1

Expression of Non-Native Genes in a Surrogate Host Organism

Dan Close, Tingting Xu, Abby Smartt,
Sarah Price, Steven Ripp and Gary Sayler
Center for Environmental Biotechnology, The University of Tennessee, Knoxville
USA

1. Introduction

Genetic engineering can be utilized to improve the function of various metabolic and functional processes within an organism of interest. However, it is often the case that one wishes to endow a specific host organism with additional functionality and/or new phenotypic characteristics. Under these circumstances, the principles of genetic engineering can be utilized to express non-native genes within the host organism, leading to the expression of previously unavailable protein products. While this process has been extremely valuable for the development of basic scientific research and biotechnology over the past 50 years, it has become clear during this time that there are a multitude of factors that must be considered to properly express exogenous genetic constructs.

The major factors to be considered are primarily due to the differences in how disparate organisms have evolved to replicate, repair, and express their native genetic constructs with a high level of efficiency. As a result, the proper expression of exogenous genes in a surrogate host must be considered in light of the ability of the replication and expression machinery to recognize and interact with the gene of interest. In this chapter, primary attention will be given to the differences in gene expression machinery and strategies between prokaryotic and eukaryotic organisms. Factors such as the presence or absence of exons, the functionality of polycistronic expression systems, and differences in ribosomal interaction with the gene sequence will be considered to explain how these discrepancies can be overcome when expressing a prokaryotic gene in a eukaryotic organism, or vice versa.

There are, of course, additional concerns that are applicable regardless of how closely related the surrogate host is to the native organism. To properly prepare investigators for the expression of genes in a wide variety of non-native organisms, concerns such as differences in the codon usage bias of the surrogate versus the native host, as well as how discrepancies in the overall GC content of each organism can affect the efficiency of gene expression and long term maintenance of the construct will be considered in light of the mechanisms employed by the host to recognize and remove foreign DNA. This will provide a basic understanding of the biochemical mechanisms responsible for genetic replication and expression, and how they can be utilized for expression of non-native constructs.

In addition, the presence, location, and function of the major regulatory signals controlling gene expression will be detailed, with an eye towards how they must be modified prior to exogenous expression. Specifically, this section will focus on the presence, location, and composition of common promoter elements, the function and location of the Kozak sequence, and the role of restriction and other regulatory sites as they relate to expression across broad host categories. Considerations relating to the potential phenotypic effects of exogenous gene expression will also be considered, especially in light of the potential for interaction with host metabolism or regulation of possible aggregation of the protein product within the surrogate host. This will provide readers with a basic understanding of how common sequences can be employed to either enhance or temper the production of a gene of interest within a surrogate host to provide efficient expression.

Finally, to highlight how these processes must be employed in concert to express non-native genes in a surrogate host organism, the expression of the full bacterial luciferase gene cassette in a human kidney cell host will be presented as a case study. This example represents a unique case whereby multiple, simultaneous considerations were applied to express a series of six genes originally believed to be functional only in prokaryotic organisms in a eukaryotic surrogate. The final expression of the full bacterial luciferase gene cassette has been the result of greater than 20 years of research by various groups, and nicely demonstrates how each of the major topic areas considered in this chapter were required to successfully produce autonomous bioluminescence from a widely disparate surrogate host. It will summarize the considerations that have been introduced, and present the reader with a clear overview of how these principles can be applied under laboratory-relevant conditions to achieve a specific goal.

2. Mechanisms of gene expression

Before exogenously expressing a gene in a foreign host organism, it is important to understand the basics behind how genes are expressed and maintained. Through this understanding of innate genetic function, it is possible to better understand the modifications that serve to enhance expression of non-native genes. Fortuitously, from a basic standpoint, all genes are subject to the same basic processes whether they are prokaryotic or eukaryotic in origin: replication, transcription, and translation. The primary differences that separate eukaryotic and prokaryotic gene expression are due to the associated proteins that are involved in each of these processes. In the end however, the objective is the same, to transcribe DNA to messenger RNA (mRNA), translate that mRNA to protein, and to have that protein carry out a function. This succession of events has

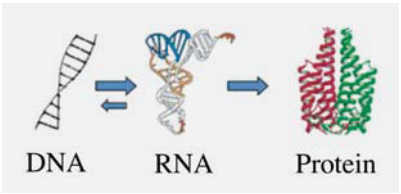

Fig. 1. The central dogma of biology shown in schematic form. DNA is transcribed to RNA and the RNA is then translated into protein. This process is the fundamental platform of our understanding of life. Adapted from (Schreiber, 2005)

become known as the central dogma of biology (Fig. 1). By understanding the differences in the genetic machinery that are employed by eukaryotes and prokaryotes, one can achieve a better understanding of why certain modifications must be made when expressing a prokaryotic gene in a eukaryotic host, and vise versa.

2.1 Replication

The end goal of the replication process is the same for all organisms, whether eukaryotic or prokaryotic: reproducing genetic information to pass on to the next generation. Replication is an especially important stage for the gene expression process not only because it provides a means for passing on genetic information, but also because any errors that occur during this period alter the genetic code and subsequently pass that alteration to future generations. The major differences in replication between prokaryotes and eukaryotes are due to the location where replication occurs and the layout of the genome itself. In prokaryotic organisms, the DNA is typically stored as a circular chromosome, located in the uncompartmentalized cytoplasm of the cell. However, in eukaryotic organisms, the DNA is packaged into linear chromosomes and stored in the nucleus of the cell. The replication of DNA, however, occurs in a similar process for both prokaryotes and eukaryotes. An origin of replication is defined where the binding of DNA helicase allows the DNA to unwind, exposing both strands of DNA and allowing them to serve as templates for replication (Keck & Berger, 2000; So & Downey, 1992). Once unwound, an RNA primer is added to the 5′ end of the DNA, and the DNA polymerase enzyme begins adding complementary nucleotides in the 5′ to 3′ direction. As DNA has an antiparallel conformation, a leading strand and lagging strand are both formed when it is unwound. The leading strand allows replication to occur continuously and therefore needs only one primer, however, the lagging strand is exposed in the 3′ to 5′ direction and forces replication to occur discontinuously. The lagging strand therefore requires multiple primers that allow the polymerase to make numerous short DNA fragments, called Okazaki fragments, which are later formed into a continuous strand (Falaschi, 2000; So & Downey, 1992). As described previously, prokaryotic DNA is housed on a circular chromosome, allowing for bidirectional replication and termination when the two replication forks meet at a termination sequence (Keck & Berger, 2000). However, because eukaryotes have linear chromosomes, termination is achieved by reaching the end of the chromosome where a telomerase enzyme then elongates the 3′ end of the chromosome so that the template DNA can complete the replication process (Zvereva et al., 2010).

2.2 Transcription

2.2.1 Transcription initiation

Transcription is the process of creating an mRNA message from a DNA template, and proceeds in three basic steps for both eukaryotic and prokaryotic organisms: initiation, elongation, and termination. One important difference is that while prokaryotes have only a single coding region for genetic information, eukaryotes have both coding and non-coding regions called exons and introns, respectively. The exons carry the genetic information that must be transcribed and translated, whereas introns break up sequences of exons with non-coding genetic sequences (Watson et al., 2008). The initiation step begins with the binding of an RNA polymerase enzyme to a specific DNA sequence that encodes the gene or genes

being expressed. This stage varies slightly between prokaryotic and eukaryotic organisms, with prokaryotes having only one RNA polymerase, whereas eukaryotes have three RNA polymerases. The prokaryotic RNA polymerase uses a specific feature called a sigma (σ) factor to recognize an upstream start site called a promoter. This region is composed of, at minimum, two DNA sequences located -35 and -10 base pairs (bp), upstream from where transcription will begin (Murakami & Darst, 2003). In addition, another DNA element called an UP-element is sometimes located further upstream within the promoter, allowing a stronger bond between the DNA template and the RNA polymerase upon binding. Immediately following the binding of the RNA polymerase, the DNA undergoes a conformational change whereby it unwinds to expose the single template strand required for the transcription process to proceed to the elongation step. This process of DNA separation generally occurs between the -11 and +3 bp positions relative to the transcription start site. Although the basic process of transcription initiation is similar in eukaryotes, different enzymes are utilized to carry out the steps described above. Unlike prokaryotes, eukaryotic organisms have three RNA polymerase enzymes called Pol I, Pol II and Pol III. Of these three enzymes, Pol II is the most predominant during routine transcription. And while prokaryotes have only the single initiation factor, the σ factor, Pol II works in conjunction with multiple general transcription factors (GTFs). Regardless of these differences, the polymerase binding process is the same, with initiation factors recognizing specific points on the promoter and allowing Pol II to bind (Ebright, 2000). In eukaryotes, the most common recognition sites are the TRIIB site, the TATA box, the initiator, or downstream promoter elements (Boeger et al., 2005). Once bound to the DNA, Pol II and the GTFs allow the DNA to unwind, preparing the way for the elongation step and the beginning of mRNA message assembly synthesis.

2.2.2 Elongation during transcription

As the elongation step begins, a conformational change allows the RNA polymerase to release from the promoter and it begins building an mRNA message as it scans along the template sequence. In prokaryotes, as the DNA template enters into the polymerase-promoter complex, it is paired with a complementary messenger sequence, producing a small transcript composed of linked mRNA nucleotides. As this process repeats, the newly formed mRNA nucleotide cannot be contained within the polymerase and must exit through a designated exit channel. This causes the σ factor to dissociate from the polymerase and likewise, the polymerase to dissociate from the template, allowing for continued elongation of the nascent mRNA message. As the mRNA is lengthened by the polymerase moving along the DNA, adding one mRNA nucleotide at a time, the DNA winds and unwinds to keep the transcription bubble that forms on the DNA template a constant size. This process is slightly different in eukaryotes, where escaping the promoter requires two steps to disconnect the GTFs from the polymerase and the polymerase from the promoter. The first step is an input of energy derived from the hydrolysis of ATP. Without the free energy released from ATP hydrolysis, an arrest period would occur that could terminate the elongation phase and thus, stop transcription altogether (Dvir et al., 1996, 2001). The second required step is the phosphorylation of Pol II. As phosphates are added to the polymerase tail, it sheds the associated GTFs and dissociates from the promoter region (Boeger et al., 2005). Once the polymerase is free of the GTFs, elongation factors are able to bind and stimulate the addition of nucleotides to the growing mRNA message.

