Chapter 2
Plant Trait Gene Expression Cassette Design

Michael Nuccio, Xi Chen, Jared Conville, Ailing Zhou and Xiaomei Liu

Introduction

Plant genetic engineering relies on promoters to develop trait genes. Many early promoters were modeled on plant pathogens, such as the cauliflower and figwort mosaic virus (CaMV, FMV; Benfey and Chua 1990; Sanger et al. 1990) and *Agrobacterium tumefaciens* (An 1986). Their activity is regarded as "constitutive" because they contained all the necessary information to produce mRNA in most plant cells. In addition, they are generally active across plant species, although quantitative performance or the amount of transgene activity produced can be variable. The basic elements required to successfully produce mRNA in plant cells are a promoter, a coding sequence, and a terminator. The promoter contains the necessary information to recruit the transcriptional machinery and initiate transcription. The coding sequence encompasses the desired trait which can take the form of a protein or RNA. The terminator provides information to end transcription and signal polyadenylation (Birch 1997). This basic structure has been in use since the inception of modern plant genetic engineering more than 25 years ago.

M. Nuccio (✉) · J. Conville · A. Zhou · X. Liu
Syngenta Biotechnology, Inc., 3054 East Cornwallis Road, Research Triangle Park, 27709 NC, Research Triangle Park, USA
e-mail: michael.nuccio@syngenta.com

J. Conville
e-mail: jared.conville@syngenta.com

A. Zhou
e-mail: ailing.zhou@syngenta.com

X. Liu
e-mail: xiaomei.liu@syngenta.com

X. Chen
Syngenta Biotechnology (China) Co. Ltd, Zhongguancun Life Science Park, No. 25, Life Science Park Road, Changping District, 102206 Beijing, China
e-mail: xi.chen@syngenta.com

K. Azhakanandam et al. (eds.), *Recent Advancements in Gene Expression and Enabling Technologies in Crop Plants,* DOI 10.1007/978-1-4939-2202-4_2

Transgene expression challenges were encountered as more scientists experimented with plant genetic engineering. These include transcriptional and posttranscriptional silencing, low or no protein accumulation, targeting transgene expression to specific cells, and enabling transgene expression under specific conditions. The transcriptional challenges were met by sourcing promoters from plants (Christensen and Quail 1996; McElroy et al. 1990) and incorporating enhancers found in certain plant pathogens (Gallie and Walbot 1992). Expression problems were addressed by introducing heterologous introns into 5′-untranslated regions (UTRs) or the trait gene coding sequence (Rose 2004), incorporating a Kozak sequence (Kozak 2002), altering codon usage (Koziel et al. 1996), introducing matrix attachment regions (Allen et al. 2000; Butaye et al. 2004), and altering terminators (Ingelbrecht et al. 1989). All of these innovations expanded the trait gene expression control toolkit, giving practitioners more flexibility (Lessard et al. 2002; Potenza et al. 2004).

The early days of plant genetic engineering included work to identify and characterize the basic plant-gene sequences required to initiate transcription (Katagiri and Chua 1992). Examples of this work include chalcone synthase, a gene in flavonoid biosynthesis pathway (Schulze-Lefert et al. 1989), the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO; Giuliano et al. 1988) in the Calvin cycle, and seed storage proteins (Jordano et al. 1989; Nunberg et al. 1994). Investigators identified regions that were responsible for environmental and hormone regulation, among other inputs. In many cases, fine-mapping work demonstrated the interaction of a specific transcriptional regulator with a specific sequence in a promoter (Gruissem 1990). This body of work defined numerous *cis*-regulatory elements, some of which have been shown to functionally enhance transcription from a given promoter (Chang and Sun 2002; Rombauts et al. 2003).

Other work focused is on the contributions of introns to trait gene expression control (Luehrsen and Walbot 1991). For example, the maize Shrunken-1 intron greatly improves trait protein production when incorporated into its corresponding promoter (Maas et al. 1991). Some plant-gene promoters require at least the first intron in order to function correctly in transgenic plants (Rose et al. 2008; Sieburth and Meyerowitz 1997). More recent work revealed a correlation between the physical properties of introns and protein production from transgenes (Korf and Rose 2009). Now, there is evidence that introns with specific properties can be used in a heterologous context to increase protein production from transgenes (Bartlett et al. 2009; Emami et al. 2013; Parra et al. 2011).

