2

Optimizing Recombinant Protein Expression in Soybean

Laura C. Hudson, Kenneth L. Bost and Kenneth J. Piller
University of North Carolina at Charlotte and SoyMeds, Inc.
United States of America

1. Introduction

The production of biopharmaceuticals represents the fastest growing segment in the pharmaceutical industry. Currently, the majority of biopharmaceuticals are produced in recombinant microbe expression systems. However, microbes have certain limitations in the classes of proteins that can be economically produced, and in the post-translational processing that can be achieved. To solve this problem, insect and mammalian cell cultures have also been utilized for eukaryotic protein production (Lubiniecki and Lupker, 1994; Kost and Condreay, 1999). However, a significant problem with these systems, in particular cell cultures, is that production costs are prohibitively high for many proteins. Therefore, an increased demand for biopharmaceuticals will require improved and cost effective manufacturing practices as well as practical transportation and delivery methods for a global community. Over the past two decades, there has been substantial research on the expression of heterologous proteins in plants as a means to produce biopharmaceuticals that can meet current and future global needs. Several excellent review articles describe these recent advances (Streatfield 2007; Karg and Kallio 2009; Franconi et al., 2010).

While numerous plant systems have been shown to support expression of heterologous proteins, we believe that soybean is perhaps the most practical of these systems. Soybean is often overlooked as an expression system in part due to a technically challenging, lengthy, and costly transformation process. However, there is enormous potential for the use of transgenic soybean as a factory for the economic production of pharmaceutical proteins. The soybean system has distinct advantages which offer a practical alternative to existing expression systems.

First, while soybeans are traditionally thought of as high oil seeds, they are also high protein seeds. Soybeans contain nearly 40% protein by dry mass, and represent one of the richest natural sources of protein known. Given this high protein content, it is therefore possible to express in excess of a milligram of transgenic protein in a single soybean seed. There are few, if any, host systems (plant or non-plant) that can produce such levels of foreign protein based on weight. Second, soybean is a relatively easy and inexpensive plant to grow. Therefore, the production of biopharmaceuticals in soybeans is extremely cost-effective. For example, given the high protein content of seeds and a foreign protein expression level of 1%-4% of total soluble protein (TSP), large greenhouses could produce millions of doses of a vaccine for pennies per dose. Also contributing to the low cost of heterologous protein expression in soy are the well-known and established procedures for processing soybeans into palatable products. Lastly, soybeans are safe and do not pose significant risk for humans or livestock since soy is a major food staple and an integral component of most diets. Numerous products containing soy are currently consumed by a large number of humans and animals across the globe (Lusas and Riaz, 1995).

Soybean based pharmaceuticals could become extremely beneficial in developing countries. The World Health Organization estimates that nearly 3 million people die each year from vaccine preventable disease (Thomson 2006). In developing countries, people cannot afford the high cost of vaccines, but the use of cost-effective, soy-derived pharmaceuticals could address this issue. Also, in developing areas there is often a lack of infrastructure, including access to refrigeration, making it difficult to transport and store pharmaceuticals. Soy-based pharmaceuticals, whether shipped as processed powder or as a soymilk formulation, would have the potential to eliminate the current requirement for a cold chain and therefore offer practical alternatives. With soymilk formulations, this is possible because formulations could be lyophilized or dehydrated in a manner similar to that for making powdered dry milk, and then packaged, shipped, and stored without a requirement for a cold chain. Aliquoted doses could then be reconstituted on site and administered when needed.

As a plant, soybean offers another advantage over the use of microbial systems in that they possess the necessary machinery for generating large and complex proteins. Plant cells are capable of all eukaryotic post-translational modifications (Hood et al., 1997), including disulfide bond formation, glycosylation, and subunit protein assembly. Often, these post-translational processes do not have the drawbacks associated with other expression systems. For example, glycosylation in plant systems can be achieved without the hyperglycosylation observed in other systems such as yeast (Nakamura et al., 1993; Karnoup et al., 2005). Furthermore, plant cells do not harbor human or zoonotic pathogens, making them a safe host for the production of pharmaceuticals for both human and agricultural use. Finally, the concept of expressing a foreign protein in soybean has been proven by several groups who have established and attained feasible levels of protein expression (Piller et al., 2005 ; Ding et al., 2006 ; Morevec et al., 2007 ; Garg et al., 2007 ; Schmidt et al., 2008).

In this chapter, we will focus on incorporating our current knowledge for optimizing expression in soybeans and to evaluate this unique host as a platform for the expression of a wide variety of pharmaceutical proteins such as vaccine subunits.

2. Selection of a practical promoter to drive transgene expression

A critical first step for efficient production of heterologous protein in any recombinant system is to maximize the levels of foreign protein expression, as doing so will ultimately decrease production costs. However, reaching high levels of expression is a key limiting factor in the production of foreign proteins in a host system. Since transcription is one of the earliest steps in the process of protein production, it is an obvious place to focus when attempting to increase recombinant protein yield. To increase transcription levels, it is important to choose a promoter that can drive optimal transcription of a desired protein. Promoter sequences must be added for genes to be correctly translated into proteins, and because of this critical regulatory activity, promoters have been studied extensively in efforts to improve the efficiency of expression in transgenic host systems.

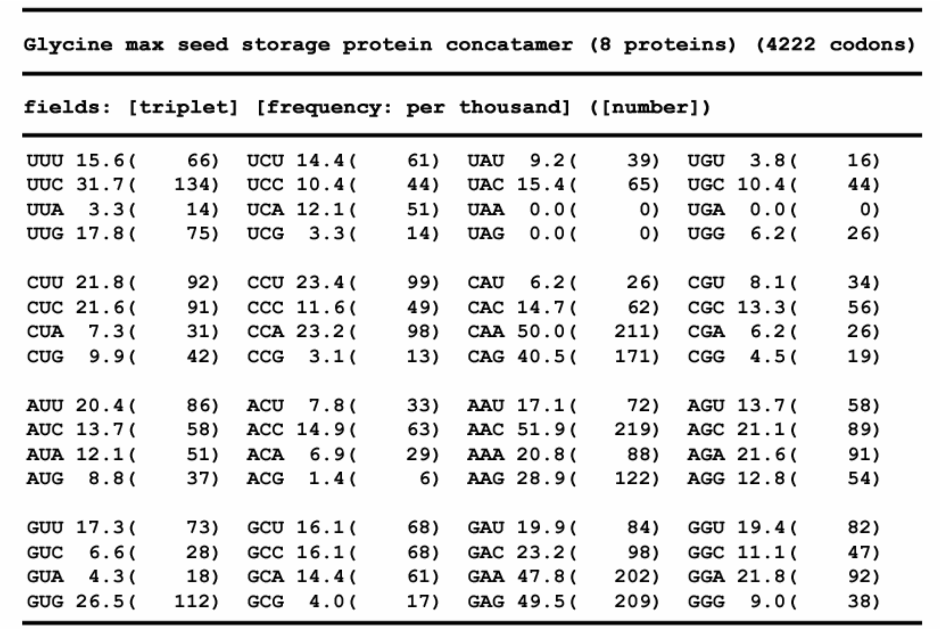
To date, some of the most common promoters used to express foreign proteins in transgenic crops have been constitutive in nature. Constitutive promoters drive gene expression in the majority of plant tissue types as well as throughout all life stages of plant, flower, and seed maturation. One of the most common constitutive promoters used for plant biotechnology applications is the cauliflower mosaic virus 35S promoter (35S CaMV; Odell et al., 1985) which has been shown to result in high levels of recombinant protein expression (Fiedler et al., 1997; Gutierrez-Ortega et al., 2005). While constitutive promoters have many advantages, there are also potential negative factors associated with these promoters that should be considered when designing a gene expression cassette. For example, constitutive promoters can consume unnecessary energy and resources, and can interfere with important agronomic factors such as growth habit and the ability to photosynthesize. Furthermore, expression in unwanted locations may require additional biohazard disposal (e.g. to remove unwanted parts of the plant) or containment (e.g. to prevent out crossing from pollen, etc.), both of which could negatively impact the cost-effectiveness of the system.

In contrast, tissue-specific promoters express spatially and temporally, allowing foreign proteins to be expressed in a desired location or at a desired point in the host life cycle. Seed-specific promoters are one example of tissue-specific promoters that can be used to target expression to different stages of seed development. Typical seed-specific promoters used for heterologous expression in soybean include the ß-conglycincin (7S) and glycinin (11S) promoters (Eckert et al., 2006; Nielsen et al., 1989). Several groups have used these promoters to successfully express recombinant proteins in soybean seeds (Ding et al., 2006; Moravec et al., 2007). While constitutive promoters can also drive expression in seeds (Zeitlin et al., 1998, Piller et al., 2005), it is likely that higher levels of expression in seeds can be achieved using seed-specific promoters.

3. Molecular optimization of synthetic genes

There are many challenges associated with expressing a functional protein outside of its native environment. The genetic information encoded in an open reading frame goes far beyond simply stating the order of amino acids in a protein. For example, different organisms prefer different codons when making a functional protein. The genetic code is comprised of 64 codons which encode 20 structural amino acids and three stop codons to terminate translation. Each codon is read in the ribosome by complementary tRNAs that transfer specific amino acids to a growing polypeptide chain. The overabundance in the number of codons allows many amino acids to be encoded by more than one codon. Different organisms often show particular preferences for one of the several codons that encode for the same amino acid. This phenomenon, called codon bias, has been identified as an important factor in gene expression. The degree to which any given codon appears in the genetic code can vary significantly between organisms. This is due to the fact that preferred codons correlate with the abundance of cognate tRNAs available within the cell. When a foreign gene is expressed in a non-native host, sequences that may contain expression-limiting factors and low-frequency codons in the host should be eliminated. There is a definite correlation between the frequency of rare codons and the level of foreign protein expression. Thus, it is generally understood that rare codons, especially at the N-terminus of a protein, should be avoided in the design of heterologous genes. Current improvements in the cost and speed of gene synthesis can facilitate the complete redesign of an entire gene sequence to maximize the likelihood of high protein expression. In this case, a gene of interest can be synthesized by replacing infrequently used condons with those that are preferred by soybean without changing the amino acid sequence.

