library. *Exp Cell Res*. **251**, 299-306.

Gale, J.M., Romero, C.P., Tafoya, G.B. and Conia, J. (2003). Application of optical trapping for cells grown on plates: optimization of PCR and fidelity of DNA sequencing of p53 gene from a single cell. *Clin Chem*. **49**, 415-24.

Gasca, S., Canizares, J., De Santa Barbara, P., Menean, C., Poulat, F., Berta, P. and Boizet-Bonhoure, B. (2002). A nuclear export signal within the high mobility group domain regulates the nucleocytoplasmic translocation of SOX9 during sexual determination. *Proc Natl Acad Sc USA*. **99**, 11199-204.

Gevaert, K. and Vandekerckhove, J. (2000). Protein identification methods in proteomics. *Electrophoresis*. **21**, 1145-54

Gossen, M. and Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc Natl Acad Sc USA*. **89**, 5547-51.

Gossen, M., Freundlieb, S., Bender, G., Muller, G., Hillen, W. and Bujard, H. (1995). Transcriptional activation by tetracyclines in mammalian cells. *Science*. **268**(5218), 1766-9.

Grez, M., Akgun, E., Hilberg, F. and Ostertag, W. (1990). Embryonic stem cell virus, a recombinant murine retrovirus with expression in embryonic stem cells. *Proc Natl Acad Sci USA*. **87**, 9202-6.

Guo, L.H. and Wu, R. (1982). New rapid methods for DNA sequencing based in exonuclease III digestion followed by repair synthesis. *Nucleic Acids Res*. **10,** 2065-84.

Harry, J.L., Wilkins, M.R., Herbert, B.R., Packer, N.H., Gooley, A.A. and Williams, K,L. (2000). Proteomics: capacity versus utility. *Electrophoresis*. **21**, 1071-81.

Henikoff, S. (1987). Unidirectional digestion with exonuclease III in DNA sequence analysis. *Methods Enzymol*. **155**, 156-65.

Hoheisel, J.D. (1993). On the activities of Escherichia coli exonuclease III. *Anal Biochem*. **209**, 238-46.

Huber, L.A., Pfaller, K. and Vietor, I. (2003). Organelle proteomics: implications for subcellular fractionation in proteomics. *Circ Res.* **92**, 962-8.

Ivanova, N.B., Dimos, J.T., Schaniel, C., Hackney, J.A., Moore, K.A. and Lemischka, I.R. (2002). A stem cell molecular signature. *Science*. **298**(5593), 601-604.

Karniely, S. and Pines, O. (2005). Single translation--dual destination: mechanisms of dual protein targeting in eukaryotes. *EMBO Rep*. **6**, 420-5.

Kawai, J, and Hayashizaki, Y. (2003). DNA book. *Genome Res*. **13**, 1488-95.

Kinsella, T.M. and Nolan, G.P. (1996). Episomal vectors rapidly and stably produce high-titer recombinant retrovirus. *Hum Gene Ther*. **7**, 1405-13.

Knoblich, J.A., Jan, L.Y. and Jan, Y.N. (1997). The N terminus of the Drosophila Numb protein directs membrane association and actin-dependent asymmetric ocalization. *Proc Natl Acad Sci USA.* **94**, 13005-10.

Koroleva, O.A., Tomlinson, M.L., Leader, D., Shaw, P. and Doonan, J.H. (2005). High-throughput protein localization in Arabidopsis using Agrobacterium-mediated transient expression of GFP-ORF fusions. *Plant J.* **41**, 162-74.

Liu, M. and Price, D.H. (1997). In vitro Transcription on DNA Templates Immobilized to Streptavidin Coated MagneSphere Paramagnetic Particles. *Promega Notes*. **64**, 21-26.

Lodish, H., Berk, K., Zipursky, S.L., Matsudaira, P., Baltimore, D. and Darnell, J. (2000). *Molecular Cell Biology*. WH Freeman, NY.

Lu, B., Ackerman, L., Jan, L.Y. and Jan, Y.N. (1999). Modes of protein movement that lead to the asymmetric localization of partner of Numb during Drosophila neuroblast division. *Mol Cell*. **4**, 883-91.

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D.S., Poulin, B., Anvik, J., Macdonell, C. and Eisner, R. (2004). Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*. **20**, 547-56.

Majoul, I., Straub, M., Duden, R., Hell, S.W. and Soling, H.D. (2002). Fluorescence resonance energy transfer analysis of protein-protein interactions in single living cells by multifocal multiphoton microscopy. *J Biotechnol*. **82**, 267-77.

Makalowski, W., Zhang, J. and Boguski, M.S. (1996). Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res*. **6**, 846-57.