Terminators have received much less attention with respect to control of trait gene expression because the role of a gene's 3′-sequences in expression control has not been investigated extensively (Hunt 2008; Rothnie 1996). Investigations on sequences that contribute to the formation of 3′-end of mRNAs (Hunt 1994) and sequences that contribute to mRNA stability (Lidder et al. 2005) demonstrate that elements associated with plant transcriptional terminators contribute to overall gene activity. Another group demonstrated that a plant mRNA's 3′-UTR interacts with a metabolite, influencing its stability and ability to recruit ribosomes (Wachter et al. 2007). There has been an effort to map functional sequences within 3′-UTRs (Wachter et al. 2007), and the basic properties of transcription terminators have

been described (Xing et al. 2010). Much more work has been done in yeast and animal systems. For example, evidence shows that a gene's 3′-sequence functions in the formation of transcription loops that lead to the production and processing of mRNA (Moore and Proudfoot 2009).

The transcriptional terminator derived from the *Agrobacterium* nopaline synthase (NOS) gene was among the first used in plant transgenes, and remains in wide use today (Gleave 1992). Some investigators have shown that substituting a different terminator for the NOS terminator can influence trait protein production (Ingelbrecht et al. 1989). In one case, an *Arabidopsis* embryo-specific promoter was found to be nonfunctional when coupled to the NOS terminator, and functional when used with the gene's corresponding terminator (Nuccio 1997). These individual cases suggest that terminators potentially contribute far more gene expression control information than currently understood. The influence terminators exert on overall gene activity needs more attention. New methods should facilitate this work (Zhao et al. 2011).

Much of the early plant genetic engineering work revealed ways to manage individual trait gene expression problems (Koziel et al. 1996). More recent demands require multiple transgenes. Investigations over the years indicate that repeated use of a specific expression cassette, like the CaMV35S/NOS cassette, may not be an ideal or workable solution (Kebeish et al. 2007). In addition, trait gene stacking reveals a distinct shortage of reliable expression cassettes. There are two important aspects to this problem. The first is physically assembling multigenic vectors (Gibson et al. 2008). The second is identifying trait gene expression cassettes that effectively cooperate to enable the trait (Peremarti et al. 2010). Work to facilitate multigenic trait construction at the industry level is well underway (Que et al. 2010).

The trait gene assembly problem can be managed to some extent with current recombinant DNA methodology (Sambrook and Russell 2001), although efficiencies suffer as vector size extends beyond 20 kb. Large DNA molecules are sheared more easily when handled using common manual techniques. The chromatography kits routinely used to isolate DNA molecules do not work well with molecules larger than 20 kb. Furthermore, the DNA ligases used in early recombinant DNA work do not efficiently join large DNA molecules to produce the intended products. Ultimately, these early recombinant DNA methods fail with vectors larger than 50 kb. DNA recombination systems like Gateway™ are very useful in producing large DNA assemblies (Chen et al. 2006). More recently, a combination of DNA synthesis and *in vivo* homologous recombination demonstrated assembly of very large DNA molecules, including a complete microbial genome (Gibson et al. 2010).

Several prototype multigenic expression-control systems have been described for plant applications (Halpin 2005; Jiang et al. 2013; Zhu et al. 2008). Early work sought to string CaMV35S/NOS expression cassettes in tandem. This rarely worked well if the repetitive unit was viral in origin. This is not well understood. One theory suggests that sequence homology of more than 90 bp between two promoters in a transgenic plant leads to transcriptional gene silencing (Flavell 1994). This indicates that using the same promoter many times may lead to homology-based transgene silencing (Vaucheret and Fagard 2001). Another possibility is that the

CaMV35S promoter contains a recombination hotspot that can lead to unintended trait gene rearrangement (Kohli et al. 1999).