Codon tables are available for most organisms. Such tables are compiled by tabulating the total number of codons present in all known genes for a particular organism, and then showing totals as a percentage or frequency. While the majority of available codon tables represent codon frequency of all genes within a particular organism, custom codon tables can be generated to represent codon frequency within a specific organ type. For example, Figure 1 shows a custom codon table that our laboratory generated for heterologous protein expression in soybeans. Since the majority of soybean protein is comprised of 7S and 11S storage protein, and our laboratory utilizes the soybean 7S and 11S promoters to target expression to seeds, we chose the eight most abundant proteins belonging to the 7S and 11S gene families to comprise a codon table based on these gene sequences. Thus, the custom table in Figure 1 reflects the codon bias of proteins residing in the environment that we are targeting for expression.



```
Glycine max seed storage protein concatamer (8 proteins) (4222 codons)

fields: [triplet] [frequency: per thousand] ([number])

UUU 15.6(    66)  UCU 14.4(    61)  UAU  9.2(    39)  UGU  3.8(    16)
UUC 31.7(   134)  UCC 10.4(    44)  UAC 15.4(    65)  UGC 10.4(    44)
UUA  3.3(    14)  UCA 12.1(    51)  UAA  0.0(     0)  UGA  0.0(     0)
UUG 17.8(    75)  UCG  3.3(    14)  UAG  0.0(     0)  UGG  6.2(    26)

CUU 21.8(    92)  CCU 23.4(    99)  CAU  6.2(    26)  CGU  8.1(    34)
CUC 21.6(    91)  CCC 11.6(    49)  CAC 14.7(    62)  CGC 13.3(    56)
CUA  7.3(    31)  CCA 23.2(    98)  CAA 50.0(   211)  CGA  6.2(    26)
CUG  9.9(    42)  CCG  3.1(    13)  CAG 40.5(   171)  CGG  4.5(    19)

AUU 20.4(    86)  ACU  7.8(    33)  AAU 17.1(    72)  AGU 13.7(    58)
AUC 13.7(    58)  ACC 14.9(    63)  AAC 51.9(   219)  AGC 21.1(    89)
AUA 12.1(    51)  ACA  6.9(    29)  AAA 20.8(    88)  AGA 21.6(    91)
AUG  8.8(    37)  ACG  1.4(     6)  AAG 28.9(   122)  AGG 12.8(    54)

GUU 17.3(    73)  GCU 16.1(    68)  GAU 19.9(    84)  GGU 19.4(    82)
GUC  6.6(    28)  GCC 16.1(    68)  GAC 23.2(    98)  GGC 11.1(    47)
GUA  4.3(    18)  GCA 14.4(    61)  GAA 47.8(   202)  GGA 21.8(    92)
GUG 26.5(   112)  GCG  4.0(    17)  GAG 49.5(   209)  GGG  9.0(    38)
```

Fig. 1. Codon Bias Table for *Glycine Max* (soybean) seeds. This custom table was assembled from codons present in highly expressed soybean 11S protein (G1, G2, G3, G4, G5) and 7S protein (α, α'and β subunits of β-conglycinin).

It is generally understood that the use of low frequency codons (e.g. those used <5-10% of the time by the host) can result in decreased protein yields. Such low frequency codons should be avoided and replaced with the preferred codon, or alternatively re-distributed at frequencies that represent the remainder of higher frequency codons. The elimination of rare codons, along with the utilization of a codon bias table, will greatly increase the probability for heterologous protein expression.

Another factor related to codon bias is the variation in %G+C composition between organisms. This difference is thought to contribute to variation in selection, mutational bias, and biased recombination-associated DNA repair. The design and use of synthetic genes offers a mechanism by which researchers can gain much greater control of heterologous protein expression. For example, the Bt gene for insect resistance is of bacterial origin and has a higher percentage of A-T nucleotide pairs compared to plants, which prefer G-C nucleotide pairs. Codon optimization of the Bt gene for expression in plants (e.g. substituting AT-rich codons with GC-rich codons that encode the same amino acid) resulted in greater expression of Bt protein in tobacco (Jabeen et al., 2010).

Although codon bias plays a large role in gene expression, other factors are also important. The nucleotide sequence surrounding the start codon, referred to as the Kozak sequence, is extremely important for maintaining optimal gene expression (Kozak 1984). The Kozak sequence is present in most eukaryotic mRNAs and plays a major role in the initiation of translation (De Angioletti et al., 2004). The Kozak sequence varies among eukaryotes, but has a general consensus of AACAAUGGC in plants (Lütcke et al., 1987). Therefore, any gene designed for expression in soybeans should contain a Kozak sequence or an acceptable variation.

While designing a synthetic gene or expression cassette, there are additional parameters that should be considered. Parameters such as destabilization sequences, cryptic splice sites, and premature polyadenylation sites frequently limit heterologous gene expression in plants. This makes it necessary to adapt and optimize the gene of interest towards the genetic requirements of the host organism. By taking these parameters into consideration when having a gene synthesized, it is possible to remove any sequences that may be detrimental to protein expression in plants. Many gene synthesis companies incorporate these parameters into the gene optimization algorithms that are used, and can design a gene with the maximum potential for high protein expression.

By having a gene of interest synthesized, it is possible to control other features of a gene. For example, restriction enzyme sites can be added or removed from synthetic sequences. Such sites can be used as either a diagnostic marker, or to facilitate downstream manipulations such as subsequent cloning. Furthermore, consideration should be given to any Southern blot analyses that may be necessary to evaluate the complexity of T-DNA insertion, and restriction sites should be added or removed accordingly. Sequences encoding histidine tags can also be added to synthetic genes, as these may be beneficial for downstream diagnostic detection (e.g. western analyses) or purification (using nickel resin).

When synthesizing a gene, it is also possible to generate inexpensive gene variants. Such gene variants can include alternative signal peptide sequences, organelle targeting and/or retention signals, and purification tags. If the gene is being synthesized by one of the many companies offering gene synthesis services, then variants containing such modifications (usually located at the 5′ or 3′ ends of the gene) can be generated simultaneously. Since variant modifications are typically small in size (e.g. 50-100 nucleotides) gene variants can be generated from a "master clone" for little additional cost. These variants can be used in independent transformation experiments to evaluate stability at multiple subcellular locations.

Finally, enhancers and leader sequences can be used to increase the ultimate yield of a transgenic protein. These sequences can be cloned adjacent to synthetic genes, or alternatively engineered into the gene design for ultimate synthesis by a commercial service provider. Our laboratory often utilizes the tobacco etch virus (TEV) leader sequence to enhance translation of foreign genes expressed in soybean. The TEV leader is naturally uncapped and yet a highly competitive mRNA during translation. In plants, the TEV leader has been shown to mediate cap-independent translation (Gallie and Walbot, 1992).

4. Cellular and subcellular targeting of heterologous proteins

One factor to consider when designing a gene cassette is that a protein can be targeted to a specific plant organ, tissue type, cell type, or subcellular target. The selection of a tissue or cell type is important since tissues and cells have specific biochemical microenvironments (e.g. pH, endogenous proteases, etc.) that may support, or alternatively be detrimental to heterologous protein expression. Since cells have a finite number of tRNAs, translation of recombinant mRNA transcripts can be limited. The endogenous protein content of cells comprising the targeted tissue may also impact stability. Subcellular targeting is another strategy that can be used to increase the yield of recombinant proteins since these compartments and locations can influence the interrelated processes of folding, assembly and post-translational modifications. Choices of subcellular compartments may include the cytosol, apoplast, endoplasmic reticulum (ER), chloroplast, or vacuole.

Subcellular targeting can be accomplished with the inclusion of a signal peptide. To date, numerous attempts to predict the correct subcellular location of proteins using machine learning techniques have been developed. Computational programs are publicly available that predict N-terminal signal peptides by predicting the presence of signal peptidase I cleavage sites. One convenient program that can be used to predict a signal peptide is SignalP. This program is easily accessed via the internet (http://www.cbs.dtu/services/SignalP/) and has a user-friendly interphase allowing the direct cut and paste of sequences to be analyzed with various readouts. In addition to predicting potential cleavage site(s) for a signal peptide, the SignalP program also provides a prediction score for the presence of a signal peptide. Thus, it is possible to compare the strengths of various signal peptide sequences when designing a gene.

Several groups have compared protein expression levels when targeting foreign proteins to different cellular and subcellular locations. For example, immunoglobulins and single-chain antibody fragments (scFv) were comparatively targeted to subcellular compartments and the greatest levels of expression were detected when these proteins were synthesized via the secretory pathway as opposed to being synthesized in the cytosol (Zimmermann et al., 1998; Schillberg et al., 1999). In this example, the secretory pathway appeared more suitable for the folding and assembly of the scFv protein and resulted in greater expression. Another group analyzed expression of two scFv proteins containing a plant signal peptide with and without the addition of an ER retention signal. In this case, a higher level of protein expression was achieved when the protein was retained within the endoplasmic reticulum by the addition of a KDEL retention sequence when compared to the same protein which lacked the retention signal and accumulated in extracellular spaces (Fiedler et al., 1997). A comparison of the human growth hormone (hGH) expressed in the apoplast, chloroplast, and the cytosol was performed by Gils et al., 2005 who reported the highest levels of hGH were achieved when targeted to the apoplast. Comparative studies with Lt-B targeted to either the cell surface, vacuole, ER, nucleus, plastids, or cytoplasm of maize showed a 20,000-fold difference in expression level with vacuoles being the preferred location for Lt-B accumulation (Streatfield et al., 2003).

While there are often choices that can be selected for the design and evaluation of what appears to be a rational target for heterologous expression on paper, there are often times when a heterologous protein dictates a specific target. For example, the EPSPS enzyme that confers resistance to glyphosphate in Roundup Ready crops is expressed in chloroplasts since the shikimate pathway - the target for glyphosate-based herbicide sprays – resides within chloroplasts. Similarly, expression of a vaccine antigen intended for oral delivery using soymilk formulations may require expression of the antigen in seed to avoid costly purification. Since each transgenic protein will have its own intrinsic properties, a preferential location for heterologous expression may not always be obvious, and therefore may need to be determined by systematic comparison.