2.2.3 Termination of transcription

After the complete mRNA has been synthesized, transcription ends in the termination step. As suggested by the name, the purpose of the termination step is to stop the production of mRNA after the template gene has been transcribed. Prokaryotes have two different termination methods, Rho-dependent and Rho-independent. Rho binding sequences are DNA sequences that signal the end of elongation and allow the polymerase to dissociate from the DNA. The Rho protein is made up of six identical subunits that have a high affinity for C-rich RNA sequences. It becomes active in transcription termination once the ribosome has slowed translation to a point where it can bind to the RNA between the RNA polymerase and the ribosome (Richardson, 2003). The presence of a Rho binding region allows the corresponding Rho protein to bind to the RNA, after it has exited the polymerase. The intrinsic ATPase activity of the Rho protein then terminates elongation, stopping the production of RNA (Richardson, 2003). Rho-independent terminators do not require binding of the Rho protein to initiate termination of RNA production. Instead, the DNA template sequence encodes an inverted repeat and a series of AT base pairs that, when transcribed to RNA, form a hairpin that is followed by a series of AU base pairs. The formation of this secondary structure causes termination of RNA production and releases the nascent mRNA message from the polymerase (Abe & Aiba, 1996). In eukaryotes, this termination process is again different from that of prokaryotes because there are three RNA processing events that lead to termination: capping, splicing, and polyadenylation. As the mRNA message exits the polymerase, capping occurs through the addition of a methylated guanine to the 5′ end of the nascent mRNA (Wahle, 1995). Next, splicing occurs where the non-coding regions of the mRNA are removed, and finally, the 3′ end of the mRNA is polyadenylated, allowing it to dissociate from polymerase and end transcription. The major differences in the transcription process between prokaryotes and eukaryotes are summarized in Table 1.

Prokaryotes	Eukaryotes
Occurs in cytoplasm	Occurs in nucleus
Single polymerase	Pol I, Pol II, and Pol III
-10, -35, and UP recognition elements	TATA box and TRIIB recognition elements
Single coding region	Multiple coding regions: exons and introns
Rho dependent and independent termination	RNA processing 5′ capping, splicing, and 3′ polyadenylation

Table 1. Comparison of the transcriptional process in prokaryotes and eukaryotes

2.3 Translation

After transcription has been successfully completed, the mRNA is ready to be translated; a process that takes the mRNA message and uses it to produce a string of amino acids, known as a protein. Just as with the transcriptional process, there are subtle, but important, differences in how this is performed in prokaryotes and eukaryotes. In eukaryotes, whereas the transcriptional process takes place in the nucleus, translation takes place in the

cytoplasm. This means that the previously produced mRNA must move across the nuclear membrane to the cytoplasm before translation can occur. Since the transcriptional process in prokaryotes occurs in the uncompartmentalized cytoplasm, this is an unnecessary step and translation can occur as soon as the mRNA exits the polymerase during transcription. Regardless of if this process occurs in a prokaryote or eukaryote, there are four major components involved: mRNA, transfer RNA (tRNA), aminoacyl-tRNA synthetases, and ribosomes. The mRNA component is composed of codons, three nucleotide long elements, which are joined together end to end to form open reading frames (ORFs). While the genes of eukaryotes usually only have one ORF per mRNA sequence, it is not uncommon for prokaryotes to contain two or more ORFs per mRNA sequence (Watson et al., 2008). These multi-ORF mRNA sequences are referred to as polycistronic mRNAs and can encode multiple proteins from a single sequence of mRNA. In order for the amino acids to recognize and bind to the mRNA template, tRNA is used as a mediator. tRNAs are complementary to specific codons via their anti-codons and, upon recognition of their specified codon, incorporate the corresponding appropriate amino acid for that codon (Kolitz & Lorsch, 2010). Once the corresponding amino acid is bound to the tRNA, the complex is referred to as an aminoacyl-tRNA synthetase, which then binds to the complement mRNA to allow the appropriate amino acid to be added to the peptide chain. The final component of the translational process, the ribosome, is the enzyme responsible for catalyzing the pairing of mRNA and tRNA, leading to the formation of the polypeptide chain. Ribosomes are composed of two individual subunits, the small and large subunits, and contain three binding sites, the A site, the P site and the E site (Ramakrishnan, 2002). These three binding sites work together to allow protein synthesis. Similar to the transcriptional process, these components work together to perform the initiation, elongation, and termination phases of translation.

2.3.1 Initiation of translation

The translational initiation stage for prokaryotes and eukaryotes involves similar steps, but each performs these steps using different enzymes. For prokaryotes, the initiation step involves the recruitment of the ribosome to the mRNA through a ribosomal binding site that is located just upstream of the start codon on the previously synthesized mRNA. This process can occur as soon as the nascent mRNA has exited the polymerase, with three translation initiation factors (IF1, IF2, IF3) binding to the A, E and P sites of the ribosome and directing the placement of the initiator tRNA to the start codon of mRNA (Ramakrishnan, 2002). Following binding, the initiation factor bound to the E site releases, allowing the large ribosomal subunit to unite with the small subunit, creating a 70S initiation complex. This binding causes the hydrolysis of GTP and subsequent release of all additional initiation factors. Following disassociation of the initiation factors, the ribosome/mRNA complex is then ready to enter the elongation phase.

Due to the intrinsic compartmentalization in eukaryotic organisms, translation is a completely separate event from that of transcription because the nuclear membrane prevents the mRNA from interacting with the ribosome until it is released into the cytoplasm. However, once in the cytoplasm, the 5′ methylated guanine cap attached to the eukaryotic mRNA binds to the ribosome and the process begins. The eukaryotic ribosome is similar to its prokaryotic counterpart in that it too has A, P and E binding sites and utilizes initiation factors to achieve correct attachment of associated tRNA (Figure 2). However,

unlike the prokaryotic ribosome, the small subunit of the eukaryotic ribosome must bind to the initiator tRNA before coming into contact with mRNA (Watson et al., 2008). After the tRNA is bound, the ribosome then recognizes the mRNA template and begins scanning for an AUG start codon. Once identified, the initiator tRNA binds to the mRNA through hydrolysis of GTP, causing the release of the first set of initiation factors and introduction of a second set (Acker et al., 2009). This allows the large subunit to bind, initiating another GTP hydrolysis event that dissociates the remaining initiation factors and creates an 80S initiation complex. After the complete ribosome initiation complex is formed the ribosome/mRNA complex is ready to enter the elongation phase of translation.

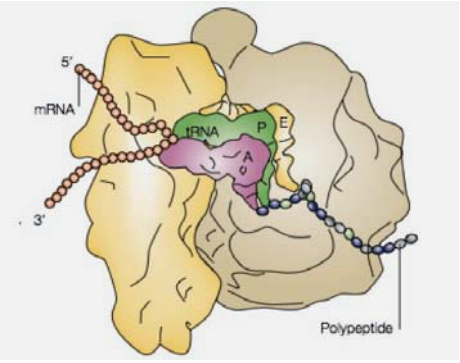

Fig. 2. The ribosome is responsible for translating mRNA into protein. Used with permission from (Lafontaine & Tollervey, 2001)

2.3.2 Elongation during translation

Elongation is where the resultant protein encoded by a specific gene first begins to take form. During elongation, each tRNA codon associates with the appropriate amino acid through a 3´ ester bond. Once the amino acid is attached, the aminoacyl-tRNA containing that amino acid binds to the A site of the ribosome. The ribosome then forms a peptide bond between the amino acid of the incoming tRNA and the peptide chain attached to the peptidyl-tRNA in the P site. Binding of the amino acid to the peptide chain causes the aminoacyl-tRNA to become a peptidyl-tRNA and forces translocation of this tRNA from the A site to the P site. This transfer then forces the peptidyl-tRNA that was previously present at the P site to exit through the E site, forming a growing chain of polypeptides that will form the final protein originally encoded by the gene being expressed. This process is carried out with the help of elongation factors. In prokaryotes there are three elongation factors (EF-Tu, EF-G, and EF-T), whereas eukaryotes utilize only two elongation factors (eEF-1 and eEF-2) (Lavergne et al., 1992; Nilsson & Nissen, 2005; Oldfield & Proud, 1993). The prokaryotic elongation factor EF-Tu and eukaryotic elongation factor eEF-1 work in a similar fashion to bind to aminoacyl-tRNAs and escort them to the A site of the ribosome (Nilsson & Nissen, 2005; Oldfield & Proud, 1993). Once the aminoacyl-tRNA is in the A site, the peptide chain from the peptidyl-tRNA attaches to the amino acid on the aminoacyl-tRNA, and this complex is ready to be translocated. Translocation involves either the EF-G factor in prokaryotic systems or the eEF-2 factor in eukaryotic systems. Both of these factors are able to associate with the peptidyl-tRNA at the P site once the peptide chain has been

transferred to the aminoacyl-tRNA at the A site, causing the hydrolysis of GTP that allows for the now peptidyl-tRNA of the A site to translocate to the P site and the peptidyl-tRNA that was in the P site to exit through the E site (Nilsson & Nissen, 2005; Riis et al., 1990; Watson et al., 2008). The final elongation factor, EF-T, found in prokaryotes and having no eukaryotic homologue, is responsible for the removal of EF-Tu and EF-G from the ribosome so that the A site is again able to bind to a new aminoacyl-tRNA and continue the elongation process (Nilsson & Nissen, 2005). This cycle of amino acid addition continues until all mRNA codons have been translated to protein.

2.3.3 Termination of translation

After successful completion of the protein synthesis process, the elongation phase must be terminated, effectively ending the growth of the polypeptide chain and marking the formation of a complete protein product. The elongation of the polypeptide product will continue until a stop codon is read from the mRNA template. In both prokaryotes and eukaryotes, there are three stop codons that can be employed to stop translation: UAG, UGA, or UAA. Once a stop codon has been recognized in the A site of the ribosome, a set of release factors (RFs) are called into action to allow the synthesized protein to be released. In prokaryotes there are two Class I release factors, RF1 and RF2, that recognize the UAG and UGA stop codons respectively and the UAA stop codon universally, and one Class II release factor, RF3, that allows the Class I release factors to dissociate from the ribosome after the protein has detached (Moreira et al., 2002). In contrast, eukaryotes have only one Class I release factor, eRF1, which recognizes all three stop codons and one Class II release factor eRF3 for dissociation (Moreira et al., 2002). Regardless of which release factor is used, when the stop codon is recognized, hydrolysis of the peptide chain begins and the newly synthesized protein and all termination elements are released from the ribosome. A summary of the host protein machinery active during translation is presented in Table 2.

	Prokaryotes	Eukaryotes	Function
Initiation	IF-1	eIF-1	Blocks the A site from initiation t-RNA
	IF-2	eIF-2	Binds to initiator t-RNA
	IF-3	eIF-3	Blocks the E site
	N/A	eIF-4	Ribosomal recognition of mRNA
	N/A	eIF-5	Blocks the E site
Elongation	EF-Tu	eEF-1	Binds aminoacyl-tRNA to the A site
	EF-G	eEF-2	Translocation
	EF-T	N/A	Releases elongation factors
Termination	RF-1	eRF-1	Recognizes the UAA and UAG stop codons
	RF-2		Recognizes the UAA and UGA stop codons
	RF-3	eRF-2	Releases all translation factors

Table 2. Host proteins active during translation

3. Considerations for the expression of exogenous DNA

Although nucleic acids serve as the universal genetic material and the central dogma applies to all organisms, exogenous expression of foreign genes is not as straightforward as delivering the target sequence into host cells and waiting for it to be expressed. This is because the gene expression machinery in certain species has evolved in such a way as to manipulate its own genetic material more efficiently than genomic material from other species, a fact that is especially true when the exogenous genetic material is from a very distantly related species. Any discrepancies, such as the genomic characteristics of GC content and codon usage patterns between the native and surrogate hosts will play an important role in the efficiency of exogenous gene expression. In addition, some organisms have also evolved to recognize and remove or silence foreign genetic sequences in order to protect themselves from the deleterious effects of foreign DNA expression. It is only through mimicking, circumventing, or deactivating these mechanisms that it becomes possible to efficiently express a foreign gene in a surrogate host. Therefore, by understanding how these mechanisms work, it increases the likelihood that a strategy can be developed for effective exogenous gene expression.