Manabe, R., Whitmore, L., Weiss, J.M. and Horwitz, A.R. (2002). Identification of a novel microtubule-associated protein that regulates microtubule organization and cytokinesis by using a GFP-screening strategy. *Curr Biol*. **12**, 1946-51

Misawa, K., Nosaka, T., Morita, S., Kaneko, A., Nakahata, T., Asano, S. and Kitamura, T. (2000). A method to identify cDNAs based on localization of green fluorescent protein fusion products. *Proc Natl Acad Sci USA*. **97**, 3062-6.

Mizuguchi, H., Xu, Z.L., Sakurai, F., Mayumi, T. and Hayakawa, T. (2003). Tight positive regulation of transgene expression by a single adenovirus vector containing the rtTA and tTS expression cassettes in separate genome regions. *Hum Gene Ther*. **14**, 1265-77.

Moss, G.P. (2005). *International Union of Biochemistry and Molecular Biology Recommendations on Biochemical & Organic Nomenclature, Symbols & Terminology etc*. http://www.chem.qmul.ac.uk/iubmb/. Accessed August 9, 2005.

Nagy, A., Rossant, J., Nagy, R., Abramow-Newerly, W. and Roder, J.C. (1993) Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proc Natl Acad Sci USA*. **90**, 8424-8.

Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*. **24**, 34-6.

New England Biolabs. (2005) *New England Biolabs 2005-06 Catalog and Technical Reference*. New England Biolabs, Beverly, MA.

No, D., Yao, T.P. and Evans, R.M. (1996). Ecdysone-inducible gene expression in mammalian cells and transgenic mice. *Proc Natl Acad Sci USA*. **93**, 3346-51.

Nolan, G. (2005). *Troubleshooting and Optimization of Retroviral Transduction Parameters.* http://www.stanford.edu/group/nolan/protocols/pro_optimiz.html. Accessed June 17, 2005.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature.* **420**(6915), 563-73.

Osborne, C.S., Pasceri, P., Singal, R., Sukonnik, T., Ginder, G.D. and Ellis, J. (1999) Amelioration of retroviral vector silencing in locus control region beta-globin-transgenic mice and transduced F9 embryonic cells. *J Virol*. **73**, 5490-6.

Ozawa, T., Nishitani, K., Sako, Y. and Umezawa, Y. (2005). A high-throughput screening of genes that encode proteins transported into the endoplasmic reticulum in mammalian cells. *Nucleic Acids Res*. **33**(4), e34.

Patanjali, S.R., Parimoo, S. and Weissman, S.M. (1991). Construction of a uniform-abundance (normalized) cDNA library. *Proc Natl Acad Sci USA.* **88**, 1943-7.

Pepperkok, R., Simpson, J.C. and Wiemann, S. (2001). Being in the right location at the right time. *Genome Biol*. **2**, 1024.

Perez-Iratxeta, C., Palidwor, G., Porter, C.J., Sanche, N.A., Huska, M.R., Suomela, B.P., Muro, E.M., Krzyzanowski, P.M., Hughes, E., Campbell, P.A., et al. (2005). Study of stem cell function using microarray experiments. *FEBS Lett.* **579**, 1795-801.

Perucho, M., Hanahan, D. and Wigler, M. (1980). Genetic and physical linkage of exogenous sequences in transformed cells. *Cell.* **22**, 309-17.

Putney, S.D., Benkovic, S.J. and Schimmel, P.R. (1981). A DNA fragment with an alpha-phosphorothioate nucleotide at one end is asymmetrically blocked from digestion by exonuclease III and can be replicated in vivo. *Proc Natl Acad Sci USA*. **78**, 7350-4.

Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C. and Melton, D.A. (2002). "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science*. **298**(5593), 597-600.

Rao, M. (2004). Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells. *Dev Biol.* **275**, 269-86.

Rao, R.R. and Stice, S,L. (2004). Gene expression profiling of embryonic stem cells leads to greater understanding of pluripotency and early developmental events. *Biol Reprod*. **71**, 1772-8.

Roegiers, F. and Jan, Y.N. (2004). Asymmetric cell division. *Curr Opin Cell Biol*. **16,** 195-205.

Rogers, S.G. and Weiss, B. (1980). Exonuclease III of Escherichia coli K-12, an AP endonuclease. *Methods Enzymol*. **65**, 201-11.

Rolls, M.M., Stein, P.A., Taylor, S.S., Ha, E., McKeon, F. and Rapoport, T.A. (1999). A visual screen of a GFP-fusion library identifies a new type of nuclear envelope membrane protein. *J Cell Biol*. **146**, 29-44.

Sambrook, J. and Russell, D. (2001). *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Press, NY.

Sasaki,Y., Ayusawa, D. and Oishi, M. (1994). Construction of a normalized cDNA library by introduction of a semi-solid mRNA-cDNA hybridization system. *Nucleic Acids Res.* **22**, 987-992.

Schneider, G. and Fechner, U. (2004). Advances in the prediction of protein targeting signals. *Proteomics*. **4,** 1571-80.