Work was more successful with plant gene-based expression cassettes (Naqvi et al. 2009). Another approach incorporated a protease signal, enabling construction of polycistronic protein coding sequence (Halpin et al. 1999). This enabled multiprotein production from a single promoter. More recent work combines unique expression cassettes together to form multigenic trait constructs, and utilizes well-characterized promoters (Fujisawa et al. 2009). Even here, some promoters are used more than once. It is easy to see that global expression profiling data might be leveraged to identify genes that share an activity profile. While promising, this area requires more investigation.

A simplified expression cassette development strategy for trait work is needed to meet today's trait expression control requirements. The focus here will be to use native plant-gene transcription units as a source for applications in maize. Global transcription profiling data simplify the identification of plant genes that possess desirable expression profiles (Wolfinger et al. 2001; Zimmermann et al. 2004). Furthermore, the ever-increasing availability of plant genome data provides the basic information required to design effective expression cassettes.

A Plant-Gene-Based Expression Cassette Design Strategy

The challenge is that little is known about how specific gene regulatory sequences work, or what sequences are necessary and sufficient to recapitulate a gene's expression profile. A method is necessary to leverage poorly characterized plant genes for expression of cassette development. An approach to address this issue (Nuccio et al. 2012) is described below. We elected to simplify plant-gene annotation into five basic units. They include the promoter which is 1.0–2.0 kb of sequence upstream of the transcription start site, the 5′-UTR or the sequence from the transcription start site to the translation start codon, the coding sequence which comprises most of the exons and introns, the 3′-UTR or the sequence from the translation stop codon to the end of the transcript, and 3′-downstream sequence which extends up to 1.0 kb past the translation stop codon. In isolation, these components have been shown to contribute trait gene expression control. The hypothesis is that these components possess the majority of a given plant gene's expression-control information, and could be combined to form robust and reliable expression cassettes without any direct knowledge of the exact sequences that regulate the donor gene's activity. The effectiveness of this approach is illustrated in the subsequent sections of this chapter.

Accurate sequence of both the gDNA and cDNA of a donor plant gene is required to be useful for expression cassette development. This information is widely available for many plant species. Gene annotation may also be available in public databases or genome browsers (Duvick et al. 2008; Karolchik et al. 2003; Liang et al. 2008; Ouyang et al. 2007). If not, the sequence data can be generated from donor plant tissue. The largest open reading frame in an mRNA sequence typically

defines the gene's protein-coding sequence, as identified by the translation start and stop codons. The protein sequence can support the accuracy of the sequence data. If necessary, techniques such as 5′- and 3′-rapid amplification of cDNA ends (RACE) can identify the mRNA's termini, which represent the transcription start and stop sites (Das et al. 2001). The gDNA and cDNA sequences can be aligned in several software tools to define the gene's basic architecture as outlined above (Wheelan et al. 2001). This information is sufficient to design expression cassettes based on most plant genes.

The objective is to develop expression cassettes that comprise the components listed above and are simple to use. In order to do this, a two-component regulatory system consisting of a gene's 5′- and 3′-regulatory sequence is defined. The 5′-regulatory sequence contains the promoter exon 1, intron 1, and part of exon 2. The 3′-regulatory sequence contains sequence downstream of the translation stop codon and 3′-nontranscribed sequence. From this point forward, these will be referred to as the promoter and terminator, respectively. This approach casts a wide net to capture most, if not all, regulatory sequence necessary to recapitulate a gene's expression profile.

Natural gene sequences present several challenges that limit their direct use in expression cassettes. First, they likely contain restriction endonuclease sites that prevent manipulation by standard recombinant DNA methodology. To address this, a standard restriction endonuclease profile for each expression cassette is defined. The promoter is flanked by SanDI on the 5′-end and NcoI on the 3′-end. The terminator is flanked by SacI on the 5′-end and RsrII on the 3′-end. Expression cassettes are designed to be assembled in an intermediate vector, and this configuration enables the cassette's mobilization into other vectors, such as binary vectors, as SanDI/RsrII fragments which can be ligated into either a SanDI or RsrII site. SanDI (aka. KflI) recognizes GG^GWCCC and RsrII recognizes CG^GWCCG. To produce compatible ends, W needs to be either A in both sites, or T in both sites. The SanDI site remains intact when a SanDI/RsrII fragment is ligated to a SanDI site. This enables a subsequent SanDI/RsrII flanked expression cassette to be ligated adjacent to the previous insert.