3.1 GC content

The term GC content refers to the percentage of G and C bases in a DNA sequence. It can be used to describe a gene, a chromosome, a genome, and even any region of a particular DNA sequence. Different organisms can vary significantly in their genomic GC content. For example, *Plasmodium falciparum* has an extremely GC-poor genome, with a GC content of approximately 20%, while *Streptomyces coelicolor* possess a GC content as high as 72%. The GC contents of commonly used laboratory organisms are listed in Table 3.

Species	Genomic GC content (%)
Escherichia coli	51
Saccharomyces cerevisiae	38
Arabidopsis thaliana	36
Caenorhabditis elegans	36
Drosophila melanogaster	33
Homo sapiens	41

Table 3. GC content varies among common organisms

Due to the difference in thermodynamic stability between the GC bonding pairs and the AT bonding pairs, GC content can affect the formation and stability of both DNA and RNA secondary structures, which are important factors in the regulation of gene expression (Kubo & Imanaka, 1989; Kudla et al., 2009). In bacteria, the Shine-Dalgarno ribosome binding site that is located in the 5′ untranslated region of the mRNA is relatively AU-rich. The presence of this high AT abundance and low secondary structure stability at the 5′ end of a coding region has been found to contribute significantly to producing high translation efficiency in bacteria (Allert et al., 2010; Desmit & Vanduin, 1990). Furthermore, Kudla et al.

have demonstrated that the addition of these types of AU-rich leader sequences to the 5′ untranslated region of mRNAs can improve the expression levels of otherwise poorly expressed proteins (Kudla et al., 2009). In a recent systematic study of 340 genomes from various groups of organisms including bacteria, archaea, fungi, plants, insects, fishes, birds, and mammals, Gu and colleagues discovered a trend of reduced mRNA stability near the start codon in most organisms except birds and mammals and that this reduction results in changes in mRNA stability that are correlated with genomic GC content (Gu et al., 2010).

In birds and mammals, however, the genome-wide trend of reduced mRNA stability near the translation initiation site has not been observed, even though the GC content in these organisms is not significantly different from the species where such a trend was originally observed (Gu et al., 2010). The authors speculate that this difference is due to the isochore-type structure in the genomes of these organisms. An isochore is the result of a high variation in GC content over large-scale sequences within a genome (Bernardi, 1995). Within an isochore structure, however, the GC content is generally homogeneous regardless of the heterogeneous nature of the remainder of the genome (Figure 3) (Eyre-Walker & Hurst, 2001). It is important to note that, unlike in *E. coli*, high GC content within the coding region usually increases expression in mammalian cells (Kudla et al., 2006). Kudla and colleagues have found that GC-rich genes in mammalian cells were transcribed more efficiently than alternate, GC-poor versions of the same gene, leading to higher protein production. In fact, the 5′ cap and Kozak consensus sequence located on the 5′ untranslated region normally have a GC-rich composition in eukaryotic genes (Kozak, 1987).

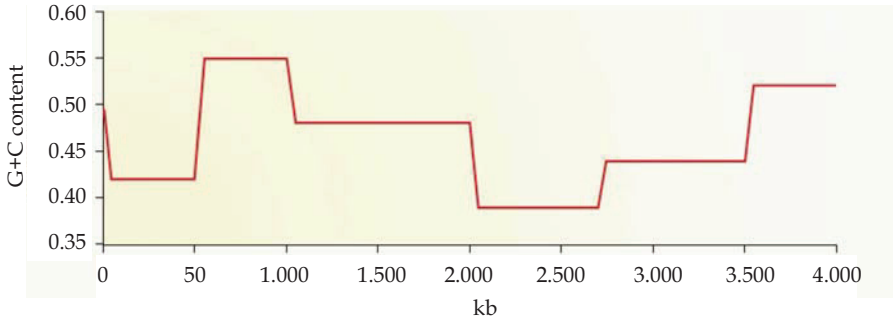

Fig. 3. The classic isochore model of genomic GC content. Used with permission from (Eyre-Walker & Hurst, 2001)

It is widely accepted that genomic GC content has co-evolved with the gene expression machinery to ensure optimal expression for the fitness of the host (Andersson & Kurland, 1990; Kudla et al., 2009). Therefore, with regards to expression of exogenous genes, the difference in the GC contents between the target genes, especially at the 5′ end, and the expression host can also impact the expression level of foreign genes. The difficulty in expressing *Plasmodium falciparum* genes in *E. coli* is hypothesized to be attributed to its extreme low GC content and the possibility of degradation of mRNA by ribonuclease E (McDowall et al., 1994; Plotkin & Kudla, 2011). Plotkin and Kudla have also predicted that more than 40% of human genes would be expressed poorly in *E. coli* without modification due to the relatively high GC content in the 5′ end of mRNA and subsequent low 5′ folding energy (Plotkin & Kudla, 2011).

3.2 Codon usage bias

In addition to determining mRNA stability and secondary structure organization, another feature of every genome that is impacted by GC content is its codon usage profile. The 20 amino acids commonly found in protein sequences are all encoded from a series of 61 different nucleotide triplets. The redundancy of this coding system necessarily allows the same amino acid to be encoded by several different codons. For example, the amino acids alanine and serine can be encoded using either four or six codons, respectively (Table 4). This innate degeneracy that is built into the genetic code has evolved to play a role in protecting DNA sequences from otherwise deleterious mutations by preserving their resultant protein sequences despite the inevitable incorporation of mutations at the genetic level, effectively silencing these mutations. However, the available synonymous codons are not used at equal frequencies across all species, nor across different regions within the same genome, and sometimes not even within the same gene (Andersson & Kurland, 1990; Kurland, 1991). Predictably, the discrepancy of codon usage profiles is greatest between remotely related species, while more closely related species are more likely to share similar codon preferences. Although the mechanistic processes underlying how an organism develops a specific codon bias has not been completely resolved (Chamary et al., 2006; Hershberg & Petrov, 2008), the GC content of the preferred codon chosen is thought to be the single most important factor determining codon usage biases across genomes (Plotkin & Kudla, 2011).

		Second Position									
		U		**C**		**A**		**G**			
		Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid		
First Position	**U**	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	**U**	**Third Position**
		UUC		UCC		UAC		UGC		**C**	
		UUA	Leu	UCA		UAA	**STOP**	UGA	**STOP**	**A**	
		UUG		UCG		UAG		UGG	Trp	**G**	
	C	CUU		CCU	Pro	CAU	His	CGU	Arg	**U**	
		CUC		CCC		CAC		CGC		**C**	
		CUA		CCA		CAA	Gln	CGA		**A**	
		CUG		CCG		CAG		CGG		**G**	
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	**U**	
		AUC		ACC		AAC		AGC		**C**	
		AUA		ACA		AAA	Lys	AGA	Arg	**A**	
		AUG	Met	ACG		AAG		AGG		**G**	
	G	GUU	Val	GCU	Ala	GAC	Asp	GGU	Gly	**U**	
		GUC		GCC		GAC		GGC		**C**	
		GUA		GCA		GAA	Glu	GGA		**A**	
		GUG		GCG		GAG		GGG		**G**	

Table 4. Redundancy in the genetic code allows more than one codon to specify a particular amino acid

Although it was initially believed that synonymous codon substitutions were simply examples of fortuitous silent mutations, more recent research has revealed that codon usage patterns can directly affect important cellular processes such as the efficiency of transcription and translation, the accuracy of protein translation and even the process of protein folding (Angov, 2011; Zhang et al., 2009). It is therefore conceivable that the specific codon usage pattern of an organism has co-evolved along with other cellular machinery in order to provide for optimal gene expression and protein function of the host genes within their natural environment (Grantham et al., 1981). In prokaryotes, for example, the frequency of a codon being used correlates positively with the intracellular abundance of its corresponding tRNA (Bulmer, 1987; Dong et al., 1996). It therefore follows that the expression of non-native genes is hampered by the existence of variation in their respective codon usage pattern compared to the host organism. This hypothesis has been supported throughout the long history of exogenous gene expression, revealing that the same DNA sequence is often expressed at different efficiencies in different organisms (Gustafsson et al., 2004). This is due to the foreign DNA sequence containing codons that are rarely used in the host, a situation that leads to low levels of translational efficiency and protein expression (Kane, 1995; Kim & Lee, 2006; Rosano & Ceccarelli, 2009) due to a reduced translation elongation rate caused by the imbalance between the codons used in the target gene sequence and the available pool of charged tRNA in the host. These expression problems are then compounded by any incompatibility between the host translation machinery and the mRNA secondary structure due to changes in GC content from alternate codon usage patterns (Kim & Lee, 2006; Wu et al., 2004).

To overcome these problems, a common strategy aimed at enhancing the expression of non-native genes in a surrogate host is that of codon optimization. This process encompasses the replacement of rare codons within the DNA sequence in order to closely match the host codon usage bias while retaining 100% identity to the original amino acid sequence. This process of codon optimization also allows for the simultaneous modification of predicted mRNA secondary structures that could result from changes in the GC content. This process is especially helpful in eliminating structures at the 5′ end of coding regions, where they have an increased likelihood of interfering with downstream protein expression (Wu et al., 2004) *Cis*-acting negative regulatory elements within the coding sequence are also eliminated in order to reduce the chance of repression, therefore improving expression (Graf et al., 2000). The codon optimization process can be achieved experimentally either through multiple stages of site-directed mutagenesis on directly cloned DNA, or by resynthesis of the target gene *de novo*. The former method may be preferred if there are a limited number of codons that must be changed, however, the later method has become more and more practical due to improvements in the gene synthesis process that have both reduced the cost and time required to generate synthetic DNA sequences. In general, the codon optimization process has been shown to increase expression of a typical mammalian gene five- to fifteen-fold when expressed in an *E. coli* host (Burgess-Brown et al., 2008; Gustafsson et al., 2004). Similarly, expression of prokaryotic genes in eukaryotic cells can be improved significantly using this method as well (Patterson et al., 2005; Zolotukhin et al., 1996; Zur Megede et al., 2000).