Simpson, J.C., Wellenreuther, R., Poustka, A., Pepperkok, R. and Wiemann, S. (2000). Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep*. **1**, 287-92.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L. and Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci USA*. **91**, 9228-32.

Stratagene. (2002) *Complete Control Retroviral Inducible Expression System Instruction Manual.* Stratagene, La Jolla, CA. Revision#082005.

Suzuki, Y., Kagawa, N., Fujino, T., Sumiya, T., Andoh, T., Ishikawa, K., Kimura, R., Kemmochi, K., Ohta, T. and Tanaka, S. (2005). A novel high-throughput (HTP) cloning strategy for site-directed designed chimeragenesis and mutation using the Gateway cloning system. *Nucleic Acids Res*. **33**(12):e109.

Szebenyi, G., Smith, G.M., Li, P. and Brady, S.T. (2002). Overexpression of neurofilament H disrupts normal cell structure and function. *J Neurosci Res*. **68**, 185-98.

Tanaka, T., Ogiwara, A., Uchiyama, I., Takagi, T., Yazaki, Y. and Nakamura, Y. (1996). Construction of a normalized directionally cloned cDNA library from adult heart and analysis of 3040 clones by partial sequencing. *Genomics*. **35**, 231-235.

Thiede, B., Hohenwarter, W., Krah, A., Mattow, J., Schmid, M., Schmidt, F. and Jungblut, P,R. (2005). Peptide mass fingerprinting. *Methods.* **35**, 237-47.

Toews, M.W., Warmbold, J., Konzack, S., Rischitor, P., Veith, D., Vienken, K., Vinuesa, C., Wei, H. and Fischer, R. (2004). Establishment of mRFP1 as a fluorescent marker in Aspergillus nidulans and construction of expression vectors for high-throughput protein tagging using recombination in vitro (GATEWAY). *Curr Genet.* **45**, 383-9.

Tonkin, C.J., van Dooren, G.G., Spurck, T.P., Struck, N.S., Good, R.T., Handman, E., Cowman, A.F. and McFadden, G.I. (2004). Localization of organellar proteins in Plasmodium falciparum using a novel set of transfection vectors and a new immunofluorescence fixation method. *Mol Biochem Parasitol.* **137**, 13-21.

Tsien, R.Y. (1998). The green fluorescent protein. *Annu Rev Biochem.* **67**, 509-44.

Unwin, R.D., Gaskell, S.J., Evans, C.A. and Whetton, A.D. (2003). The potential for proteomic definition of stem cell populations. *Exp Hematol*. **31**, 1147-59.

Walker, D., Htun, H. and Hager, G.L. (1999). Using inducible vectors to study intracellular trafficking of GFP-tagged steroid/nuclear receptors in living cells. *Methods.* **19**, 386-93.

Wakita, K., McCormick, F. and Tetsu, O. (2001). Method for screening ecdysone-inducible stable cell lines. *Biotechniques*. **31**, 414-8.

Weiss, B. (1976). Endonuclease II of Escherichia coli is exonuclease III. J Biol Chem. **251**, 1896-901.

Wellenreuther, R., Schupp, I., Poustka, A., Wiemann, S. and The German cDNA Consortium. (2004). SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones. *BMC Genomics*. **5**, 36.

Wesche, J., Malecki, J., Wiedlocha, A., Ehsani, M., Marcinkowska, E., Nilsen, T. and Olsnes, S. (2005). Two nuclear localization signals required for transport from the cytosol to the nucleus of externally added FGF-1 translocated into cells. *Biochmistry*. **44**, 6071-80.

Wigle, D.A., Rossant, J. and Jurisica, I. (2001) Mining mouse microarray data. *Genome Biol*. **2**, 1019.

Wyborski, D.L., Bauer, J.C. and Vaillancourt, P. (2001). Bicistronic expression of ecdysone-inducible receptors in mammalian cells. *Biotechniques*. **31**, 618-20, 622, 624.

Yao, S., Osborne, C.S., Bharadwaj, R.R., Pasceri, P., Sukonnik, T., Pannell, D., Recillas-Targa, F., West, A.G. and Ellis, J. (2003). Retrovirus silencer blocking by the cHS4 insulator is CTCF independent. *Nucleic Acids Res*. **31**, 5317-23.

Zeng, X., Chen, J., Sanchez, J.F., Coggiano, M., Dillon-Carter, O., Petersen, J. and Freed, W.J. (2003). Stable expression of hrGFP by mouse embryonic stem cells: promoter activity in the undifferentiated state and during dopaminergic neural differentiation. *Stem Cells*. **21**, 647-53.

Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. and Siebert, P.D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques.* **30**, 892-7.

Ziauddin, J. and Sabatini, D.M. (2001). Microarrays of cells expressing defined cDNAs. *Nature*. **411**(6833), 107-10.