5. Cellular and subcellular targeting of FanC in soybean

Almost a decade ago our laboratory set out to demonstrate the feasibility and practicality of using soybean as a platform for heterologous protein expression. This demonstration involved the targeting of different proteins to various cellular and subcellular locations, and characterizing the resulting expression. At the time these studies were initiated, few heterologous proteins had been expressed in soybean and an ideal location for maximal expression of heterologous proteins in this host was not known.

The model protein chosen for studies in soybean was FanC, a relatively small (~18 kDa) structural protein that is a major component of bacterial fimbrae. Analysis of the native FanC gene and protein sequences revealed several potential issues that could negatively impact expression in soybeans. These included (a) relatively low GC content, (b) codon bias for bacterial proteins, (c) potential destabilization sequences, (d) a cryptic splice junction, and (e) a cryptic polyadenylation sequence. Therefore, a synthetic FanC gene encoding the mature FanC protein was designed to ameliorate these concerns. A synthetic FanC gene was created by annealing complementary oligonucleotide primers (~100 nucleotides in length and containing ~30 nt of overlapping sequence) and using DNA polymerase to synthesize complementary strands. High fidelity DNA polymerase was used in a polymerase chain reaction (PCR) to create the full-length synthetic gene which was subcloned and then sequenced to verify integrity. While this assembly process was relatively long and tedious, the only other alternative was to have the entire gene synthesized and such synthesis services were cost-prohibitive. However, within the past decade the cost for gene synthesis services has decreased significantly, and now entire genes can be synthesized for less than 50¢ /base pair with amazing speed (1-2 week turnover time).

Three different gene cassettes were designed to target synthetic FanC to various tissues and subcellular locations. The first gene cassette (pKP7) utilized the 35S constitutive promoter and TEV leader to drive constitutive cytosolic expression of FanC. The second gene cassette (pKP8) utilized the same 35S promoter and TEV leader but also contained a sequence encoding a chloroplast targeting peptide (CTP) fused to the 5′ end of FanC. A third gene cassette (pKP23) contained the β-conglycinin 7S seed-specific promoter and TEV leader to target expression to maturing seeds. All three constructs utilized the polyadenylation site present in the 35S terminator sequence. A summary of the elements used in these constructs is shown in Figure 2.

Following *Agrobacterium*-mediated transformation, transgenic plants were taken to maturity and transgenic seed was collected. Genomic DNA was extracted from T_1 leaf material and used to verify the presence of the T-DNA while protein was extracted from different tissues and used for detection of soy-derived FanC in western blot experiments. For plants transformed with pKP7, leaf and seed (cotyledon) tissues were both examined. Seventeen initial T_0 events were generated, and 10 of these expressed FanC protein in leaves at various levels. To quantify expression, recombinant FanC protein containing a C-terminal histidine tag (GGHHHHHH) was overexpressed in *E. coli*, purified, quantified, and then included in

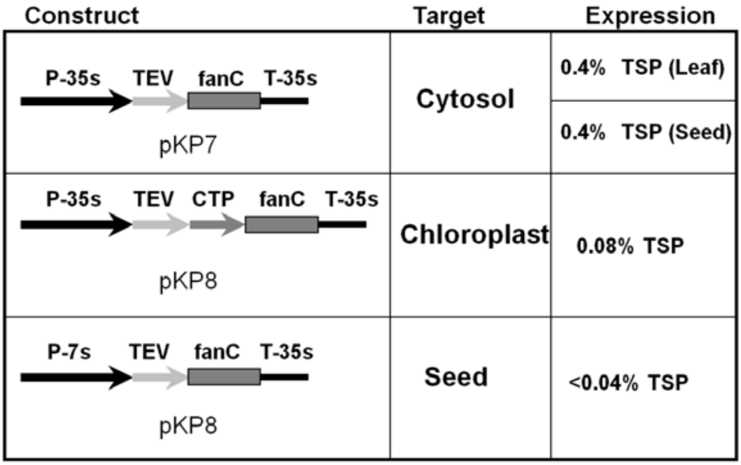


Fig. 2. FanC construct design and summary of protein expression. Four previously designed expression cassettes for expression of FanC in different cellular locations with different promoters and expression elements.

western analyses as a quantification standard. X-ray films of western blot experiments were subjected to densitometry and the amount of soy-derived FanC from a known volume of protein extract was used to calculate overall expression as a percentage of total soluble protein (TSP). Figure 3 shows that cytosolic expression of FanC represented ~0.4% of TSP in leaf extracts. Expression at 0.4% TSP was respectable and several-fold higher than other reports of heterologous protein expression in leaves (Kapusta et al., 1999; **A**ndrews and Curtis 2005). Since the 35S promoter drives expression in most tissue types, we also examined protein derived from T_1 seed (cotyledon) chips for the presence of FanC. Western analyses showed that FanC also accumulated similarly in seed tissue, to levels representing ~0.4% of seed TSP (Piller et al., 2005).

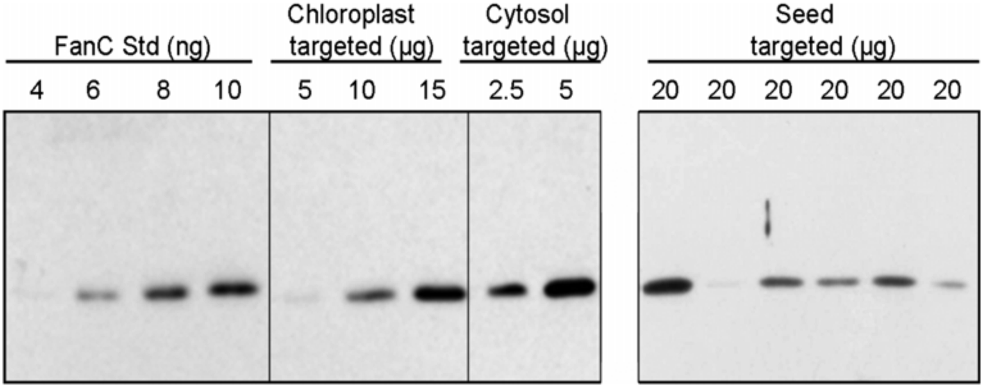


Fig. 3. Comparison of FanC protein expression levels when targeted to chloroplasts, cytosol and seed (amounts represent total soluable protein of seed). Known amounts of bacterially-derived FanC were included as standards for quantification.

To examine FanC expression in plants transformed with the pKP7 gene construct, T_1 leaf material from seven transgenic lines was characterized. To ensure that protein was being correctly targeted, chloroplasts were isolated from maturing leaves. The chloroplasts were then lysed and chloroplast protein was subjected to western analyses for the presence and quantification of FanC. Figure 3 shows the relative level of FanC in one representative chloroplast extract. Quantification against known standards of FanC revealed that FanC represented ~0.08% of chloroplast TSP or ~ 5-fold less than leaf extracts.

To examine expression of FanC in plants transformed with the pKP23 gene construct, protein extracts were prepared from T_1 seed chips (derived from 11 independent events), and subjected to western analysis. The right panel of Figure 3 shows the relative levels of FanC protein in T_1 seed extracts derived from six independent lines. One of the lines expressed a greater level of FanC relative to the other five. Quantification of FanC protein in that line showed transgenic expression representing ~0.04% seed TSP. In this case, the level of expression was approximately 10 fold less than the expression previously observed in seeds transformed with the pKP7 construct.

In each event characterized, anti-FanC antibodies recognized a novel protein with mobility consistent with that for mature FanC (~18 kDa). The mobility of chloroplast-targeted FanC is of particular relevance since that particular protein was engineered with an N-terminal chloroplast targeting peptide (CTP) fusion protein. Figure 3 showed the mobility of chloroplast FanC was identical to that of leaf and seed-derived FanC suggesting that the CTP was recognized and properly cleaved. Since bacterially-derived FanC contains a C-terminal histidine tag, it has a slightly slower mobility (~19 kDa) relative to soy-derived FanC (~18 kDa). This ~1 kDa difference was observed in Figure 3, and further supported the likelihood that all soy-derived FanC proteins contained identical or nearly identical N-termini.

A summary of FanC expression in various T_1 protein extracts following subcellular targeting is summarized in Figure 2. The lowest levels of FanC were detected in chloroplasts (0.08% of chloroplast TSP) while the greatest levels of transgenic protein were observed in the leaf and seed when targeted to the cytosol (up to 0.4% of seed TSP). It should be noted that except for events transformed with pKP7, quantification of FanC was determined using T_1 generation chloroplast (pKP8) and seed (pKP23) protein extracts. Since homozygosity was not analyzed in the T_1 generation for these events, it is possible that the tissue chosen for analysis and FanC quantification was derived from a heterozygous genotype. While it is possible that FanC expression could decrease in subsequent generations, based on our experience, it is more likely that FanC expression would increase once homozygosity is achieved and protein expression is stabilized (see Section 9).

The increased FanC expression detected in seeds transformed with pKP23 led us to conclude that seed-specific promoters may significantly impact heterologous protein expression. Other groups have used seed-specific promoters to achieve heterologous protein expression levels approaching 2-3% of total soluble seed protein (Ding et al., 2006; Moravek et al., 2007). Moravek et al., 2007 used electron microscopy to analyze subcellular localization of a heterologous protein, and in this case, a signal peptide was included to target recombinant protein expression to the secretory pathway for ultimate accumulation in protein bodies. It should be noted that the FanC gene used in our studies did not contain a signal peptide and, therefore, was not likely to be synthesized through the secretory pathway. As a result of this design, it is likely that all versions of FanC protein remained cytosolic. Analysis of the native FanC protein sequence deposited into GenBank revealed an ORF containing additional N-terminal amino acids not present in the mature sequence.

