Promoter analysis by saturation mutagenesis

Nitin S. Baliga[1*]

[1]Institute for Systems Biology, 4225 Roosevelt way NE, Suite 200, Seattle, WA 98105

*To whom correspondence should be addressed: Baliga, NS. Institute for Systems Biology, 4225 Roosevelt way NE, Suite 200, Seattle, WA 98105. Tel: 206-732-1266; Fax: 206-732-1299; Email: nbaliga@systemsbiology.org

Abbreviations: *bop*: bacterioopsin

ABSTRACT

Gene expression and regulation are mediated by DNA sequences, in most instances, directly upstream to the coding sequences by recruiting transcription factors, regulators, and a RNA polymerase in a spatially defined fashion. Few nucleotides within a promoter make contact with the bound proteins. The minimal set of nucleotides that can recruit a protein factor is called a cis-acting element. This article addresses a powerful mutagenesis strategy that can be employed to define cis-acting elements at a molecular level. Technical details including primer design, saturation mutagenesis, construction of promoter libraries, phenotypic analysis, data analysis, and interpretation are discussed.

INTRODUCTION

Saturation mutagenesis is a method implemented to generate a library of mutations within a targeted DNA sequence. This approach allows for rapid and unbiased identification and functional evaluation of cis-acting element(s) within a promoter.

A promoter by definition is the DNA sequence, generally directly upstream to the coding sequence, required for basal and/or regulated transcription of a gene. However, only a few nucleotides within a promoter are absolutely necessary for its function. Saturation mutagenesis is a simple method that can be used to scan a promoter for cis-acting elements and subsequently define the critical nucleotides within and the consensus sequence for these elements. This method of analysis is most effective when the minimal DNA sequence constituting the promoter is experimentally defined (Fig. 1) (7, 13). Therefore, some groundwork is essential prior to applying saturation mutagenesis to address a scientific problem. This article focuses on the strategy and use of saturation mutagenesis to identify and derive consensus sequence(s) for
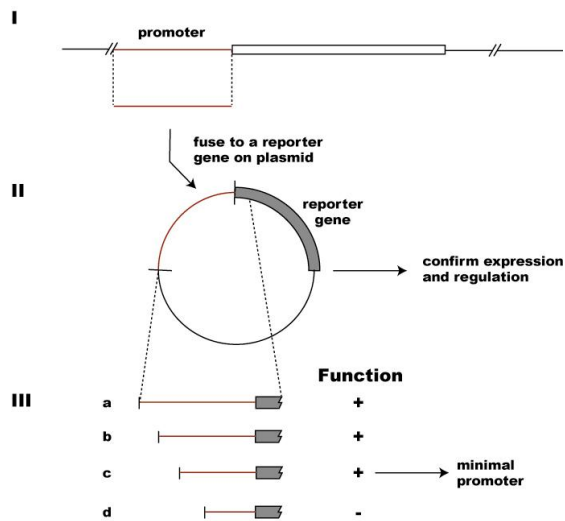
cis-acting-element(s).

Fig. 1: Identification of the minimal promoter. I & II, The promoter is spliced to a reporter gene on a plasmid. III Nested deletions with exonuclease III coupled to functional analysis can be used to identify the minimal sequence that comprises a functional promoter.

METHODS AND DISCUSSION

Saturation mutagenesis

Reporter systems
Simultaneous saturation mutagenesis of multiple nucleotides can yield an astronomical number of mutants making the analysis tedious. Therefore, analysis of the mutated-promoter libraries is significantly facilitated if functional sequences yield a scorable phenotype. For example, since the bacterioopsin (*bop*) gene product provides a colorimetric screen, for mutagenesis of the *bop* promoter the natural function of the

gene was exploited for screening functional sequences (Fig. 2). Fusion of a reporter gene to the minimal promoter can also facilitate this process by allowing for high-throughput selection or screening of the library mutants prior to biochemical analysis (Fig. 1). Table 1 lists a few genes that have been successfully used as reporters in saturation mutagenesis analyses. Direct transcription analysis can also be used when the number of mutants is not too large (9, 11, 12). [The section on mutant analysis details the method that was used to characterize the *bop* promoter using a colorimetric screen.]
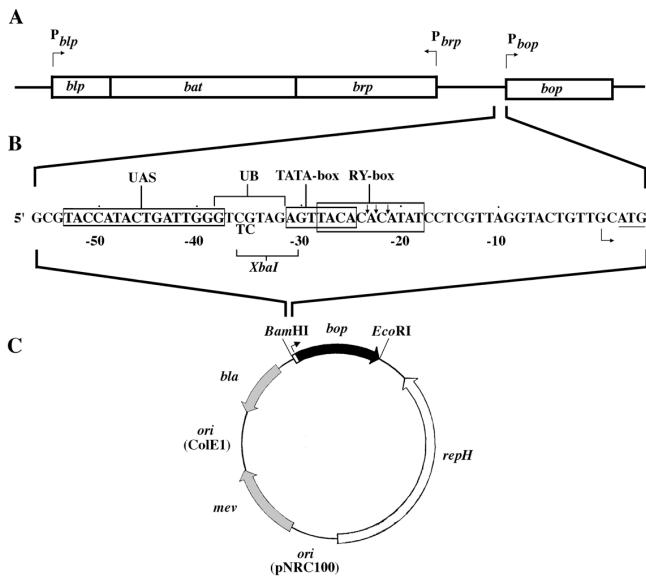