Furthermore, this configuration provides NcoI/SacI sites to insert a gene of interest. These sites are added to the plant-gene sequence. Internal restriction sites that interfere with the standard sites are eliminated by a single-point mutation in each site. This is easily done by substituting G for C, and A for T, and vice versa. Point mutations can be introduced using polymerase chain reaction (PCR) methods such as Stratagene's QuikChange® Site-Directed Mutagenesis Kit. Point mutations may disrupt the functionality of the target gene regulatory elements, but this is the least-invasive approach available at the moment. The terminator generally does not require more than this for incorporation into expression cassettes.

Fig. 2.1 outlines several additional changes to the promoter sequence necessary to make it useful in expression cassettes. The engineered NcoI site is the 5′-CDS ligation site and it provides the translation start site. This is located after the first 10–15 nucleotides of exon 2. The reason is that sequence proximal to the intron/exon junction may be required for proper intron excision. Also a maize-preferred

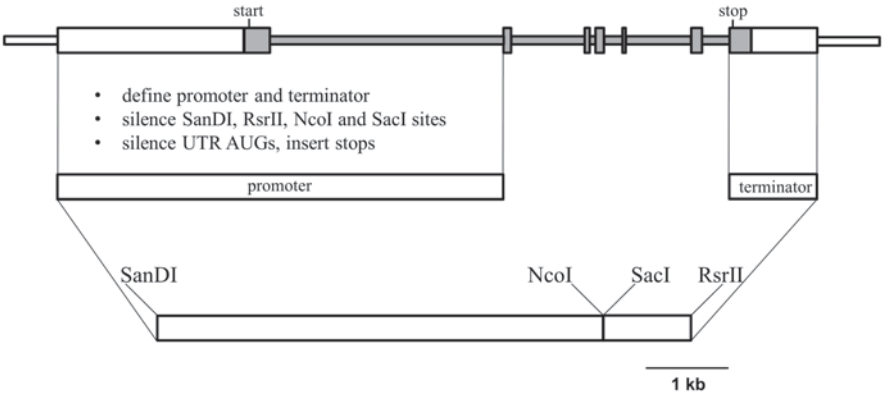

Fig. 2.1 The expression cassette design process. The gDNA at the top is minimally modified to define a promoter and a terminator. The design process captures as much regulatory sequence as possible and supports a standardized recombinant DNA framework. The primary considerations include location of the expression cassette translation initiation codon at the *NcoI* site and a restriction endonuclease arrangement to support industrial applications. In an assembly (or intermediate) vector, each expression cassette is flanked by *SanDI/RsrII* sites that can be mobilized into either a *SanDI* or *RsrII* site. This enables expression cassettes to be stacked into a binary vector. The coding sequence is directionally inserted into the *NcoI/SacI* sites. In the gDNA, the open boxes represent nontranscribed sequence, *gray* is transcript, *large gray* boxes are exons, *narrow gray* boxes are introns. The translation start and stop codons are also indicated

Kozak site, which is defined as AAAACCATGG is typically incorporated. In most cases, the promoter will possess codons that need to be eliminated. Exon 1 and exon 2 are examined for methionine codons in the three possible sequence contexts. The ATG codons are altered by a single point mutation that does not introduce an unwanted restriction endonuclease site. The same approach is used to ensure that at least one translation stop codon is in each frame upstream of the NcoI site. These manipulations ensure that translation initiates at the ATG in the NcoI site.

A final consideration for expression cassettes is proximity of donor genes to other genes. This is relatively straight forward for donor genes derived from plants with well-annotated genomes like *Arabidopsis,* rice, or sorghum. The purpose is to determine if the nontranscribed promoter and terminator sequence overlaps with an adjacent gene. This is common in compact genomes like that of *Arabidopsis*. If genome information is not available, simply BLAST (Altschul et al. 1990) the nontranscribed sequences against transcript databases. Investigate all high-quality hits. If overlap is found, it should be eliminated.