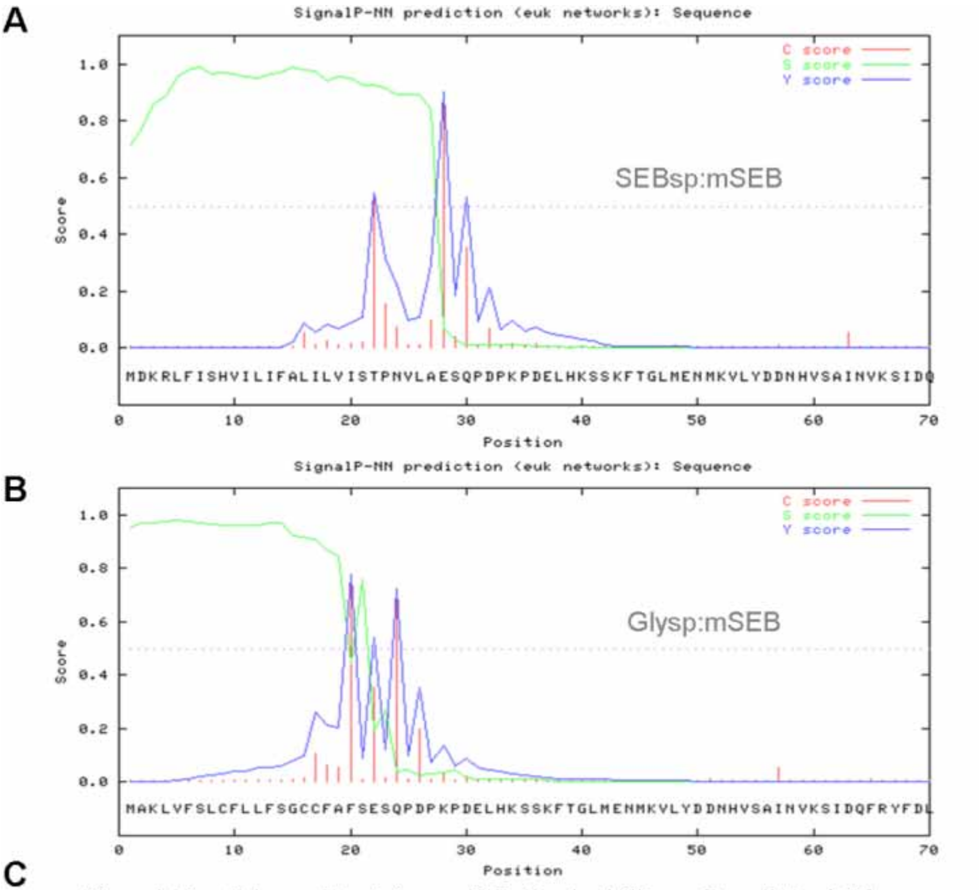
Signal peptide software programs suggested that the N-terminal amino acids fit the criteria of eukaryotic signal peptide sequences. Thus, if the N-terminal sequence of FanC was present and recognized as a signal peptide then translation and maturation would have occurred via the protein secretory pathway. Proteins derived from this pathway might then accumulate in protein bodies, vesicles, or the apoplast. These environments are biochemically different than the cytosolic and chloroplast environments and could greatly impact overall stability and accumulation of FanC. While we have not determined that targeting FanC to the secretory pathway increases expression, the inclusion of a signal peptide will likely maximize stability and accumulation of most proteins targeted to the seed (see section 7).

6. Subcellular targeting in soybean seed using different signal peptides

Our group has used the SignalP program to predict the putative presence, location, and strength of two different signal peptides attached to a mutated form of bacterial Staphylococcal enterotoxin B (mSEB), which is currently under development for use as a potential soy-based vaccine candidate in our laboratory. By nature, SEB is a secreted protein. Analysis of the native SEB sequence revealed a 5′mRNA sequence that could potentially encode a signal peptide. Others have shown that bacterial sequences can be recognized by the plant machinery and result in correctly processed proteins (Moeller et al., 2009). To determine whether the native SEB signal peptide sequence might function in plants, the amino acid sequence of the mSEB ORF was examined using the SignalP software program (Figure 4 A). This program indicated a high probability (0.94) for the presence of an N-terminal signal peptide and indicated potential cleavage sites, with the predominant cleavage site being NVLA/ESQP (Figure 4 C). Since the reported N-terminus of mature mSEB begins with ESQP, we concluded that inclusion of the 5′ signal peptide may allow transgenic protein synthesis to occur via the secretory pathway and, therefore, result in a protein that was correctly processed and either secreted or packaged internally.

In theory, if transgenic mSEB is capable of passing through the membrane, it may localize within the apoplastic space; alternatively, if it is not capable of translocating to the membrane, it would likely be retained within the cell. Both locations represent very different biochemical microenvironments, and recombinant forms of mSEB may prefer one environment over the other. Such a preference could result in differential stability and/or abundance of mSEB. To further explore these ideas, a gene cassette was designed in which the 7S β-conglycinin promoter and TEV leader elements were used to drive expression of a soy codon-optimized mSEB gene containing the native 5′ signal peptide. This gene expression cassette was designated 7S-mSEB and is shown in Figure 5A.

For comparison, a second mSEB gene cassette was designed using a known plant signal peptide sequence (Figure 5B). The soybean 7S and glycinin promoters are both strong seed-specific promoters used within the industry to target heterologous protein expression to seeds. Since others have successfully used the glycinin promoter and endogenous signal peptide to overexpress proteins in plant systems (Ding et al., 2006; Moreavec et al., 2007), we chose to utilize these same elements in our evaluation of subcellular targeting. The amino acid sequence comprising the glycinin signal peptide fused to mature mSEB was analyzed using the SignalP program to determine whether the fused glycinin signal would be recognized as a putative signal peptide. The program indeed predicted a signal peptide and with high probability (Figure 4 B). However, the program did not predict a cleavage at the



C

Signal Peptide	Protein	SP Probability	Predicted Cleavage
SEB	mSEB	0.942	NVLA-ESQP
GLY	mSEB	1.000	CCFA-FSES

Fig. 4. Analysis of signal peptide sequences using the SignalP program. SignalP graphical analysis output of presence and location of signal peptide cleavage sites in amino acid sequence of mSEB with the native SEB signal peptide(A) and mSEB with a soybean glycinin signal peptide (B). C. Table representing the signal peptide probability score along with the predicted signal peptide cleavage sites.

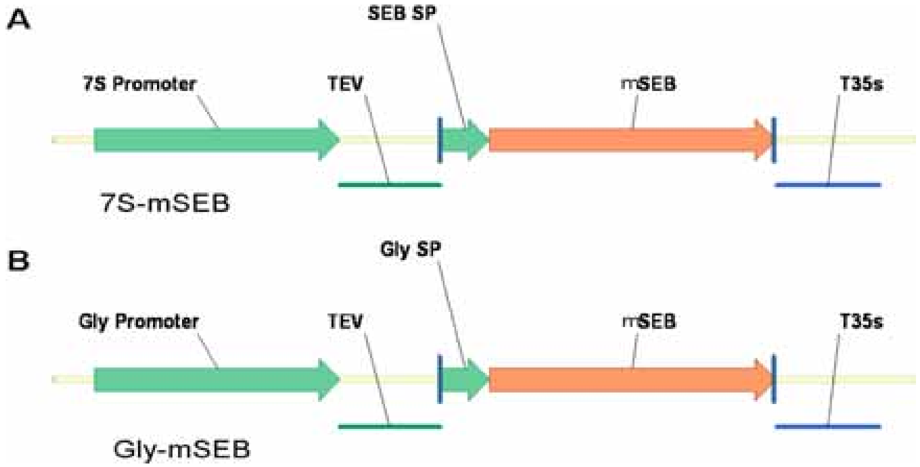


Fig. 5. Gene cassettes used to target mSEB to the secretory pathway. A. The 7S-mSEB gene cassette contains the 7S seed-specific promoter, TEV leader, native SEB signal peptide, mSEB gene, and 35S terminator elements. B. The Gly-mSEB gene cassette contains the glycinin seed-specific promoter, TEV leader, soybean glycinin signal peptide, mSEB gene, and 35S terminator elements.

precise glycinin/mSEB junction designed on paper. Instead the major cleavage was predicted to occur adjacent (two amino acids) to the fusion junction (CCFA/FSEQ) resulting in a protein containing the amino acids “FS” at the N-terminus. It is unclear how these N-terminal residues might impact stability of the protein. The SignalP program also predicted an alternate cleavage at the FAFS/ESQP junction (Figure 4 C), suggesting that a mature protein with a correctly processed N-terminus (ESQP) might also be generated. Therefore, we did not attempt to alter the surrounding amino acids. If the SignalP program did not predict a signal peptide, or alternatively predicted a cleavage site that was not anticipated or intended, manual changes to the protein could have been made in an effort to identify a strong cleavage site in the desired region. In the past we have had to make such changes in designing other gene cassettes in an effort to increase the probability of obtaining a desired mature protein product. In those cases, alterations of only 1-2 amino acids were incorporated and the sequences were continuously run through the SignalP program until an acceptable output was achieved. Since soybean glycinin resides in protein bodies, we anticipated that mSEB fused to the glycinin signal peptide may also be targeted to protein bodies.

Following transformation of soybean with each of the above cassettes, resulting transgenic seed were identified and protein expression was characterized. Both constructs resulted in the accumulation of mSEB to levels >1% of total soluble protein in T_1 seeds. Visualization of the recombinant protein was carried out using western blot analyses. The predicted molecular mass of synthetic mutant SEB is ~29 kDa, and this is the precise size of transgenic protein detected following SDS-PAGE and western analysis (Figure 6).