3.3 Mechanisms for removal and silencing of exogenous genes

For an exogenous gene to be expressed in a non-native host, the foreign DNA must be physically delivered into the host cell and then properly integrated into the gene expression

and regulation network within the host. Decades of research in the fields of molecular and cellular biotechnology have provided many effective techniques for the introduction of genetic material into both prokaryotic and eukaryotic hosts, however, after the gene has been transferred into the host cell, it needs to be recognized and processed by the host cells replication, transcription and translation machinery before it can be expressed as a functional protein. However, because expression of a foreign gene is often deleterious to host survival under wild-type conditions, many organisms have evolved defense mechanisms that remove or silence foreign DNA in order to protect themselves from this potentially detrimental process. In bacteria, for example, the invading foreign DNA can be cleaved by restriction endonucleases that recognize specific, non-self, nucleotide sequences, in a phenomenon referred to as restriction. In this process the native genetic material is often methylated at certain positions by methylase enzymes, therefore preventing recognition and degradation by the restriction endonucleases, and ensuring the maintenance and expression of native DNA sequences. This restriction modification system was first discovered in the 1960s and since that time has been demonstrated to be common in many bacterial species (Wilson & Murray, 1991). The restriction system, however, is not the only defense mechanism that has been developed to protect the host from expression of foreign genetic material. It has been demonstrated that Gram-negative bacteria are capable of selectively repressing horizontally acquired genes through their interaction with a histone-like nucleoid structuring (H-NS) protein. This phenomenon, termed xenogeneic silencing, was first discovered in 2006 by Navarre, Lucchini, Oshim and colleagues (Lucchini et al., 2006; Navarre et al., 2006; Oshima et al., 2006). The H-NS protein responsible for xenogeneic silencing belongs to a family of nucleoid-associated proteins that bind to AT-rich DNA sequences with low sequence specificity. In the case of xenogeneic silencing, H-NS protein targets the laterally acquired sequence because it exhibits a lower GC content than the host genome, allowing it to selectively repress the expression of exogenous DNA.

Unlike the prokaryotic approaches for silencing of exogenous DNA sequences, no mechanism for the direct removal of foreign genetic material has yet been proposed to function in eukaryotic organisms. Nonetheless, the expression of exogenous DNA in plants and mammalian cells often suffers from low efficiency due to epigenetic modification. These modifications lead to unstable expression and, in extreme cases, silencing of the transgene over time. Silencing can occur at either the transcriptional or post-transcriptional level through changes in the methylation status of the sequence, histone modification, or RNA interference (Pal-Bhadra et al., 2002; Pikaart et al., 1998; Riu et al., 2007). Regardless of the protective measures taken, these mechanisms are all employed by the host to regulate expression of exogenous genes and protect it from deleterious effects. One final concern that cannot yet be controlled for is that, due to the random integration following chromosomal introduction of an exogenous gene into the host chromosome, expression of the transgene can be highly dependent on the site of insertion. Depending on the location of integration, various position effects and epigenetic events often result in high variation of the expression level between individual expression attempts. While there is no way to reliably control for genomic insertion position of exogenous genes in the majority of cases, several elements have been proposed that can help to counteract the resultant position effects and achieve sustained transgene expression. These elements are discussed in section 4.4.

4. Regulatory sequences that must be considered for optimal expression

By developing a comprehensive understanding of the mechanisms underlying gene expression and appreciating how factors such as GC content and codon usage bias influence protein expression in non-native hosts, investigators can begin to develop theoretical guidelines for the rational design of DNA sequences optimally tuned for heterologous expression in their target organism. This approach is especially attractive, with the reduced time and cost of gene synthesis allowing for *de novo* production of complete genes and even entire expression cassettes making it possible to simply design a gene sequence and begin working. However, there are additional concerns that must be addressed prior to successful expression of an exogenous gene sequence. Besides the optimization of the coding region, regulatory sequences that are not transcribed or translated should also be taken into consideration in order to achieve optimal expression. Although not expressed in the final protein product, these elements are involved in the transcription, translation and long-term maintenance of target genes in the surrogate host, making their optimization just as important as optimization of the coding sequence itself.

4.1 Regulatory elements involved in transcription

The process leading from a gene to a functional protein starts with transcription by RNA polymerase. Therefore transcription initiation is often an important point of control for exogenous protein expression. The driving force behind recruiting and binding the polymerase that will transcribe the DNA to mRNA is the promoter sequence that is required to recruit the host's transcription machinery. Even though the promoter itself is not transcribed or translated, choosing a promoter that can be efficiently processed by the host's machinery therefore has a significant impact on the success of the design strategy. Commonly, strong, constitutive promoters that are normally used to drive the expression of endogenous housekeeping genes in the expression host are chosen for high level expression of exogenous genes. For example, the T7, alcohol dehydrogenase 1 (ADH1) and human elongation factor 1 α (EF1α) promoters are commonly employed for heterologous protein expression in *E. coli*, *S. cerevisiae* and mammalian cells, respectively. Viral promoters such as the cytomegalovirus immediate early (CMV IE) promoter and the Simian virus 40 (SV40) regulatory sequence are also used to drive transgene expression in mammalian cells as well. It is important to note, however, that while the strength of the promoter used can at least partially determine the level of transgene expression, different promoters can have variable rates of transcription across different cell lines. For this reason, the selection of an appropriate promoter should be determined on a case-by-case basis. Recent studies have systematically compared many of the commonly used promoters in a variety of cell types (Norrman et al., 2010; Qin et al., 2010) (Figure 4). These types of references are an excellent source of information when designing constructs with specific expression needs.

It is also important to remember that promoter sequences can be designed *de novo* similar to gene sequences, and that designing a specific primer upstream of a gene construct may be beneficial if no native alternative promoter sequences are available. Analysis of a large number of prokaryotic and eukaryotic promoters has revealed that many promoters contain a conserved core sequence that is essential for recognition and binding of RNA polymerase and its cofactors. Through incorporation of these conserved sequences, it may be possible to specifically design a promoter sequence, allowing one to tailor expression of their genetic

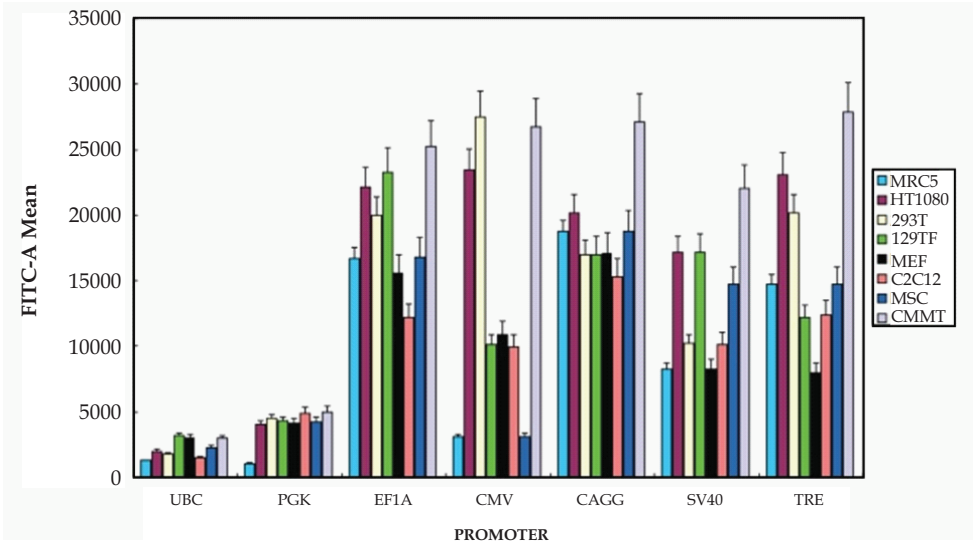

Fig. 4. Systematic comparison of different promoters in different mammalian cell types. Originally published in (Qin et al., 2010)

construct to their specific needs. In prokaryotes, this conserved sequence is known as the Pribnow box, and consists of a consensus sequence of six nucleotides, TATAAT (Pribnow, 1975). In addition, there is another conserved element often found 17 bp upstream of the Pribnow box. This upstream region has a consensus TTGACAT sequence that has been shown to be crucial for transcription initiation (Rosenberg & Court, 1979). In eukaryotic organisms, the counterpart to the Pribnow box is the TATA box with a consensus sequence of TATAAA. Besides recruiting the associated transcription machinery, these core promoter elements are also crucial in defining where RNA synthesis starts. In prokaryotes, RNA synthesis usually begins 10 bp downstream of the Pribnow box, whereas the first transcribed nucleotide is located approximately 25 bp downstream of the TATA box in eukaryotes. Therefore in addition to the use of an appropriate core promoter sequence, the location of that promoter sequence relative to the coding region should also be carefully considered to ensure complete transcription of the target genes.

It is important to note that although this minimal core promoter is essential for transcription, it alone is often not adequate to drive high level protein expression. In eukaryotes, DNA elements known as enhancers are often employed in tandem with the core promoter to enhance gene expression through the recruitment of additional transcription factors. These enhancers can be found at various locations, including upstream of the core promoter, within the introns of the gene driven by the core promoter, and downstream of the genes it regulates as well (Levine & Tjian, 2003). Although the mechanistic function of most enhancers is still not well understood, some well-studied viral enhancer elements are often included in common expression vectors as a means to increase the transcription efficiency of exogenous sequences. For example, the CMV IE enhancer has been shown to be capable of improving gene expression levels by 8- to 67-fold in lung epithelial cells when combined with several weak promoters (Yew et al., 1997) and Li and colleagues have further

demonstrated that adding an SV40 enhancer to the CMV IE enhancer/promoter or 3′ end of the polyadenylation site can increase exogenous gene expression in mouse muscle cells by up to 20-fold (Li et al., 2001).

4.2 Regulatory elements involved in translation

Just as with the requirement of a core promoter sequence for the initiation of transcription, the presence of certain, conserved sequences at the 5′ untranslated region of mRNA sequences are essential for the initiation of translation. In prokaryotic organisms, the Shine-Dalgarno sequence on the transcribed mRNA serves this function by acting as the ribosome binding site (RBS). This consensus sequence is composed of six nucleotides, AGGAGG, which are complementary to the anti-Shine-Dalgarno sequence located at the 3′ end of the 16S rRNA in the ribosome. During the initiation of translation the ribosome is recruited to the mRNA by this complementary base paring between the RBS and the 16S rRNA. For this reason, the classic RBS is included as a standard element in the Registry of Standard Biological Parts (http://partsregistry.org/). Also included in the registry is a collection of constitutive prokaryotic RBSs containing the Shine-Dalgarno sequence as well as flanking sequences that are known to affect translation. These sequences are invaluable when designing promoter and gene sequences, as their incorporation is required for efficient expression of the synthetic construct.

In eukaryotes, the 40S ribosomal subunit helps to serve this purpose by attaching to initiation factors that assist in the process of scanning the mRNA, with the Kozak sequence acting as the main initiator for translation (Kozak, 1986, 1987). This translational process most commonly begins at the AUG codon closest to the 5′ end of the mRNA, however, this is not always the case. Kozak et al. have demonstrated that the distance from the 5′ end, the sequence surrounding the first AUG codon, and its steric relationship with the 40S ribosomal subunit all contribute to determining the actual initiation site location. However, it has been routinely demonstrated that placing the promoter and Kozak sequence upstream of the initiating codon serves to induce increased expression of target gene sequences (Morita et al., 2000).