This strategy is designed to develop trait gene expression control technology with minimal effort. It is made possible by public and private investment in plant genome information. It enables expression cassette development in the absence of specific details regarding elements that control plant-gene expression. Syngenta applied this approach to many trait gene expression control problems over the years and found it to be reliable and robust (Lee et al. 2013; Nuccio 2013; Nuccio and

Richmond 2013; Nuccio et al. 2012). Several examples from work in maize are outlined below to illustrate the utility of this approach.

Ear-Specific Expression Cassettes Based on Rice MADS Genes

Trait development may require trait gene expression to be limited to specific cells at specific times. One challenge was to target trait gene expression to pedicel tissue in early development. The MADS (MCM1, AGAMOUS, DEFICIENS, SRF) transcription factor gene family was targeted because the plant-gene family members are often active in early flower development (Alvarez-Buylla et al. 2000; De Bodt et al. 2003). There is evidence that an orchid MADS gene, DoMADS3 (Yu and Goh 2000), was active in pedicels. Corresponding gDNA sequence was not available, so the DoMADS3 protein sequence was used to screen rice genome data (Goff et al. 2002) for candidates. This genomics approach identified candidate rice genes as the basis for expression cassettes that express transgenes in young developing ears. The rationale was that the rice DoMADS3 ortholog's expression profile would reflect that of DoMADS3. This led to the identification and annotation of 34 rice MADS genes, which are rank-ordered by similarity to DoMADS3 in Table 2.1.

OsMADS5, -6, -7, -8, -13, and -14 were selected for expression cassette development based on their rank in Table 2.1. The uncertainty with respect to the location of critical regulatory elements led to the development of the cassette-design strategy outlined in the previous section. Fig. 2.2 illustrates the structure of each gene. Thick lines below each annotation depict sequence incorporated into expression cassettes. Each expression cassette was fused to the β-glucuronidase (GUS) reporter gene for evaluation in transgenic maize.

Each expression cassette was characterized by histochemical localization of GUS activity in transgenic maize. Fig. 2.3 illustrates that each expression cassette produces a unique profile in developing ears. The OsMADS5 cassette is active in vasculature traversing the spikelet and the cob. OsMADS8 is active in the lemma and the nonvascular cob cells. OsMADS6 is active in the cob and spikelet vasculature and, to some extent, the glume. Fig. 2.3c shows that OsMADS6 is also active in ear node, shank vasculature, the inner bundle of the internode extending above the ear node, and basal shank vasculature. OsMADS13 is active in the central spikelet tissues including the vasculature and most likely the carpels. It is slightly active in cob vasculature. Fig. 2.3d shows that OsMADS13 is also slightly active in the ear node and discrete files within the shank. OsMADS7 is active in the embryo sac. The difference in histochemical deposition between Fig. 2.3g and 2.3h delineates the contribution of OsMADS7's first intron to its expression pattern. Inclusion of the intron (Fig. 2.3h) limits activity to the embryo sac. OsMADS14 (specifically OsMADS14 l) is active in the cob and cob vasculature as well as the spikelet vasculature. It is also active in the embryo sac. The lack of histochemical deposition in