It is important to note that there is only one single band diagnostic of the recombinant mSEB protein, and the mobility of this novel protein is consistent with the predicted size of mature SEB (~29 kDa) suggesting that soy-derived mSEB was correctly processed. It is important to note that while both gene cassettes predicted pre-proteins with different mobilities (due to

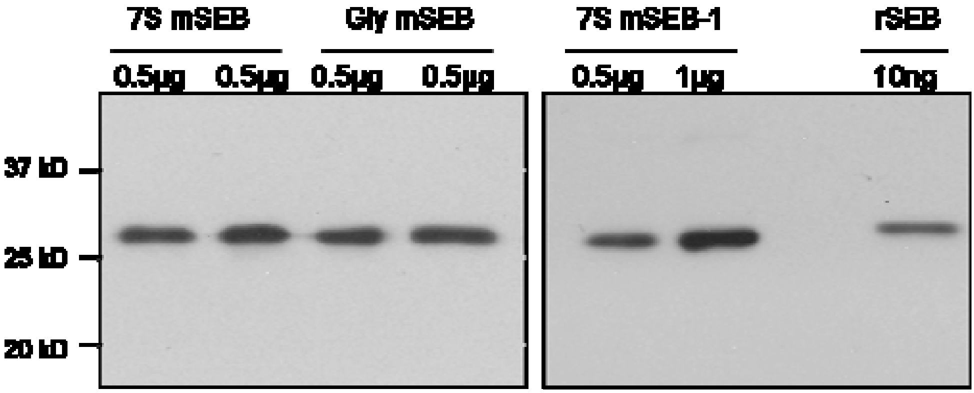


Fig. 6. Western blot analysis showing protein expression of 7S-mSEB and Gly-mSEB. Bacterially-derived mutant SEB (rSEB) served as a positive control and relative quantification standard. Amounts represent total soluable protein of seed.

the length of the signal peptide), the detected proteins showed similar mobility when separated in SDS-PAGE experiments. If correct processing of the N-terminus had not occurred, the putative signal peptide would still be attached to the N-terminus resulting in molecules showing altered motility when separated in SDS-PAGE gels. Furthermore, bacterially-expressed recombinant mutnat SEB (rSEB) containing a histidine tag (GGHHHHHH) was included as a positive control. The rSEB is predicted to migrate slightly slower than soy-derived mSEB due to the presence of the histidine tag at the C-terminus. This difference was also detected in western analyses (Figure 6). While these data suggest that both soy-derived proteins were correctly processed, definitive conclusions can only be drawn following analytical characterization of the N-termini. In this respect, N-terminal sequencing was performed on mSEB isolated from T_1 seeds transformed with the 7S-mSEB cassette. Results from this experiment verified the composition of 11 amino acids present at the N-terminus which were identical to those reported in the GenBank database (K. Piller, unpublished results). N-terminal sequencing is currently in progress to identify terminal residues in the glycinin-mSEB fusion protein.

We also visualized localization of the soy-derived mSEB proteins under control of the two different promoters and signal peptides using immunohistochemistry and confocal microscopy. Seed coats were removed and cotyledons were fixed in formaldehyde and embedded in paraffin. The tissue was incubated with a rabbit anti-SEB antibody followed by an Alexaflour^{488} goat anti-rabbit IgG-HRP conjugated secondary antibody. Images were collected with a LSM 710 Spectral Confocor 3 Confocal Microscope using a 20X objective and a 405nm laser to visualize DAPI stained nuclei along with a 488nm laser to collect emitted fluorescence. Figure 7A shows that mSEB protein derived from the 7S promoter and native SEB signal peptide construct was not retained within the plant cell and was instead secreted into the apoplastic spaces between cells. The localization of mSEB to apoplastic spaces was anticipated. Figure 7B shows a wild type control seed section for comparison. The mSEB protein derived from seeds transformed with the glycinin promoter and glycinin signal peptide, in contrast, remained intracellular and was localized throughout the cytoplasm as shown in Figure 7C, with a wild type control in Figure 7D. While this result was not completely surprising, it suggested that amino acid sequences present within the glycinin signal peptide may have contained "address tags" which dictated final localization (Mølhøj and Del Degan, 2004).

Collectively, these data show the utility of signal peptides and prediction software to aid in the design of gene cassettes targeting expression to subcellular locations. In the case of mSEB, we found that a native bacterial sequence and a known soybean signal peptide both functioned as signal peptides in vivo. Furthermore, both gene constructs allowed mSEB to accumulate to levels representing >2% seed TSP. While relatively small numbers of transformed lines (<10) were screened in this comparative study, the data nonetheless show

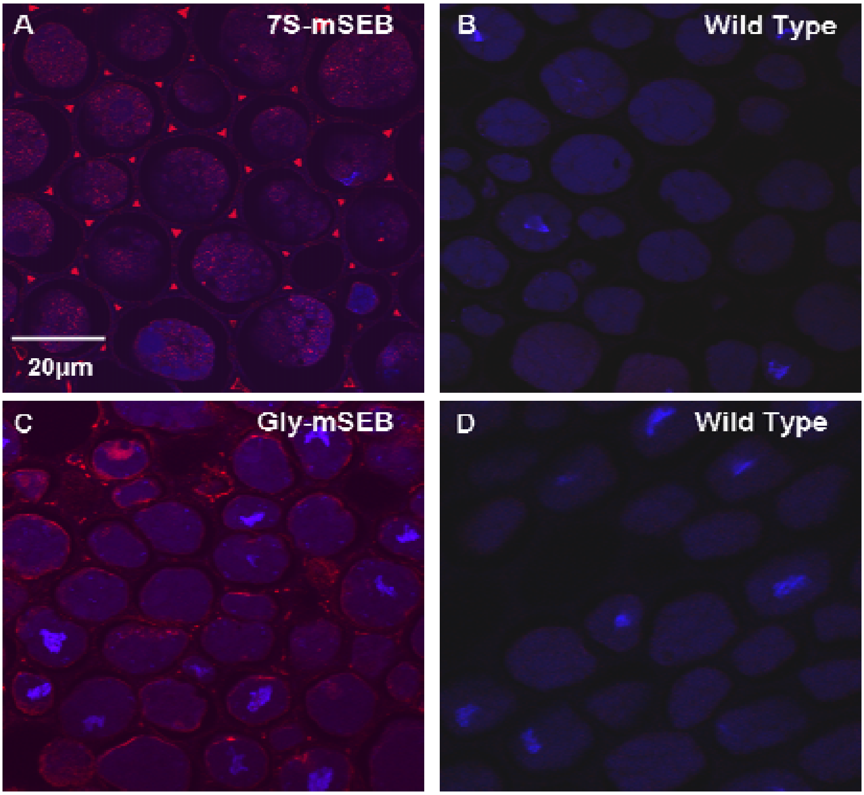


Fig. 7. Immunohistochemical detection of mSEB. 10µm sections of paraffin-embedded transgenic seed tissue were processed and incubated with rabbit serum containing anti-SEB antibodies followed by an Alexafluor goat anti-rabbit IgG secondary antibody. Nuclear material was stain with DAPI shown in blue. Samples were viewed at 20X magnification using confocal microscopy. Identical parameters on the microscope were used for photography on all tissue samples. A. 7S-mSEB refers to the gene cassette containing the 7S promoter, TEV leader, native SEB signal peptide, mSEB, and 35S terminator gene cassette. B. Wild Type control seed section under same conditions as A. C. Gly-mSEB refers to the gene cassette containing the glycinin promoter, TEV leader, glycinin native signal peptide, mSEB, and 35S terminator. D. Wild Type control under same conditions as C.

that practical levels of mSEB can accumulate both intracellularly and extracellularly. This may be due to the fact that mSEB is a toxin and, therefore, is stable under a variety of conditions. It is probable that other transgenic proteins may favor one environment over the other, and if true, then multiple gene constructs would need to be designed and tested.

7. Configuration of gene of interest cassettes in binary vector design

Another factor that can potentially impact expression of a foreign gene is the orientation of gene cassettes within a transformation vector. Most binary vectors designed for use with *Agrobacterium*-mediated transformation of soybean contain at least two independent gene cassettes. One of these cassettes is used for plant selection and typically contains either a reporter gene such as GUS for substrate detection (Jefferson et al., 1987), GFP for fluorescence (Chalfie et al., 1994), a selectable marker gene that allows selection in the host by conferring resistance to antibiotics (e.g. *kan* for kanamycin selection) or herbicides (e.g. *bar* for bialophos selection, *cp4 epsps* for glyphosate selection). Constitutive promoters such as 35S (from cauliflower mosaic virus), NOS (from nopaline synthase), and FMV (from figwort mosaic virus) drive expression of the plant selectable markers (PSM). The PSM gene cassette is located between short sequences referred to as the right border (RB) and left border (LB) repeats. The RB and LB sequences are derived from *Agrobacterium* and define the borders of the T-DNA segment that will be transferred from the bacteria to the host genome. The PSM gene cassette is usually placed adjacent to the left border sequence. Most binary vectors have been engineered to contain a short linker sequence located between the PSM cassette and the RB. This linker sequence usually contains one to several restriction enzyme sites that allows for cloning of additional gene cassettes. Collectively, these elements comprise a plant binary vector backbone.

Gene cassettes designed to express a particular gene of interest (GOI) can be cloned into a binary vector backbone in either of two directions. In one scenario, the positioning of the GOI and selectable marker cassettes exhibit the same directionality with respect to regulatory elements and the genes being transcribed. In this type of cloning, the terminator element of the GOI cassette is adjacent to the promoter of the PSM cassette. Our laboratory refers to these events as being cloned in the “forward” direction. In a second scenario, the positioning of the GOI and PSM cassettes exhibit opposite directionality with respect to regulatory elements and the genes being transcribed. In those cases the GOI and PSM promoters are adjacent and drive transcription of their respective genes in opposite directions. We refer to such events as those in which the GOI is cloned in the “reverse” direction.