Besides the optimization of the codon usage pattern in the coding region, additional considerations must be taken into account when expressing prokaryotic genes in eukaryotic hosts or vice versa. Genes cloned directly from the genomic library of a eukaryotic organism usually cannot be expressed successfully in a prokaryotic host due to the presence of intervening, non-coding regions within the sequence. Unlike eukaryotes, prokaryotes lack the RNA splicing mechanisms required to remove these intron sequences and produce a mature mRNA. Therefore, any introns present within the expression construct must be eliminated prior to introduction into the prokaryotic host.

4.3 Elements for simultaneous expression of multiple genes in eukaryotes

Conversely, a significant obstacle towards the expression of genomically cloned bacterial genes in a eukaryotic host is the inability of the host to synthesize proteins polycistronically from a single mRNA. Unlike in prokaryotes, where translation of multiple adjacent genes from one promoter is common, translation in eukaryotic cells normally requires the presence of a methyl-7-G(5′)pppN cap at the 5′ end of the mRNA prior to recognition by the

translation initiation complex at the start of peptide synthesis (Pestova et al., 2001). There are strategies, however, that allow for co-expression of two or more genes in eukaryotic cells. On the most basic level, it is possible to express each gene independently from its own promoter, either through the introduction of multiple vectors, or introduction of a single vector containing multiple promoters. An alternate approach is expression of the multiple genes using a polycistronic expression vector that takes advantage of either IRES (Internal Ribosomal Entry Site) or 2A elements. Derived from a viral linker sequence, the IRES element allows for 5′-cap-independent ribosomal binding and translation initiation directly at the start codon of the downstream gene, thus enabling translation of multiple ORFs from a single mRNA (Jackson, 1988; Jang et al., 1988). Although known IRES sequences vary in length and sequence, certain secondary structures have been shown to be conserved and important for the function of the elements (Baird et al., 2006). The most widely used IRES sequence for expression in mammalian cells is the one derived from encephalomyocarditis virus (EMCV) (de Felipe, 2002). Similar to the IRES elements, 2A elements are viral sequences that can also be used as a short linker region to provide translation of two or more genes driven off of a single promoter. Translation of the 2A element causes an interaction between the newly synthesized sequence and the exit tunnel of the ribosome. This interaction causes a "skipping" of the last peptide bond at the C terminus of the 2A sequence. Despite this missing bond, the ribosome is able to continue translation, creating a second, independent protein product. To ensure continuous translation, the stop codon of the ORF upstream of the 2A element must be mutated to avoid unnecessary termination. By using a combination of various IRES and 2A elements, investigators have demonstrated polycistronic expression of five genes simultaneously from a single promoter in mammalian cells (Szymczak & Vignali, 2005), illustrating how they can be used to simulate the polycistronic expression of some bacterial genes.

4.4 Elements for sustained maintenance and expression

Integration of exogenous DNA sequences into a host chromosome is usually required for sustained transgene expression in mammalian cells. Because the insertion event preceding expression is largely random, the expression level of the integrated gene can be greatly impacted by the surrounding sequences and chromatin structure. As a consequence, unstable expression and high variability between individual clones are the two major issues associated with transgene expression. In addition, if insertion of the exogenous genes occurs within or in close vicinity to a required host gene, the health or survivability of the host can be negatively impacted. To aid in controlling for this type of negative regulation, several DNA elements capable of preventing these types of position effects and stabilizing transgene expression have been discovered (Table 5). These DNA elements are naturally found in mammalian genomes and are crucial for regulating the proper expression of endogenous genes. The locus control regions (LCRs) can enhance transcription of linked genes and also enable copy number-dependent gene expression (Li et al., 2002), however, their large size and tissue-specific nature constrain their application in a variety of mammalian cell types (Kwaks & Otte, 2006). Insulators, also known as barriers or enhancer-blocking elements, are DNA sequences that can protect genes from the transcriptionally inactive heterochromatin or the action of enhancers and repressors (Recillas-Targa et al., 2004). As an example, the best-characterized insulator, cHS4 (chicken β-globin hypersensitive site 4), has been shown to stabilize transgene expression over a long period

of time (Pikaart et al., 1998) and facilitate efficient integration of expressed sequences (Recillas-Targa et al., 2004). Similar to insulators, STARs (stabilizing and antirepressor elements) are specifically used to block repression. Another type of DNA sequence, known as the ubiquitous chromatin opening element (UCOE) is derived from promoters of ubiquitously expressed genes. These elements have been shown to improve and stabilize transgene expression in a tissue-nonspecific manner, most likely through the maintenance of an active chromatin structure (Williams et al., 2005). Matrix attachment regions (MAR) are elements that mediate the attachment of the chromosome to the nuclear matrix and, as such, are also widely used in DNA for sustained transgene expression. These elements have also been shown to counteract position dependent insertion effects and prevent transgene silencing in a variety of cell types and transgenic animals (reviewed by Harraghy and colleagues (Harraghy et al., 2008)).

Element	Size	Increased expression	Stability of expression	Cell type-specific	Copy number-dependent	Position-independent
LCR	16 kb	Unknown	Yes	Yes	Yes	Yes, if powerful enough
Insulator	1.2-2.4 kb	Unknown	Yes	Unknown	No	Majority Yes
UCOE	2.5-8 kb	Yes	Yes	No	Unknown	Yes
MAR	~3 kb	Yes	Yes	No	No	Majority Yes
STAR	0.5-2 kb	Yes	Yes	No	Yes	Yes

Table 5. Many different elements can be used to enhance and stabilize transgene expression in mammalian cells. Modified from (Kwaks & Otte, 2006) and (Harraghy et al., 2008)

5. Mammalian expression of the bacterial luciferase gene cassette: A case study in exogenous expression

Over the years there have been myriad examples of exogenously expressed genes. A recent example that highlights many of the considerations discussed here is the adaption of the bacterial luciferase gene cassette to function autonomously in a human cell line. The bacterial luciferase gene cassette, commonly referred to as the *lux* cassette, had been utilized in prokaryotic systems for almost 20 years prior to its first successful expression in a eukaryotic cell, and, even then, required almost another decade before it was successfully expressed in a human cell line. By following the development of the *lux* system from a strictly bacterial genetic system through its development into a eukaryotic reporter cassette, it is possible to review not only the genetic modifications that are required for exogenous gene expression, but also the thought process that leads researchers to implement these modifications.

5.1 Bacterial luciferase background

The bacterial luciferase (*lux*) gene cassette is a series of five genes whose protein products synergistically work together to produce a luminescent signal at 490 nm in the blue range of the visible spectrum (Close et al., 2009). Two of the five genes (*luxA* and *luxB*) form the heterodimeric luciferase protein, while the remaining three genes (*luxC*, *luxD*, and *luxE*) are responsible for the production of a long chain aliphatic aldehyde co-substrate upon which the luciferase protein acts (Meighen, 1991). The remaining co-substrates, $FMNH_2$ and O_2, are naturally present within the host and can be directly scavenged by the enzyme. Upon binding of the substrate complex to the luciferase dimer, the complex becomes oxidized and releases a photon at 490 nm (Figure 5). The turnover of this reaction is extremely slow, with the process taking as long as 20 sec at 20°C (Hastings & Nealson, 1977).

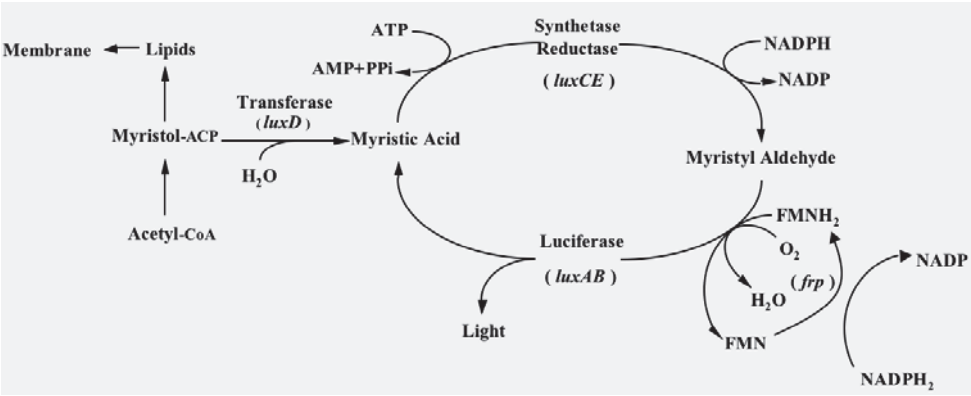

Fig. 5. The bioluminescent reaction catalyzed by the bacterial luciferase gene cassette. Reproduced with permission from (Close et al., 2009)

While these genes are widely distributed in prokaryotic organisms, the bioluminescent system they encode for is quite distinct from those commonly found in eukaryotes, such as the firefly or *Renilla* luciferase systems. Unlike these eukaryotic bioluminescence systems, the *lux* system is organized as a single operon, with all of the genes required for bioluminescent production driven from a single promoter. In addition, its prokaryotic origin means that it is optimized for function in a cellular background that is free from extensive compartmentalization. It is therefore not surprising that extensive genetic modifications were required prior to successful expression in the distantly related human cellular background. These modifications present an interesting case study of the considerations that must be made when exogenously expressing any gene in a non-native host organism.

5.2 Initial attempts at exogenous expression

The first attempts to express the *lux* system outside of bacteria started in the 1980's. After realizing the benefits offered by the fully autonomous expression of light as a bioluminescent reporter system in bacterial species, there was an increasing interest in evolving this system to function in a wider variety of organisms in order to take advantage of it usefulness across an increasingly broad range of circumstances. These initial attempts focused on expression of only the *luxA* and *luxB* genes rather than full cassette expression,

seeking to first determine how to make the luciferase function and then apply the lessons learned to expression of the remaining *lux* genes.

Because eukaryotic organisms are not capable of polycistronic expression, the first modification made for the expression of the *luxA* and *luxB* genes was to place them each under the control of independent promoters (Koncz et al., 1987). This strategy allowed for the transcription of each mRNA sequence to occur independently. However, since each was placed on the same plasmid, their physical location of expression in the host should be proximal. This expression strategy circumvents the need for polycistronic expression, while simultaneously maximizing the chance that the *luxA* and *luxB* protein products will associate *in vivo* to produce a functional heterodimer. When this system was expressed in plants, cell extracts were capable of producing light in response to treatment with an aldehyde substrate. While this demonstrated the ability to exogenously express at least a portion of the *lux* cassette, it was still far from practical in terms of autonomous bioluminescent expression.

Moving forward from this dual promoter system in plants, several groups began experimenting with expressing the *luxA* and *luxB* genes as fusion products in yeast (Boylan et al., 1989; Kirchner et al., 1989), *Drosophila* (Almashanu et al., 1990), and even murine cell lines (Pazzagli et al., 1992). Regardless of the host origin, the results of these experiments were generally met with similar outcomes. When tested in yeast cells, the bioluminescent expression upon treatment with the aldehyde substrate was detectable above background, however, not as prevalent as bioluminescence from alternate prokaryotic systems tested under the same conditions (Boylan et al., 1989). When expression using this strategy was attempted using higher eukaryotic hosts such as *Drosophila* and murine cell lines, an interesting problem was encountered; bioluminescence was detectable but was determined to be highly temperature sensitive.