Fig. 2: Map of the *bop* gene cluster, the minimal *bop* promoter and promoter mutagenesis. (A) The relative sizes and arrangement of the 4 genes, *brp*, *bat*, *blp*, and *bop*, in a gene cluster on the chromosome is shown in boxes with position and orientation of the promoters (indicated by arrows above). (B) The sequence of nucleotides within the minimal *bop* promoter and a few surrounding nucleotides with the TATA box, UAS, and RY box, boxed. The start codon is underlined and the transcription start site (indicated by an arrow) is numbered +1. (C) *E.coli-Halobacterium* shuttle plasmid (pNB series) with the cloned *bop* gene (black arrow) and promoter (white box). The plasmid contains origins of replication for *E.coli* (ori ColE1) and *Halobacterium* (ori pNRC100, *repH*) in addition to selectable markers (*bla* for Amp^r in *E.coli*, and *mev* for Mev^r in *Halobacterium* [shaded arrows]).

The reporter gene downstream to the un-mutated wild type promoter should serve as a control to account for anomalous results attributable to mutations outside the targeted region or the altered behavior, if any, of the promoter on a plasmid backbone.

Oligonucleotide design

The primer used for mutagenesis should have clamps of at least 15-bp on either side of the randomized sequence to ensure high annealing specificity (for e.g the primer used for saturation mutagenesis of 7 bp of the *bop* gene TATA box had 22 and 24 bp clamps on either side). Use of degenerate primers with smaller clamps may result in non-specific amplification or frameshift mutations.

Table 1. Reporter genes used for promoter analysis

Gene	Function	Reference
bop	purple/orange screen	Baliga and DasSarma, 1999, 2000 (1, 2)
dhfr	trimethoprim resistance	Danner and Soppa, 1996 (8)
cat	chloramphenicol resistance	Colgan and Manley, 1995 (6)
bgaH	β-Galactosidase	Patenge *et al*., 2000 (10)

The number of nucleotides mutagenized has a direct effect on the level of statistical significance at which the promoter elements are defined. DNA transformation efficiencies of the host strains used for library construction and analysis also factor into this equation. Maximal representation for all mutations can be achieved with a library population that exceeds the total number of mutants expected by at least 5-fold; for e.g. saturation mutagenesis of a 7 bp stretch yields 16,384 mutants, hence >80,000 clones are required to constitute a good library for saturation mutagenesis of a 7-bp region (Fig. 3). Furthermore, secondary mutations at non-critical nucleotides may compensate for deleterious effects of mutations at a critical nucleotide position(s). The consensus sequence derived from simultaneous mutagenesis of too long a stretch of DNA will be loosely defined with a greater abundance R, W, S, Y, etc. instead of G, A, T, and C. In other words, to have maximal representation for all mutations and to avoid interference by secondary mutations, the length of DNA mutagenized should be kept at an optimal minimum.

Mutagenesis by PCR

The mutagenesis is conducted by amplifying the promoter with the degenerate oligonucleotide and a second downstream non-mutagenic oligonucleotide. The promoter can be amplified with or without the reporter gene, in case of the latter the promoter is spliced back into the promoter-reporter plasmid construct. The mutated-promoter PCR product should be gel-purified from unincorporated nucleotides, primers, and primer-dimers, which may interfere with downstream processing and cloning.

Nucleotides within long promoters can be mutagenized by mega-primer or recombinant PCR techniques (3, 4, 5). Alternately, restriction sites can be engineered into the PCR product to splice it back into the promoter-reporter plasmid construct.

A stretch of 28 nucleotides can be scanned for cis-acting elements by mutagenesis in four separate PCR reactions using four oligonucleotides, each targeting a different set of seven nucleotides, and the same downstream oligonucleotide (Fig 3).