There has been speculation as to whether foreign gene expression can be impacted by the directionality of the GOI and PSM cassettes in the plant transformation vector. The majority of binary vectors used in studies reported in the literature favor a forward directionality in which the GOI and plant selectable markers are adjacent and transcribed in the same direction. There are theoretical explanations that propose the greater functionality of one vector over the other. For example, promoter occlusion (also known as transcriptional interference) is a phenomenon in which RNA polymerase from an upstream gene fails to terminate and interferes with transcription of a downstream gene. Evidence of this transcriptional interference has been reported in plants (Thompson and Myatt 1997). One could argue that binary vectors containing the GOI cloned in the forward direction run the risk of decreasing transformation efficiencies if promoter occlusion were to disrupt expression of the selectable marker gene. This was found to be the case by Padidam and Cao 2001, who did a direct comparison of two genes aligned in the forward orientation and found that the expression of the downstream gene was reduced 80% by the upstream gene. Another argument in support of vectors with reverse-oriented cassettes is the potential for synergistic enhancement of promoter activity. By this logic, two promoters would have a greater probability of recruiting trans-regulatory protein factors than would a single promoter. In this case, the adjacent promoters would both be at an advantage and could result in increased expression of their respective genes. However, others could argue that adjacent promoters may not be synergistic, but instead antagonistic. In such cases, a large regulatory protein or protein complex bound to one promoter could block or hinder the binding of regulatory proteins to an adjacent promoter. This could be especially true if the promoters were different with respect to the trans-regulatory factors being recruited, which is the likely case in most scenarios. Thus, there appears to be several logical reasons for the use of binary vectors containing GOI and PSM cassettes cloned in both orientations.

In a study to directly compare the effects of GOI and PSM cassette positioning with respect to foreign gene expression, we designed binary vectors with the GOI cassette cloned in

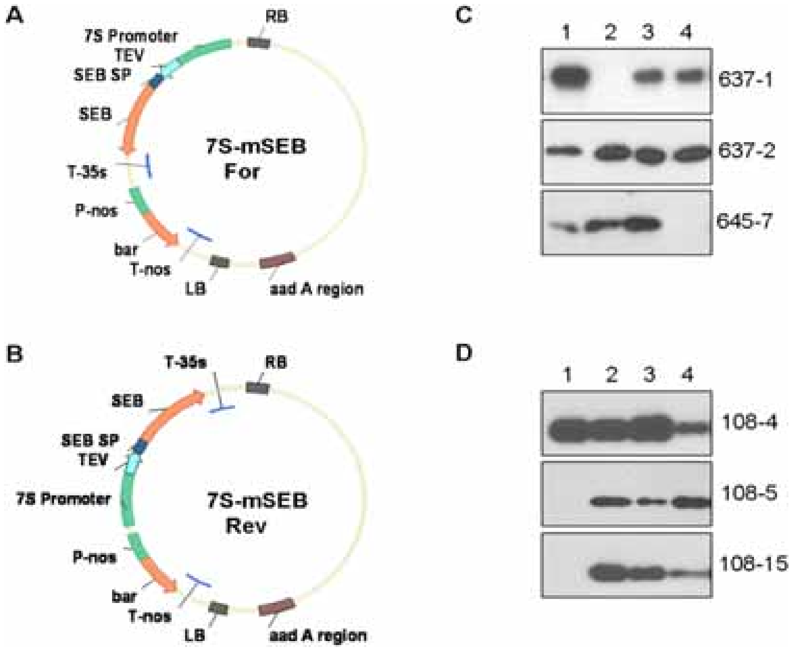


Fig. 8. Gene cassettes in opposing orientation and western analysis of T_1 progeny from events transformed with both orientations . A. The 7S-mSEB gene cassette in the forward orientation containing the 7S seed-specific promoter, TEV leader, native SEB signal peptide, mSEB gene, and 35S terminator elements. B. The 7S-mSEB gene cassette in the reverse orientation containing the same elements as the forward version. C. Western blot analysis showing protein expression of seeds from a transformation event with the cassette in the forward orientation. D. Protein expression from a transformation event with the cassette in the reverse orientation.

either the forward or reverse direction. The GOI cassette contained the 7S seeed-sprecific promoter, TEV leader, native SEB signal peptide, mutant SEB (mSEB), and 35S terminator elements. This cassette was assembled in a shuttle vector, and then excised using HindIII restriction sites flanking the cassette and cloned into the HindIII site present in the linker region of binary vector backbone. Following ligation and bacterial transformation, individual colonies were screened to identify clones with forward and reverse directionalities of the GOI cassette. These clones were then transformed into *Agrobacterium* and used for soybean transformation. Plasmid maps depicting the forward and reverse clonings are shown in Figure 8A and B.

Following transformation and regeneration, T_1 seeds were collected and protein extracts were prepared from four individual cotyledon chips. Equal amounts of protein were used in SDS-PAGE experiments, and western analysis was used for detection of recombinant mSEB protein. Results from these experiments are shown in panels C and D of Figure 8. Both binary vectors resulted in production of mSEB protein. Segregation of the transgene was observed in four of the six events, indicative of low T-DNA insert complexity. Additional evidence for segregation of the transgene was suggested by the variation in mSEB expression within lines. Conclusions regarding the ultimate level of mSEB expression cannot be drawn at this time since homozygosity has not yet been achieved in lines transformed with the reverse orientation cassette. We are currently in the process of generating T_2 and T_3 seed for the ST108 lines. Once homozygosity for the T-DNA has been achieved, more meaningful results regarding foreign protein accumulation levels can be concluded. Based on the western results shown in Figure 8, and our prior experiences, it is likely that recombinant protein expression levels derived from either construct will be similar. If this is the case, the directionality of the GOI cassette relative to the PSM cassette may not significantly impact expression levels of this particular foreign protein.

8. Stable recombinant protein expression over generations

Transgene expression is influenced by several factors that cannot be controlled through construct design, and these factors can lead to variable transgene expression. For practical applications, expression of a heterologous protein must remain stable from one generation to the next. However, expression levels and patterns of transgene inheritance generally show wide variation between independently transformed plants carrying the same construct (Spencer et al., 1992; Walters et al., 1992). This variation can be the result of several factors including integration site, transgene copy number, locus configuration, and epigenetic silencing mechanisms. There are several reviews that discuss these factors in detail (Iyer et al., 2000; Matzke et al., 2000). For self-pollinating species, homozygous plant lines are often the basic material used for molecular genetic studies. However, initial assessment of transgene expression and associated phenotype is often carried out using primary transformants that are segregating populations containing hemizygous progeny. Thus, it is possible that expression levels in a segregating population may not mimic expression levels in stable homozygous populations. As a result, conclusions regarding ultimate expression levels should be drawn only after homozygosity has been achieved and stable expression has been observed for multiple generations.

Agrobacterium-mediated transformation is one of the preferred methods for soybean transformation. This process was first described in Horsch et al., 1985 and has undergone numerous modifications in an effort to increase efficiency and reduce the time and cost associated with tissue and plant regeneration. Through the course of these improvements, and over time, conditions have been optimized which allow the majority of transformed events to result in single copy T-DNA insertions (Hobbs et al., 1990; Caligari et al., 1993; Bhattacharyya et al., 1994; Mlynárová et al., 1996). It has been shown that single transgene insertions are inherited in a Mendelian fashion (Chyi et al., 1986; Muller et al., 1987; Puonti-Kaerlas et al., 1992) making it possible to identify and segregate away null lines and to focus on stable homozygous lines for further analysis of protein expression levels.

Several genetic approaches have been applied to increase expression, including increasing the transgene copy number through selfing or crossing, and introducing foreign genes into germplasm suited to their overexpression. With stable transgenics, taking a single transgene to homozygosity through selfing doubles tranasgegne representation and typically increases expression (Zhong et al., 1999). Also, crossing high-expressing transgenic lines arising from independent transformation events can increase transgene copy number and expression levels. Such approaches involving crossing to increase transgene copy number appears to have a reduced risk of gene silencing when compared with transgenic events containing multiple T-DNA insertions.

In our laboratory, we typically analyze T_1 seed chips for the presence of a transgene as the first step in our screening process. Since the transformation process used to generate the transgenic events is optimized for single T-DNA insertion events, we expect to observe Mendelian segregation in many of the events. While this is often the case, there are also instances when segregation does not follow Mendelian genetics, even when it is later determined that the T-DNA insertion was a single copy insertion. Since initial screens can be misleading, all positive T_1 seeds are planted and taken to maturity as a rule of thumb. Since T_1 seeds are segregating and the majority of transgenic events are likely heterozygous with respect to the T-DNA, it is possible to see an increase in hetetrologous protein expression when T_2 seed extracts are compared to parent T_1 seed extracts. The likely explanation for this observation is that segregation in single-copy events has resulted in homozygosity with respect to the T-DNA locus, and, therefore, a doubling in the number of genes that can produce transgenic protein.

Figure 9A shows one example of increased expression over generations using the model protein mSEB. Equal amounts of protein from a parent T_1 seed extract and four individual T_2 progeny seed extracts were analyzed by western analysis for comparison of mSEB expression.

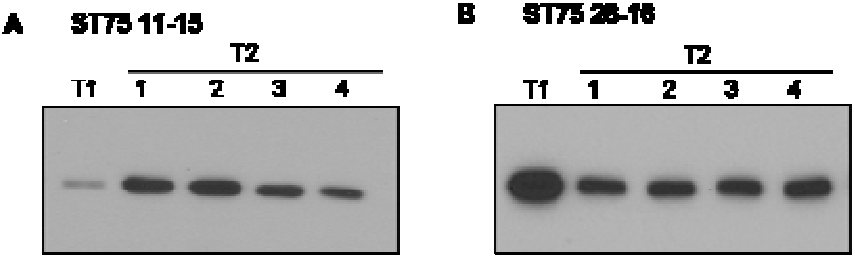


Fig. 9. Western blots showing fluctuation in transgene expression levels between generations. A. Equal amounts (1µg) of seed protein from a T_1 parent and four T_2 progeny were examined to reveal an increase in mSEB expression. B. Equal amounts (1µg) of seed protein from a T_1 parent and four T_2 progeny were examined and reveal a decrease in mSEB expression.

In this example, four different T_2 progeny showed >2-fold increases in mSEB expression compared with the T_1 parent. While a 2-fold increase in expression correlates with doubling in copy number, it is not clear why two to four-fold expression increases are often detected. One explanation for this observation is that a transgene requires a full generation to "re-set" itself while another could have to do with differences in seed quality.