Because of the higher temperatures required for growth of the murine Ltk$^-$ cell line, the *lux* luciferase proteins were not able to maintain high levels of stability following gene expression. This resulted in extremely low levels of bioluminescent production from Ltk$^-$ cells transfected with the *luxA* and *luxB* genes when grown at their optimal temperature of 37°C. When the growth temperature was decreased to a tolerable, but not ideal temperature of 30°C, bioluminescent detection increased 10-fold (Pazzagli et al., 1992). The temperature-dependent nature of this bioluminescent decrease was additionally confirmed through further testing in *E. coli*, where it was determined that hosts expressing LuxA-LuxB fusion proteins were capable of producing a greater than 50,000-fold increase in bioluminescent production when grown at 23°C compared to growth at 37°C (Escher et al., 1989). This highlights the need to not only evaluate the potential genetic hurdles to exogenous expression of a target gene, but also to consider the physiological limitations constraining expression of the protein encoded from that gene as well. This constraint proved to be a significant challenge in the development of routine eukaryotic expression of these genes, and it would be another decade before it was overcome, finally leading to expression of the full *lux* cassette in a yeast cell model.

5.3 Autonomous bioluminescent expression from the *lux* cassette in yeast

Using the lessons that were learned from expression of both dual-promoter and fusion-based expression of the *luxA* and *luxB* genes detailed above, work continued toward the

expression of the full *lux* cassette in a eukaryotic host. The first major breakthrough came from the decision to express *lux* genes from the bacterium *Photorhabdus luminescens* rather than the classical *lux* model organism, *Vibrio harveyi* (Gupta et al., 2003). Unlike the *V. harveyi* template organism used in the previous attempts, *P. luminescens* is a terrestrial rather than marine bacterium. As such, it therefore has a higher native growth temperature, which leads to the stability of its protein products at a higher temperature than those encoded by *V. harveyi,* despite performing the same function *in vivo*. This simple change in selection for the source of the exogenous genes demonstrates how important the selection process can be when expressing genes in a foreign host. Without the innate structural stability offered by the *P. luminescens* proteins, no combination of genetic modifications would have been capable of inducing high-level expression in a eukaryotic host at its preferred growth temperature.

Having overcome the intrinsic problems with gene expression at the natural yeast growth temperature, there were still additional genetic modifications that would have to be considered before the full *lux* cassette could be autonomously expressed. The first important consideration was that of how to promote constitutive, high level expression of the genes themselves. This was accomplished through the incorporation of yeast-specific promoter sequences that had previously been demonstrated to drive high-level expression under the majority of growth conditions. These promoters, the glyceraldehyde 3′ phosphate dehydrogenase (GPD) and alcohol dehydrogenase 1 (ADH1) promoters, were used in place of the native upstream regions from the wild-type bacterial species that either have an inducer binding site or AT rich region (Meighen, 1991). The replacement of this AT rich promoter region with known, host-expressible promoters ensured that there would be high levels of transcription when the genes were expressed in the yeast surrogate.

Next, it was necessary for the researchers to develop a method for the expression of the five *lux* cassette genes simultaneously within the adopted host. Because *S. cerevisiae* is a eukaryote, it is not capable of carrying out the natural polycistronic expression of the cassette as would occur under wild-type conditions in a prokaryotic host. To overcome this hurdle, the polycistronic expression system was mimicked through the incorporation of IRES sites (Gupta et al., 2003). These IRES sites function as linker regions between the individual *lux* genes and allow for expression of multiple ORFs to be transcribed to a single piece of mRNA, but then translated individually through cap-independent ribosome recruitment during translation (Lupez-Lastra et al., 2005). While there are multiple organisms that are known to harbor these IRES elements, the researchers used an IRES sequence found natively from *S. cerevisiae* to ensure it would function efficiently in this system (Gupta et al., 2003).

Even with the addition of these IRES linker regions and multiple promoters, the shear number of genes that must be expressed for autonomous light production using the *lux* cassette still presented a significant obstacle for exogenous expression. To overcome this problem, it was determined that the most efficient expression strategy was to divide expression of the *lux* cassette between two independent expression vectors (Gupta et al., 2003). This created an expression system whereby the *luxA* and *luxB* genes were expressed independently from two promoters on a single vector, while the *luxC*, *luxD*, and *luxE* genes were expressed from a second vector and linked using IRES sequences (Figure 6). While the vectors used in this example are capable of episomal expression in yeast, it is important to

note that normally eukaryotic expression occurs after chromosomal integration of the transfected gene sequences. Since this process cannot control the integration location of the gene sequences, a dual vector expression strategy could potentially lead to distal integration of the gene sequences and increase the probability that expression of the different gene groups would occur with different efficiencies despite their use of identical promoter sequences.

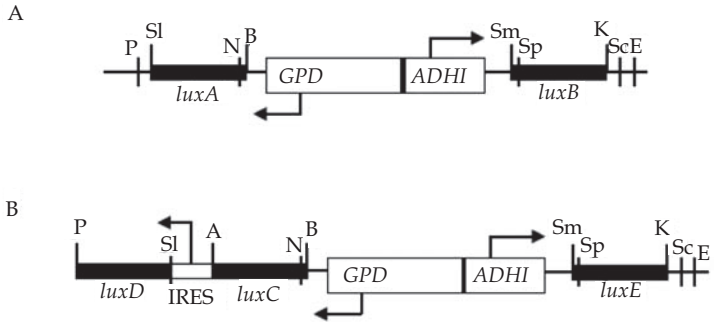

Fig. 6. Expression of the *lux* gene cassette in *S. cerevisiae* was made possible through A) independent expression of the *luxA* and *luxB* genes on one plasmid and B) expression of the remaining *lux* genes using a combination of multiple promoters and IRES linker regions from a second plasmid. Adapted from (Gupta et al., 2003)

Due to the extensive modifications performed to the *lux* cassette genes, they were capable of producing a well defined bioluminescent signal when expressed in *S. cerevisiae* (Gupta et al., 2003). This marked the first successful demonstration of *lux*-based autonomous bioluminescent production from a eukaryotic host organism. Despite this success, it was determined that the compartmentalization intrinsic to the eukaryotic nature of the yeast host was limiting access of the luciferase to its $FMNH_2$ co-substrate. Unlike prokaryotes, eukaryotes do not have large quantities of cytosolically available $FMNH_2$. This required an additional change to the *lux* expression strategy, whereby a flavin reductase gene (*frp*) was added to the *lux* cassette downstream of the *luxE* gene using the previously described IRES linker region and under control of the ADH1 promoter. This served to increase the amount of $FMHN_2$ available locally to the luciferase enzyme. This final modification both stabilized bioluminescent production and increased light output greater than 5-fold (Gupta et al., 2003). While not often considered during exogenous expression, this addition provides an excellent example of how the expression environment must be considered in addition to general genetic modifications. In the case of *lux* expression, the addition of the *frp* gene was sufficient to alter the environment to a more favorable condition; however, this may not always be the case and should be approached on a case-by-case basis.

5.4 Modification of the *lux* cassette for expression in mammalian cells

Following the successful demonstration of autonomous bioluminescence from the *lux* genes in *S. cerevisiae*, research was begun into its expression in human cell lines. It was initially believed that the modifications that had been established during development for yeast expression would be sufficient for expression in the human cellular background. If

this had been determined to be the case, it would have been possible simply to transfect human cells with the previously developed vectors and monitor bioluminescent output. Unfortunately, this was determined not to be true, and expression of the genes, even with the addition of human specific, strong promoters could not be detected at levels significantly above background (Close et al., 2010; Patterson et al., 2005). It was therefore necessary to again modify the *lux* expression system in order to promote expression in a human host cell line.

Just as with previous modification approaches, this work began by focusing on expression of only a subset of the *lux* cassette genes, *luxA* and *luxB*. Using the lessons learned from *S. cerevisiae* expression, the *luxA* and *luxB* genes were placed under the control of a strong, constitutive human promoter and linked using a human specific IRES linker region. While this did lead to the ability to detect bioluminescence from cell extracts upon supplementation with substrates, it was not a significant improvement over expression in a yeast host. With little more that could be done to improve expression through genetic organization and enhanced promoter sequences, the researchers turned to the process of codon-optimization in hopes of increasing transcriptional and translational efficiency and therefore increasing light output. The codon usage patterns for the *P. luminescens lux* genes were compared to the codon usage patterns of each amino acid for all known expressed human genes and then altered to more closely match the human codon preference. At this time, the gene sequences were also scanned for the presence of restriction and other regulatory sequences such as potential hairpins or terminator sequences. These sequences were then eliminated through the replacement of the DNA sequence with a sequence that matched the original amino acid output with 100% identity, but was computationally favored due to its closer match with human codon preferences and absence of regulatory sequences (Table 6) (Patterson et al., 2005). This codon-optimization process, along with the previously described modifications, was capable of boosting bioluminescent output 54-fold over expression of non-codon-optimized gene sequences. This significant change highlights how important the codon optimization process can be when exogenously expressing genes in a distantly related organism.

Gene	Predicted Start Position	Length (bp)	% GC	Number of Nucleotide Substitutions	Probability of Recognition as an Exon
wt*luxA*	61	1023	40%	N/A	0.70
co*luxA*	1	1083	54%	190	0.88
wt*luxB*	1	984	35%	N/A	0.97
co*luxB*	1	984	52%	188	0.99

Table 6. Comparison of the *luxA* and *luxB* genes in their wild-type (wt) and codon-optimized (co) forms. The probability of recognition as an exon was determined *in silico* using the genescan algorithm (http://genes.mit.edu). Adapted from (Patterson et al., 2005)

Based on the success of the codon-optimization process for expression of the *luxA* and *luxB* genes in a human host cell, work then immediately began on implementing expression of the full *lux* cassette for autonomous bioluminescent production from a human host. For this process, the vector that was developed for expression of *luxA* and *luxB* was maintained, and the additional *lux* genes were placed into a second vector, mimicking the strategy employed for full *lux* cassette expression in *S. cerevisiae*. One important change that was incorporated, however, was the replacement of the yeast specific glyceraldehyde 3' phosphate dehydrogenase and alcohol dehydrogenase 1 promoters with CMV and EF1-α promoters (Close et al., 2010). These promoters allowed for strong constitutive expression of the remaining *lux* genes in a way that would not be possible if the original bacterial AT rich regions or yeast promoters were used. The benefits of the codon-optimization process were again highlighted during optimization of the remaining *lux* genes. The removal of regulatory sequences had a dramatic effect on the expression of the *luxE* gene, where their presence would have moved the predicted translational start point back to the 102nd nucleotide of the DNA sequence. In addition, the GC content of each of the genes was significantly altered to more closely match that of human coding regions, aiding in the recognition, expression and stability of each of the gene sequences following transfection into the human cellular genome (Table 7). As before, the *frp* flavin reductase gene was included in these constructs as well to compensate for the diminished cytosolic availability of $FMNH_2$ in the highly compartmentalized eukaryotic host.