Table 2.1 Identification of OsMADS genes using the DoMADS3 protein sequence

Locus[a]	Representative cDNA	SwissProt ID	Study name	Protein size (AAs)	Comparison to DoMADS3 (percentage)				
					Whole protein			*MADS domain*	
					Identity	Similarity	Gaps	Identity	Similarity
LOC_Os08g41950.2	U78891	MADS7_ORYSJS	OsMADS7	236	62	75	4	94	99
LOC_Os03g11614.1	AF204063	MADS1_ORYSJS		257	60	72	3	91	98
LOC_Os09g32948.1	U78892	MADS8_ORYSJS	OsMADS8	248	60	73	6	94	99
LOC_Os03g54170.1	AB003324	MAD34_ORYSJS	OsMADS14	239	59	74	3	80	95
LOC_Os02g45770.1	U78782	MADS6_ORYSJS	OsMADS6	250	58	69	4	91	99
LOC_Os06g06750.1	AF141967	MADS5_ORYSJS.	OsMADS5	225	57	72	2	92	97
LOC_Os04g49150.1	AF095646	MAD17_ORYSJS		254	55	67	6	94	99
LOC_Os10g39130.1	AF141965	MAD56_ORYSJS		233	51	67	0	84	91
LOC_Os03g54160.2	AF139664	MAD14_ORYSJS		246	50	66	1	78	95
LOC_Os07g01820.3	AF345911	MAD15_ORYSJS		267	50	68	5	80	95
LOC_Os07g41370.1	AF139665	MAD18_ORYSJS		249	48	66	0	80	95
LOC_Os03g03100.1	AB003328	MAD50_ORYSJS		230	48	64	0	77	91

Table 2.1 (continued)

Locus[a]	Representative cDNA	SwissProt ID	Study name	Protein size (AAs)	Comparison to DoMADS3 (percentage)				
LOC_Os12g10540.4	AF151693	MAD13_ORYSJS	OsMADS13	270	46	67	1	84	94
LOC_Os08g33488.1	AY177694	MAD23_ORYSJS		159	46	66	0	66	87
LOC_Os01g10504.3	L37528	MADS3_ORYSJS		276	45	68	1	84	94
LOC_Os12g31748.1	AY250075	MAD20_ORYSJS		233	42	67	5	64	88
LOC_Os02g52340.1	AB003322	MAD22_ORYSJS		228	42	58	0	68	78
LOC_Os06g45650.1	AY174093	MAD30_ORYSJS		221	42	63	0	63	87
LOC_Os02g07430.1	AY177697	MAD29_ORYSJS		260	41	62	2	66	87
LOC_Os04g52410.1	AY177698	MAD31_ORYSJS		178	41	61	3	68	87
LOC_Os12g10520.1	AY177700	MAD33_ORYSJS		202	41	61	0	66	87
LOC_Os06g49840.2	AF077760	MAD16_ORYSJS		224	40	59	2	64	80
LOC_Os01g66030.1	AF095645	MADS2_ORYSJS		209	40	61	5	64	90
LOC_Os01g69850.1	AF141964	NP_001045235.1		164	40	60	11	66	86
LOC_Os06g49840.2	AF424549	MAD16_ORYSJS		224	39	59	2	63	87

Table 2.1 (continued)

Locus[a]	Representative cDNA	SwissProt ID	Study name	Protein size (AAs)	Comparison to DoMADS3 (percentage)				
LOC_Os01g66290.2	AY177693	MAD21_ORYSJS		265	39	60	7	80	95
LOC_Os04g23910.1	AY177695	MAD25_ORYSJS		227	39	58	0	66	87
LOC_Os08g02070.1	AY115556	MAD26_ORYSJS		222	39	61	1	60	85
LOC_Os02g36924.1	AY177696	MAD27_ORYSJS		240	38	62	4	61	87
LOC_Os02g49840.1	AY224482	MAD57_ORYSJS		241	38	59	5	70	89
LOC_Os08g41960.1	AY177701	BAD11644.1		203	37	56	5	72	90
LOC_Os01g52680.1	AY177699	MAD32_ORYSJS		196	37	59	3	63	78
LOC_Os05g34940.2	L37527	MADS4_ORYSJS		210	37	60	5	63	85
LOC_Os03g08754.2	AJ293816	MAD47_ORYSJS		237	35	52	8	65	79

[a] Each gene is identified by its standard locus, a representative cDNA, and its SwissProt ID. The Study name is an internal designation. The rice MADS protein sequence was compared along the entire protein or just at the MADS domain. The order is by similarity to DoMADS3

http://www.springer.com/978-1-4939-2201-7

Recent Advancements in Gene Expression and Enabling Technologies in Crop Plants

Azhakanandam, K.; Silverstone, A.; Daniell, H.; Davey, M.R. (Eds.)

2015, XXIV, 455 p. 55 illus., 39 illus. in color., Hardcover
ISBN: 978-1-4939-2201-7