While a two to four-fold increase in protein expression is welcome, the converse is observed and heterologous protein expression decreases between the T_1 and T_2 generations. One such example is shown in Figure 9B. In this experiment, equal amounts of seed extract from a T_1 parent and four individual T_2 progeny were also characterized by western analysis. Several factors could contribute to decrease in protein expression. For example, multiple copies of the gene cassette may have been inserted during the transformation process, and these additional copies may be responsible for gene silencing and an associated decrease in protein expression. While we observe decreases in expression such as those shown in Figure 8B, they are not as common as events which show increases in expression such as those in Figure 8A. Therefore, when possible, we prefer to quantify expression once homozygosity is achieved and recombinant protein production has had a chance to stabilize.

9. Simple methods for purification of foreign proteins

Until this point the current chapter has mainly focused on strategies to achieve optimal expression of foreign proteins using soybean as an expression platform. Given the high protein content of soybeans, and expression levels typically between 1% and 4% of total soluble seed protein, the generation of milligram quantities of transgenic protein is usually attainable. In those cases fractionation or purification of the desired protein is not necessary. However, in cases where foreign proteins are not expressed at desired levels despite exhaustive attempts using known methodologies to increase expression, there are simple time and cost-effective methods that can be employed to generate a partially purified and more concentrated product.

Any given protein may represent a small or large amount of the total protein composition within a cell. These proteins can be soluble or insoluble, membrane-bound, DNA-bound, cytoplasmic, or localized within organelles. If purification of a heterologous protein is needed, the challenge would be to separate that protein from all other components in the cell with reasonable efficiency, speed, yield, and purity, while still maintaining the biological activity and chemical integrity of the polypeptide.

Proteins vary in a number of their physical and chemical properties due to size and amino acid composition. Amino acid residues may be positively or negatively charged, neutral and polar, or neutral and hydrophobic. Proteins also have very definite secondary and tertiary structures which create a unique size, shape, and distribution of residues on their surface. It is possible to exploit the differences in properties between the protein of interest and other host proteins to establish partial purification of a heterologous protein.

One simple and cost-effective way to partially purify a heterologous protein is to take advantage of its unique isoelectric point (pI). The pI of a protein is the specific pH at which that protein has no net electrical charge. The pI affects the solubility of a protein at a given pH, with minimum solubility in solutions with a pH corresponding to their pI value. Thus, a protein usually remains soluble in a solution when the pH is either higher (net negative charge) or lower (net positive charge) than its pI value.

The majority of soluble protein in soymilk is generally classified as 11S protein, 7S protein, and whey protein (Nielsen et al., 1989). Soymilk proteins within each of these three classes have relatively low pIs. Altering the pH of a soymilk solution, therefore, allows for a relatively simple method to precipitate major classes of proteins (Thanh and Shibasaki 1976). For example, decreasing the pH of a soymilk solution to 6.4 will precipitate proteins in the 11S family while further decreasing the pH to 4.8 will precipitate proteins in the 7S family. Thus, if the isoelectric point of a heterologous protein is greater than pH 6.4, isoelectric fractionation can be used to remove 11S and 7S proteins which comprise the majority of protein in soymilk solutions.

To demonstrate the utility of isoelectric fractionation, we prepared soymilk from transgenic seeds expressing the mSEB protein. There are several software programs that can be used to predict the pI of protein based on known amino acid sequence. One of these programs was used to predict a pI 8.52 for mSEB. Based on this pI value, we anticipated that soy-derived mSEB would remain soluble in a soymilk solution as the pH of the milk was lowered, allowing precipitation of 11S and 7S proteins. To test our hypothesis, soymilk was prepared in a 1X PBS solution. The pH of the starting soymilk solution was 7.8. A dilute solution of HCL was added dropwise to lower the pH of the soymilk to pH 6.8. At this point, the milk solution became opaque, presumably due to the precipitation of 11S proteins. The precipitated milk proteins were removed from solution by centrifugation. The dilute HCL solution was used to further decrease the pH of the milk to 4.8. The decrease from pH 6.4 to 4.8 again resulted in an opaque-colored solution. The precipitated milk proteins from the second pH drop were also removed by sedimentation, and both protein pellets were resuspended in 1X PBS (pH 7.4). Figure 10A shows a schematic of the process described above.

Samples of the starting material, as well as soluble and insoluble fractions from each pH drop were subjected to SDS-PAGE and western analysis. The results of this experiment are shown in Figure 10B.

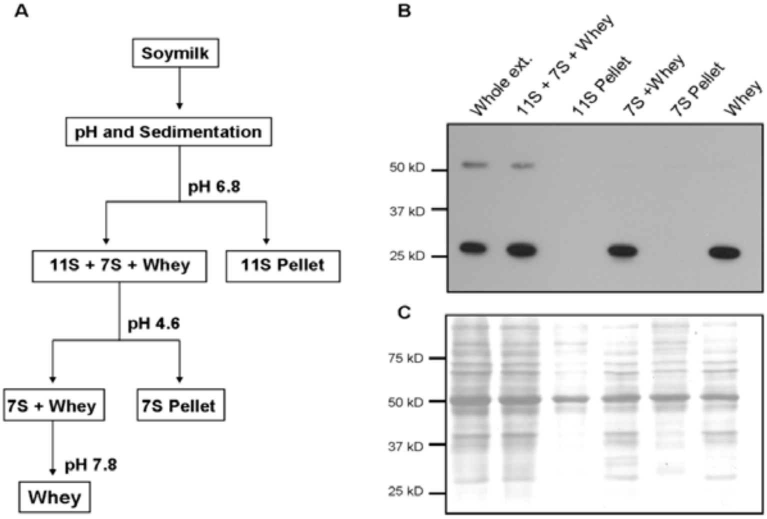


Fig. 10. Partial purification of mSEB using isoelectric fractionaton. A. Schematic of soybean protein families that precipitate at various pH. B. Western blot (top panel) and coomassie-stained membrane (bottom panel) of soymilk protein (20ug) following isoelectric fractionation.

As predicted, soy-derived mSEB remained soluble with 7S and whey proteins following initial precipitation and sedimentation of the 11S proteins. Further separation of that mixture revealed fractionation of mSEB with whey protein, while 7S proteins precipitated at pH 4.8. It should be noted that the relative amount of mSEB did not decrease during the fractionation, nor did it appear to lose immunogenicicty as detected by primary antibody in the western gel experiments. This observation was important because it confirmed the stability and immunogenicity of the end-point product.

Membranes were stained with Coomassie blue dye (Figure 10C) to visually observe the composition of proteins following fractionation. Based on the known amounts of protein loaded in each lane, as well as the relative amount of mSEB detected in the western, we were able to achieve a four-fold concentration of this model heterologous protein using isoelectric fractionation.

The above results demonstrate that partial purification of a heterologous protein can be accomplished using adjustments to pH in a soymilk formulation. This experiment took only about one hour to perform, and resulted in a four-fold purification with little detectable loss of immunogenicity. In theory, if additional concentration (with respect to volume) was needed then fractionated soymilk solutions could undergo partial lyophilization to further increase heterologous protein concentration. A five-fold reduction in volume, coupled with a four-fold increase in heterologous protein accumulation would result in a 20-fold increase in heterologous protein concentration. Such concentrations may be beneficial when attempting to address basic scientific questions with heterologous proteins that express at sub-optimal levels. Our laboratory has used the above methods to concentrate a soymilk vaccine formulation prior to efficacy testing in mice when limitations were placed on the volume of material that could be gavaged.

While the above example demonstrated fractionation of a heterologous protein with a relatively high pI, it should be noted that this simple methodology could also be used to effectively precipitate heterologous proteins with any pI value. In this manner, the heterologous protein would precipitate with host proteins sharing the same relative pI values. Our laboratory has also used isoelectric fractionation to demonstrate that a protein with a predicted pI of 5.5 could be precipitated from soymilk solutions following a drop in pH, and then reconstituted and used for downstream analyses (data not shown). While isoelectric fractionation can be applied to any host system, the physical properties of major soy protein families provide an advantage for fractionation of proteins expressed in this host.

10. Conclusion

Soybeans have the ability to assemble and accumulate many valuable proteins that can be used for the production of pharmaceuticals. This host system offers many advantages that are not present in other host systems, and shows great potential to offer safe and cost-effective protein-based products that can be used worldwide. In this chapter, we have discussed our knowledge for optimizing recombinant protein expression in soybeans. In our efforts to optimize the production of these proteins we have used codon optimization for expression in soybean. We have compared targeted expression to the whole plant and chloroplasts using constitutive promoters as well as to seeds using different seed-specific promoters. We have also analyzed the effects of a variety of plant and non-plant based signal peptides to target proteins to the secretory pathway of cells. We have also shown that protein expression levels can fluctuate until homozygosity and protein stabilization has been achieved. Protein expression levels based on the use of various combinations of promoters, leader sequences, and signal peptides resulted in protein expression that typically approached 2%-4% of total soluble protein extracted from seeds. While the optimal destination for recombinant proteins may need to be determined empirically, our results demonstrate that targeting proteins to the soybean seed through the secretory pathway resulted in the highest levels of recombinant protein accumulation.