Gene	Predicted Start Position	Length (bp)	% GC	Number of Nucleotide Substitutions	Probability of Recognition as an Exon
wt*luxC*	1	1443	37%	N/A	0.921
co*luxC*	1	1443	60%	449	0.999
wt*luxD*	1	924	38%	N/A	0.875
co*luxD*	1	924	59%	294	0.999
wt*luxE*	102	1087	38%	N/A	0.443
co*luxE*	1	1113	60%	331	0.999
wt*frp*	1	613	47%	N/A	0.715
co*frp*	1	723	64%	249	0.999

Table 7. Codon-optimization of the remainder of the *lux* genes was responsible for significant changes in both transcriptional start sites and the overall GC content. Each of these changes contributed significantly to the probability of the sequence being recognized as a coding sequence in the human host as determined *in silico* using the genescan algorithm (http://genes.mit.edu). Reproduced with permission from (Close et al., 2010)

While the changes required to induce bioluminescent production from the *lux* cassette genes in the human cellular background were extensive, they were all necessary for proper function. The failure of even a single modification would lead to cells that may be capable of expressing the genes but not maintaining expression at a high enough level to be useful as a reporter system (Figure 7).

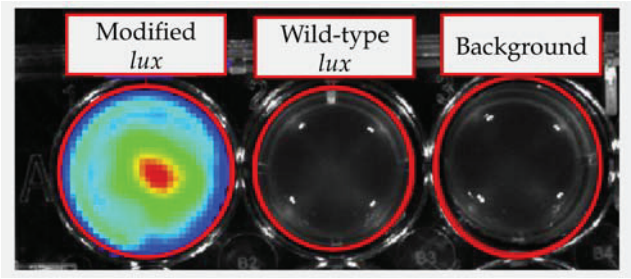

Fig. 7. Comparison of the bioluminescent expression from the *lux* genes expressed in a human host cell either following the modifications described above (modified *lux*) or without the aforementioned modifications (wild-type *lux*), and background light detection from host cells without *lux* genes (background). Adapted from (Close et al., 2010)

However, through the application of the techniques and considerations defined in this chapter, it was possible to develop not just one gene, but an entire cassette of six gene sequences from a reporter system once believed to function only in prokaryotic organisms, into a novel bioluminescent reporter system capable of being expressed in a human cell line with a signal bright enough to be seen through tissue similar to native eukaryotic genes such as firefly luciferase (Figure 8).

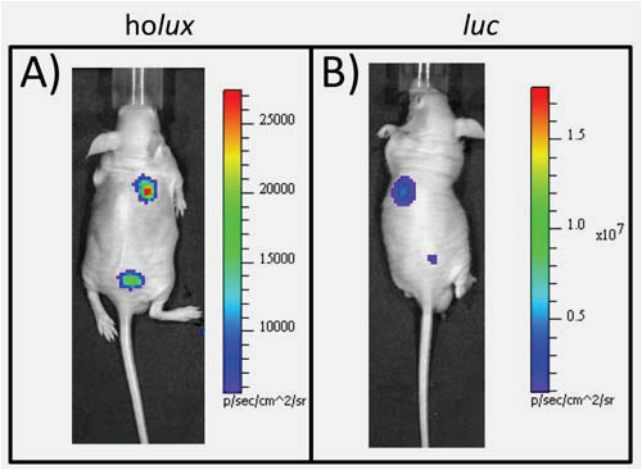

Fig. 8. Following modification of the full *lux* cassette, it was capable of being expressed in a human cell line host and producing bioluminescence at levels comparable to detection patterns of the native eukaryotic bioluminescent firefly luciferase (*luc*) gene. Adapted from (Close et al., 2011)

6. Conclusions

This chapter has detailed many of the concerns that must be considered when attempting to exogenously express a gene of interest in a foreign host. While a strong understanding of the transcriptional, translational and regulatory processes that dictate the maintenance and expression of all genes is a prerequisite for understanding the reasons why certain modifications must be performed in order to elicit high levels of exogenous expression, the examples provided here should be enough for the average researcher to begin developing an acceptable expression protocol. It is not a requirement that all of the modifications discussed in this chapter be applied to every gene, but a broad understanding of the possible changes can provide one with a wide variety of tools for expression of recalcitrant gene sequences. Just as with the *lux* cassette system, it is often necessary to perform more than one modification in order to induce acceptable levels of expression from foreign genes when expressed in a distal host organism. Often, proceeding in a step-wise fashion will yield clues as to which modifications will need to be performed, and which steps can be avoided, to save time and money when developing a new expression platform for a previously unexpressed gene sequence. It should also be noted that the methods detailed in this chapter are not all encompassing. In some cases, the host environment may simply not be suitable for expression of the target gene sequence and it may not be possible to alter that environment through the expression or deletion of additional genes. However, as the suite of exogenous expression techniques continues to grow via the discovery of new methods and our understanding of the cellular processes responsible for maintenance and expression of genes grows, the number of inexpressible genes will continue to fall.

7. Acknowledgments

Portions of this review reflecting work by the authors was supported by the National Science Foundation Division of Chemical, Bioengineering, Environmental, and Transport Systems (CBET) under award number CBET-0853780, the National Institutes of Health, National Cancer Institute, Cancer Imaging Program, award number CA127745-01, the University of Tennessee Research Foundation Technology Maturation Funding program, and the Army Defense University Research Instrumentation Program.

8. References

Abe, H., & Aiba, H. (1996). Differential contributions of two elements of rho-independent terminator to transcription termination and mRNA stabilization. *Biochimie,* 78, 11-12, pp. 1035-1042.

Acker, M. G., Shin, B.-S., Nanda, J. S., Saini, A. K., Dever, T. E., & Lorsch, J. R. (2009). Kinetic analysis of late steps of eukaryotic translation initiation. *Journal of Molecular Biology,* 385, 2, pp. 491-506.

Allert, M., Cox, J. C., & Hellinga, H. W. (2010). Multifactorial determinants of protein expression in prokaryotic open reading frames. *Journal of Molecular Biology,* 402, 5, pp. 905-918.

Almashanu, S., Musafia, B., Hadar, R., Suissa, M., & Kuhn, J. (1990). Fusion of *luxA* and *luxB* and its expression in *Escherichia coli*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*. *Journal of Bioluminescence and Chemiluminescence,* 5, 1, pp. 89-97.

Andersson, S. G. E., & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. *Microbiological Reviews,* 54, 2, pp. 198-210.

Angov, E. (2011). Codon usage: Nature's roadmap to expression and folding of proteins. *Biotechnology Journal,* 6, 6, pp. 650-659.

Baird, S. D., Turcotte, M., Korneluk, R. G., & Holcik, M. (2006). Searching for IRES. *RNA - A Publication of the RNA Society,* 12, 10, pp. 1755-1785.

Bernardi, G. (1995). The human genome: Organization and evolutionary history. *Annual Review of Genetics,* 29, 445-476.

Boeger, H., Bushnell, D. A., Davis, R., Griesenbeck, J., Lorch, Y., Strattan, J. S., et al. (2005). Structural basis of eukaryotic gene transcription. *FEBS Letters,* 579, 4, pp. 899-903.

Boylan, M., Pelletier, J., & Meighen, E. A. (1989). Fused bacterial luciferase subunits catalyze light emission in eukaryotes and prokaryotes. *Journal of Biological Chemistry,* 264, 4, pp. 1915-1918.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature,* 325, 6106, pp. 728-730.

Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., & Gileadi, O. (2008). Codon optimization can improve expression of human genes in *Escherichia coli*: A multi-gene study. *Protein Expression and Purification,* 59, 1, pp. 94-102.

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics,* 7, 2, pp. 98-108.

Close, D. M., Hahn, R., Patterson, S. S., Ripp, S., & Sayler, G. S. (2011). Comparison of human optimized bacterial luciferase, firefly luciferase, and green fluorescent protein for continuous imaging of cell culture and animal models. *Journal of Biomedical Optics,* 16, 4, pp. e12441.

Close, D. M., Patterson, S. S., Ripp, S., Baek, S. J., Sanseverino, J., & Sayler, G. S. (2010). Autonomous bioluminescent expression of the bacterial luciferase gene cassette (*lux*) in a mammalian cell line. *PLoS ONE,* 5, 8, pp. e047003.

Close, D. M., Ripp, S., & Sayler, G. S. (2009). Reporter proteins in whole-cell optical bioreporter detection systems, biosensor integrations, and biosensing applications. *Sensors,* 9, 11, pp. 9147-9174.

de Felipe, P. (2002). Polycistronic viral vectors. *Current Gene Therapy,* 2, 3, pp. 355-378.

Desmit, M. H., & Vanduin, J. (1990). Secondary structure of the ribosome binding site determines translation efficiency - A quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America,* 87, 19, pp. 7668-7672.

Dong, H. J., Nilsson, L., & Kurland, C. G. (1996). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *Journal of Molecular Biology,* 260, 5, pp. 649-663.

Dvir, A., Conaway, J. W., & Conaway, R. C. (2001). Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Current Opinion in Genetics & Development,* 11, 2, pp. 209-214.

Dvir, A., Conaway, R. C., & Conaway, J. W. (1996). Promoter escape by RNA polymerase II - A role for an ATP cofactor in suppression of arrest by polymerase at promoter-proximal sites. *Journal of Biological Chemistry,* 271, 38, pp. 23352-23356.

Ebright, R. H. (2000). RNA polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *Journal of Molecular Biology,* 304, 5, pp. 687-698.

Escher, A., Okane, D. J., Lee, J., & Szalay, A. A. (1989). Bacterial luciferase alpha-beta fusion protein is fully active as a monomer and highly sensitive *in vivo* to elevated temperature. *Proceedings of the National Academy of Sciences of the United States of America,* 86, 17, pp. 6528-6532.

Eyre-Walker, A., & Hurst, L. D. (2001). The evolution of isochores. *Nature Reviews Genetics,* 2, 7, pp. 549-555.

Falaschi, A. (2000). Eukaryotic DNA replication: a model for a fixed double replisome. *Trends in Genetics,* 16, 2, pp. 88-92.

Graf, M., Bojak, A., Deml, L., Bieler, K., Wolf, H., & Wagner, R. (2000). Concerted action of multiple *cis*-acting sequences is required for Rev dependence of late human immunodeficiency virus type 1 gene expression. *Journal of Virology,* 74, 22, pp. 10822-10826.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Research,* 9, 1, pp. R43-R74.

Gu, W. J., Zhou, T., & Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology,* 6, 2, pp. e1000664.

Gupta, R. K., Patterson, S. S., Ripp, S., & Sayler, G. S. (2003). Expression of the *Photorhabdus luminescens lux* genes (*luxA*, *B*, *C*, *D*, and *E*) in *Saccharomyces cerevisiae*. *FEMS Yeast Research,* 4, 3, pp. 305-313.

Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in Biotechnology,* 22, 7, pp. 346-353.

Harraghy, N., Gaussin, A., & Mermod, N. (2008). Sustained transgene expression using MAR elements. *Current Gene Therapy,* 8, 5, pp. 353-366.

Hastings, J., & Nealson, K. (1977). Bacterial bioluminescence. *Annual Reviews in Microbiology,* 31, 1, pp. 549-595.