12. References

Andrews L.B., Curtis W.R. (2005). Comparison of transient protein expression in tobacco leaves and plant suspension culture. *Biotechnol Prog., 21(3), 945-952.*

Bhattacharyya, M.K., Stermer, B.A., Dixon, R.A. (1994). Reduced variation in transgene expression from a binary vector with selectable markers at the right and left T-DNA borders. *The Plant Journal, 6(6), 957-968.*

Caligari, P.D.S., Yapabandara, Y.M.H.B., Paul, E.M., Perret, J., Roger, P., Dunwell, J.M. (1993). Field performance of derived generations of transgenic tobacco. *Tag Theoretical and Applied Genetics, 86(7), 875-879.*

Chalfie M., Tu Y., Euskirchen G., Ward W., Prasher D. (1994). Green fluorescent protein as a marker for gene expression. *Science 263, (5148), 802–805.*

Chyi, Y., Jorgensen, RA., Goldstein, D., Tanksley, S.D., Loaiza-Figueroa, F. (1986). Locations and stability of *Agrobacterium*-mediated T-DNA insertions in the *Lycopersicon* genome. *Molecular and General Genetics MGG, 204(1), 64-69.*

De Angioletti M., Lacerra G., Sabato V., Carestia C. (2004). Beta+45 G --> C: a novel silent beta-thalassaemia mutation, the first in the Kozak sequence. *Br J Haematol, 12,(2), 224–231.*

Ding, S.H., Huang, L.Y., Wang, Y.D., Sun, H.C., & Xiang, Z. H. (2006). High-level expression of basic fibroblast growth factor in transgenic soybean seeds and characterization of its biological activity. *Biotechnol Lett, 28(12), 869-875.*

Eckert H., LaVallee B.J., Schweiger B.J., Kinney A.J., Cahoon E.B., Clemente T. (2006). Co-expression of the borage Δ6 desaturase and the Arabidopsis Δ15 desaturase results in high accumulation of stearidonic acid in the seeds of transgenic soybean. *Planta, 224, 1050-1057.*

Fiedler U., Phillips J., Artsaenko O., Conrad U. (1997). Optimization of scFv antibody production in transgenic plants. *Immunotechnology, 3, 205-216.*

Franconi R., Demurtas O.C., Massa S. (2010). Plant-derived vaccines and other therapeutics produced in contained systems.*Expert Rev Vaccines, 9(8), 877-892.*

Gallie D.R., Walbot V. (1992). Identification of the motifs within the tobacco mosaic virus 5′-leader responsible for enhancing translation. *Nucleic Acids Res.,20, 4631–4638.*

Garg, R., Tolbert, M., Oakes, J.L., Clemente, T.E., Bost, K.L., & Piller, K.J. (2007). Chloroplast targeting of FanC, the major antigenic subunit of Escherichia coli K99 fimbriae, in transgenic soybean. *Plant Cell Rep, 26(7), 1011-1023.*

Gils M., Kandzia R., Marillonnet S., Klimyuk V., Gleba Y. (2005). High-yield production of authentic human growth hormone using a plant virus-based expression system. *Plant Biotechnol J, 3(6), 613-620.*

Gutierrez-Ortega A., Sandoval-Montes C., de Olivera-Flores T.J., Santos-Argumedo L., Gomez-Lim, M.A.(2005).Expression of a functional interleukin—12 from mouse in transgenic tomato plants. *Transgenic Res,14, 877-885.*

Hood E.A., Witcher D.P., Maddock S., Meyer T., Baszczynski C., Bailey M. , Flynn P., Register J., Marshall L. Bond D., et al. (1997). Commercial production of avidin from transgenic maize: characterization of transformant, production, processing, extraction and purification. *Mol. Breed., 3, 291-306.*

Horsch R.B., Fry J.E., Hoffmann N., Eichholtz D., Rogers, S.G., Fraley R.T. (1985). A simple and general method for transferring genes into plants. *Science, 227, 1229-1231.*

Hobbs, S.L.A., Kpodar, P., Delong, C.M.O. (1990). The effect of T-DNA copy number, position and methylation on reporter gene expression in tobacco transformants. *Plant Molecular Biology, 15(6), 851-864.*

Iyer, L.M., Kumpatla, M. B., Chandrasekharan, M.B., Hall, T.C. (2000). Transgene silencing in monocots. *Plant Molecular Biology, 43(2-3), 323-346.*

Jabeen R., Khan M.S., Zafar Y., Anjum T. (2010). Codon optimization of cry1Ab gene for hyper expression in plant organelles. *Mol Biol Rep., 37(2),1011-1017.*

Jefferson R.A., Kavanagh T. A., Bevan M. W.(1994). GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker, in higher plants. *EMBO J., 6, (13), 3901–3907.*

Kost, T.A. and Condreay, J.P. (1999) Recombinant baculoviruses as expression vectors for insect and mammalian cells. *Curr. Opin. Biotechnol., 10, 428–433.*

Kapusta J., Modelska A., Figlerowicz M., Pniewski T., Letellier M., Lisowa O., Yusibov V., Koprowski H., Plucienniczak A, Legocki AB. (1999). A plant-derived edible vaccine against hepatitis B virus. *FASEB, 13,13, 1796-1799.*

Karg S.R., Kallio P.T. (2009). The production of biopharmaceuticals in plant systems. *Biotechnol Adv.,27(6), 879-894..*

Karnoup A.S., Turkelson V., Kerr Anderson W.H. (2005). O-Linked glycosylation in maize-expressed human IgA1. *Glycobiology,15(10), 965-981.*

Kozak M.(1984).Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucl. *Acids Res.,12(2), 857-872.*

Lubiniecki, A.S. and Lupker, J.H. (1994) Purified protein products of rDNA technology expressed in animal cell culture. *Biologicals, 22, 161–169.*

Lusas, E.W. & Riaz M.N.(1995) Soy protein products: processing and use. J Nutr, 125(3) 573S-580S.

Lütcke H.A., Chow K.C., Mickel F.S., Moss K.A., Kern H.F., Scheele G.A. (1987). Selection of AUG initiation codons differs in plants and animals. *EMBO J., 6(1), 43-48.*

Matzke, M.A., Mette, M.F., Matzke, A.J.M. (2000). Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Molecular Biology, 43(2-3), 401-415.*

Mlynarova, L., Keizer, L.C.P., Stiekema, W.J., Nap, J.P. (1996). Approaching the lower limmits of transgene variability. *The Plant Cell, 8(9), 1589-1599.*

Moeller L., Gan Q., Wang K. (2009). A bacterial signal peptide is functional in plants and directs proteins to the secretory pathway. J Exp Bot., 60(12), 3337–3352.

Moravec, T., Schmidt, M.A., Herman, E.M., & Woodford-Thomas, T. (2007). Production of Escherichia coli heat labile toxin (LT) B subunit in soybean seed and analysis of its immunogenicity as an oral vaccine. *Vaccine, 25*(9), 1647-1657.

Mølhøj M. & Dal Degan F. (2004). Leader sequences are not signal peptides. *Nature Biotech.,* 22, 1502.

Muller, A.J., Mendel, R.R., Schiemann, J., Simoens, C., Inze, D. (1987). High meiotic stability of a foreign gene introduced into tobacco by *Agrobacterium*-mediated transformation. *Molecular and General Genetics MGG, 207(1), 171-175.*

Nakamura S. Takasaki H., Kobayashi K., Kato A. (1993). Hyperglycosylation of hen egg white lysozyme in yeast. *J. Biol. Chem., 268 (12), 706-712.*

Nielsen N..C., Dickinson C. D, Cho T. J., Thanh V. H., Scallon B. J., Fischer R. L., Sims T. L., Drews G. N., R. B. Goldberg R. B.(1989). Characterization of the Glycinin Gene Family in Soybean. *The Plant CellL, 1, (3) 313-328.*

Odell J.T.,Nagy F.,Chua N.H. (1985). Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. *Nature, 313, 810–812.*

Padidam, M.; Cao, Y. J. (2001). Elimination of transcriptional interference between tandem genes in plant cells. *BioTechniques, 31,328–334.*

Piller, K. J., Clemente, T.E., Jun, S.M., Petty, C.C., Sato, S., Pascual, D.W., et al. (2005). Expression and immunogenicity of an Escherichia coli K99 fimbriae subunit antigen in soybean. *Planta, 222*(1), 6-18.

Puonti-Kaerlas, J., Eriksson, T., Engstrom, P. (1992). Inheritance of a bacterial hygromycin phosphotransferase gene in the progeny of primary trangenic pea plants. *Tag Theoretical and Applied Genetics, 84(3-4), 443-450.*

Schillberg S., Zimmermann S., Voss A., Fisher R. (1999). Apoplastic and cytosolic expression of full size antibodies and antibody fragments in *Nicotiana tabacum*. *Transgenic Research, 8 (4),* 255-263.

Schmidt, M.A., and Herman, E.M. (2008). Proteome rebalancing in soybean seeds can be exploited to enhance foreign protein accumulation. *Plant Biotechnol J, 6, 832-842.*

Spencer, T. M., O'Brien, J. V., Start, W. G., Adams, T. R., Gordon-Kamm, W. J., Lemaux, P. G. (1992). Segregation of transgenes in maize. *Plant Molecular Biology, 18(2), 201-210.*

Streatfield S.J., Lane R.L., Brooks A.K, Barker D.K, Poage M. L., Mayor J.M., Lamphear B.J., Drees C.F. Jilka J.M., Hood E.E., Howard J.A. (2003). Corn as a production system for human and animal vaccines. *Vaccine, 21, 812-815.*

Streatfield S.J. (2007). Approaches to achieve high-level heterologous protein production in plants. *Plant Biotechnol J, 5(1), 2-15.*

Thanh V.H. & Shibasaki K. (1976). Major protein of soybean seeds: a straightforward fractionation and their characterization. *J Agric Food Chem., 24, 1117-1121.*

Thomson JA. (2006). Seeds for the future: The impact of genetically modified crops on the environment. Australia, Cornell University Press.

Thompson A.J. and Myatt S.C. (1997).Tetracycline-dependent activation of an upstream promoter reveals transcriptional interference between tandem genes within T-DNA in tomato. *Plant Molecular Biology, 34, 687–692.*

Walters, D.A., Vetsch, C.S., Potts, D.E., Lundquist, R.C. (1992). Transformation and inheritance of a hygromycin phosphotransferase gene in maize plants. *Plant Molecular Biology, 18(2), 189-200.*

Zeitlin L., Olmsted S.S., Moench T.R., Co M.S. (1998). A humanized monoclonal antibody produced in transgenic plants for immunoprotection of the vagina against genital herpes. *Nature Biotech, 11(1), 1361-1364.*

Zhong G.Y., Peterson D., Delaney D.E., Bailey M., Witcher D.R., Register III J.C., Bond D. , Lin C.P., Marshall L. , Kulisek E., Ritland D., Meyer D., Hood E.E. , Howard J.J. (1999). Commercial production of aprotinin in transgenic maize seeds. *Molecular Breeding, 5, 345-356.*

Zimmermann S., Schillberg S., Liao Y., Fisher R. (1998). Intracellular expression of TMV-specific single-chain Fv fragments leads to improved virus resistance in shape *Nicotiana tabacum*. *Mol Breeding, 4(4), 369–379.*