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics,* 42, 1, pp. 287-299.

Jackson, R. J. (1988). RNA translation - Picornaviruses break the rules. *Nature,* 334, 6180, pp. 292-293.

Jang, S. K., Krausslich, H. G., Nicklin, M. J. H., Duke, G. M., Palmenberg, A. C., & Wimmer, E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *Journal of Virology,* 62, 8, pp. 2636-2643.

Kane, J. F. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Current Opinion in Biotechnology,* 6, 5, pp. 494-500.

Keck, J. L., & Berger, J. M. (2000). DNA replication at high resolution. *Chemistry & Biology,* 7, 3, pp. R63-71.

Kim, S., & Lee, S. B. (2006). Rare codon clusters at 5'-end influence heterologous expression of archaeal gene in *Escherichia coli*. *Protein Expression and Purification,* 50, 1, pp. 49-57.

Kirchner, G., Roberts, J. L., Gustafson, G. D., & Ingolia, T. D. (1989). Active bacterial luciferase from a fused gene: Expression of a *Vibrio harveyi luxAB* translational fusion in bacteria, yeast and plant cells. *Gene,* 81, 2, pp. 349-354.

Kolitz, S. E., & Lorsch, J. R. (2010). Eukaryotic initiator tRNA: Finely tuned and ready for action. *FEBS Letters,* 584, 2, pp. 396-404.

Koncz, C., Olsson, O., Langridge, W. H. R., Schell, J., & Szalay, A. A. (1987). Expression and assembly of functional bacterial luciferase in plants. *Proceedings of the National Academy of Sciences USA,* 84, 1, pp. 131-135.

Kozak, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell,* 44, 2, pp. 283-292.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research,* 15, 20, pp. 8125-8148.

Kubo, M., & Imanaka, T. (1989). mRNA secondary structure in an open reading frame reduces translation efficiency in *Bacillus subtilis*. *Journal of Bacteriology,* 171, 7, pp. 4080-4082.

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., & Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biology,* 4, 6, pp. 933-942.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science,* 324, 5924, pp. 255-258.

Kurland, C. G. (1991). Codon bias and gene expression. *FEBS Letters,* 285, 2, pp. 165-169.

Kwaks, T. H. J., & Otte, A. P. (2006). Employing epigenetics to augment the expression of therapeutic proteins in mammalian cells. *Trends in Biotechnology,* 24, 3, pp. 137-142.

Lafontaine, D. L. J., & Tollervey, D. (2001). The function and synthesis of ribosomes. *Nature Reviews Molecular Cell Biology,* 2, 7, pp. 514-520.

Lavergne, J. P., Reboud, A. M., Sontag, B., Guillot, D., & Reboud, J. P. (1992). Binding of GDP to a ribosomal protein after elongation factor-2 dependent GTP hydrolysis. *Biochimica et Biophysica Acta - Gene Structure and Expression,* 1132, 3, pp. 284-289.

Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature,* 424, 6945, pp. 147-151.

Li, Q. L., Peterson, K. R., Fang, X. D., & Stamatoyannopoulos, G. (2002). Locus control regions. *Blood,* 100, 9, pp. 3077-3086.

Li, S., MacLaughlin, F. C., Fewell, J. G., Gondo, M., Wang, J., Nicol, F., et al. (2001). Muscle-specific enhancement of gene expression by incorporation of SV/40 enhancer in the expression plasmid. *Gene Therapy,* 8, 6, pp. 494-497.

Lucchini, S., Rowley, G., Goldberg, M. D., Hurd, D., Harrison, M., & Hinton, J. C. D. (2006). H-NS mediates the silencing of laterally acquired genes in bacteria. *PLoS Pathogens,* 2, 8, pp. 746-752.

Lupez-Lastra, M., Rivas, A., & BarrÌa, M. (2005). Protein synthesis in eukaryotes: the growing biological relevance of cap-independent translation initiation. *Biological Research,* 38, 121-146.

McDowall, K. J., Linchao, S., & Cohen, S. N. (1994). A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *Journal of Biological Chemistry,* 269, 14, pp. 10790-10796.

Meighen, E. A. (1991). Molecular biology of bacterial bioluminescence. *Microbiological Reviews,* 55, 1, pp. 123-142.

Moreira, D., Kervestin, S., Jean-Jean, O., & Philippe, H. (2002). Evolution of eukaryotic translation elongation and termination factors: Variations of evolutionary rate and genetic code deviations. *Molecular Biology and Evolution,* 19, 2, pp. 189-200.

Morita, S., Kojima, T., & Kitamura, T. (2000). Plat-E: An efficient and stable system for transient packaging of retroviruses. *Gene Therapy,* 7, 12, pp. 1063-1066.

Murakami, K. S., & Darst, S. A. (2003). Bacterial RNA polymerases: The whole story. *Current Opinion in Structural Biology,* 13, 1, pp. 31-39.

Navarre, W. W., Porwollik, S., Wang, Y. P., McClelland, M., Rosen, H., Libby, S. J., et al. (2006). Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science,* 313, 5784, pp. 236-238.

Nilsson, J., & Nissen, P. (2005). Elongation factors on the ribosome. *Current Opinion in Structural Biology,* 15, 3, pp. 349-354.

Norrman, K., Fischer, Y., Bonnamy, B., Sand, F. W., Ravassard, P., & Semb, H. (2010). Quantitative comparison of constitutive promoters in human ES cells. *PLoS ONE,* 5, 8, pp. e12413.

Oldfield, S., & Proud, C. G. (1993). Phosphorylation of elongation factor-2 from the lepidopteran insect, spodoptera frugiperda. *FEBS Letters,* 327, 1, pp. 71-74.

Oshima, T., Ishikawa, S., Kurokawa, K., Aiba, H., & Ogasawara, N. (2006). *Escherichia coli* histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Research,* 13, 4, pp. 141-153.

Pal-Bhadra, M., Bhadra, U., & Birchler, J. A. (2002). RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in *Drosophila*. *Molecular Cell,* 9, 2, pp. 315-327.

Patterson, S. S., Dionisi, H. M., Gupta, R. K., & Sayler, G. S. (2005). Codon optimization of bacterial luciferase (*lux*) for expression in mammalian cells. *Journal of Industrial Microbiology & Biotechnology,* 32, 3, pp. 115-123.

Pazzagli, M., Devine, J. H., Peterson, D. O., & Baldwin, T. O. (1992). Use of bacterial and firefly luciferases as reporter genes in DEAE-dextran mediated transfection of mammalian cells. *Analytical Biochemistry,* 204, 2, pp. 315-323.

Pestova, T. V., Kolupaeva, V. G., Lomakin, I. B., Pilipenko, E. V., Shatsky, I. N., Agol, V. I., et al. (2001). Molecular mechanisms of translation initiation in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America,* 98, 13, pp. 7029-7036.

Pikaart, M. I., Recillas-Targa, F., & Felsenfeld, G. (1998). Loss of transcriptional activity of a transgene is accompanied by DNA methylation and histone deacetylation and is prevented by insulators. *Genes & Development,* 12, 18, pp. 2852-2862.

Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics,* 12, 1, pp. 32-42.

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences of the United States of America,* 72, 3, pp. 784-788.

Qin, J., Zhang, L., Clift, K., Hulur, I., Xiang, A., Ren, B., et al. (2010). Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One,* 5, 5, pp. e10611.

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. *Cell,* 108, 4, pp. 557-572.

Recillas-Targa, F., Valadez-Graham, V., & Farre, C. M. (2004). Prospects and implications of using chromatin insulators in gene therapy and transgenesis. *Bioessays,* 26, 7, pp. 796-807.

Richardson, J. P. (2003). Loading Rho to terminate transcription. *Cell,* 114, 2, pp. 157-159.

Riis, B., Rattan, S. I. S., Clark, B. F. C., & Merrick, W. C. (1990). Eukaryotic protein elongation factors. *Trends in Biochemical Sciences,* 15, 11, pp. 420-424.

Riu, E. R., Chen, Z. Y., Xu, H., He, C. Y., & Kay, M. A. (2007). Histone modifications are associated with the persistence or silencing of vector-mediated transgene expression *in vivo*. *Molecular Therapy,* 15, 7, pp. 1348-1355.

Rosano, G. L., & Ceccarelli, E. A. (2009). Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. *Microbial Cell Factories,* 8, 1, pp. 41.

Rosenberg, M., & Court, D. (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. *Annual Review of Genetics,* 13, 1, pp. 319-353.

Schreiber, S. L. (2005). Small molecules: The missing link in the central dogma. *Nature Chemical Biology,* 1, 2, pp. 64-66.

So, A. G., & Downey, K. M. (1992). Eukaryotic DNA replication. *Critical Reviews in Biochemistry and Molecular Biology,* 27, 1-2, pp. 129-155.

Szymczak, A. L., & Vignali, D. A. A. (2005). Development of 2A peptide-based strategies in the design of multicistronic vectors. *Expert Opinion on Biological Therapy,* 5, 5, pp. 627-638.

Wahle, E. (1995). 3'-End cleavage and polyadenylation of mRNA precursors. *Biochimica et Biophysica Acta - Gene Structure and Expression,* 1261, 2, pp. 183-194.

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M., & Losick, R. (2008). *Molecular Biology of the Gene* (6 ed.). Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Williams, S., Mustoe, T., Mulcahy, T., Griffiths, M., Simpson, D., Antoniou, M., et al. (2005). CpG-island fragments from the *HNRPA2B1/CBX3* genomic locus reduce silencing and enhance transgene expression from the hCMV promoter/enhancer in mammalian cells. *BMC Biotechnology,* 5, 1, pp. 17.

Wilson, G. G., & Murray, N. E. (1991). Restriction and modification systems. *Annual Review of Genetics,* 25, 1, pp. 585-627.

Wu, X. Q., Jornvall, H., Berndt, K. D., & Oppermann, U. (2004). Codon optimization reveals critical factors for high level expression of two rare codon genes in *Escherichia coli*: RNA stability and secondary structure but not tRNA abundance. *Biochemical and Biophysical Research Communications,* 313, 1, pp. 89-96.

Yew, N. S., Wysokenski, D. M., Wang, K. X., Ziegler, R. J., Marshall, J., McNeilly, D., et al. (1997). Optimization of plasmid vectors for high-level expression in lung epithelial cells. *Human Gene Therapy,* 8, 5, pp. 575-584.

Zhang, G., Hubalewska, M., & Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nature Structural & Molecular Biology,* 16, 3, pp. 274-280.

Zolotukhin, S., Potter, M., Hauswirth, W., Guy, J., & Muzyczka, N. (1996). A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *Journal of Virology,* 70, 7, pp. 4646-4654.

Zur Megede, J., Chen, M. C., Doe, B., Schaefer, M., Greer, C. E., Selby, M., et al. (2000). Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 *gag* gene. *Journal of Virology,* 74, 6, pp. 2628.

Zvereva, M., Shcherbakova, D., & Dontsova, O. (2010). Telomerase: Structure, functions, and activity regulation. *Biochemistry,* 73, 13, pp. 1563-1583.