Cloning of the PCR product: library construction

The mutated-promoter library should be constructed in an appropriate vector if phenotypic analysis is desired. The PCR product can be used directly for cloning as a blunt-ended insert using the Sureclone ligation kit (Amersham Pharmacia Biotech) or digested with restriction enzymes for directional cloning (if unique restriction sites are incorporated at the 5' ends of the primers).
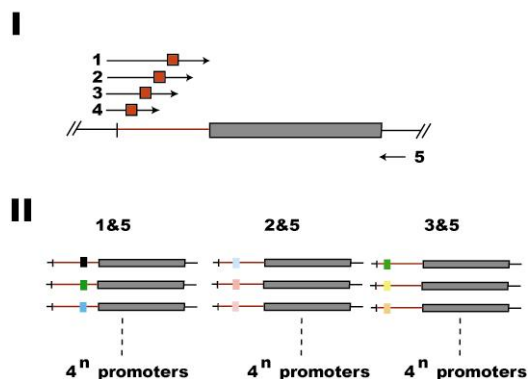

Fig. 3: Promoter scanning by saturation mutagenesis. I Tandem oligonucleotides with randomized bases are designed to span the region of interest. II PCR amplification with a primer pair (of which one is mutagenic) yields a collection of copies of the gene downstream to an equally diverse set of promoter sequences. The number of mutated promoters is equal to the nth exponent of 4 (where n is the number of nucleotides mutated and the randomized base is N = G, A, T, C).

For most restriction enzymes it may be necessary to include additional bases 5' to the restriction site to improve the efficiency of cutting (see NEB technical resource for details).
To construct the library electroporate 0.5 µl of the ligation into 20 µl of Electromax DH10B cells (Invitrogen Lifetechnologies Inc.). Repeat the procedure for the entire ligation mix (20 electroporations for a 10 µl ligation reaction). Plate 50 µl of one transformation to estimate the size of the library and inoculate 1L LB + 100µg/ml ampicillin with the rest of the transformations (after the 1 hr incubation step). Culture overnight at 37°C, prepare plasmid DNA: this is your library of mutated-promoter plasmids. Sequence plasmid DNA prepared from randomly selected transformants from the efficiency plate for assessing the randomness of the library (see below).

Confirming the randomness of the library.

It is essential to rule out non-uniform representation of all four nucleotides in the mutagenized region in the library that may arise from a bias in (i) the coupling efficiency during synthesis

of the degenerate oligonucleotide or (ii) priming efficiency of the mutagenic primer on the wild-type DNA template. The mutagenized region should be sequenced from a random set of library plasmids and compared to sequences post-selection/screening. In absence of screening or selection the library should have an even distribution of nucleotides at each position (Fig. 4). A distinct bias in nucleotide distribution post-selection/screening is indicative of a requirement for specific nucleotides for promoter function.

Analysis of mutants

Phenotypic analysis

Phenotypic analysis allows for a quick estimation of the number of functionally critical nucleotides within the mutagenized region. The mutated nucleotides are not important for promoter function if all transformants have functional phenotypes. On the other hand, presence of only 25% phenotypically positive transformants is indicative of one highly conserved nucleotide position. The number of active promoters decreases exponentially with mutations in each additional critical nucleotide.

A

	-31	-30	-29	-28	-27	-26	-25
1	G	A	T	C	C	C	T
2	T	T	G	C	T	A	G
3	C	A	C	G	G	G	T
4	T	C	G	C	T	T	G
5	A	G	G	G	A	T	T
6	C	G	G	A	A	A	T
7	A	T	T	T	C	C	C

B

	-31	-30	-29	-28	-27	-26	-25	Total
G	1	2	4	2	1	1	2	13
A	2	2	0	1	2	2	0	9
T	2	2	2	1	2	2	4	15
C	2	1	1	3	2	2	1	12

C

	-31	-30	-29	-28	-27	-26	-25	Total
G	5	1	0	0	1	0	2	9
A	5	3	2	1	3	3	8	25
T	2	6	6	13	10	10	1	48
C	2	4	6	0	0	1	3	16
	n	t	y	T	T/a	T	a	

Fig. 4: Sequence analysis of pNBTATA library plasmids from unscreened (E. coli transformants) vs. screened (*Halobacterium* SD23 transformants) colonies. A. Tabulation of nucleotide position and identity of the TATA box in seven plasmids prepared from E. coli transformants before the purple/orange screening step. B. Tally of nucleotides (G, A, T, and C) at the seven positions in the TATA box. C. Tally of nucleotides in the TATA box from mutated but active promoters identified through a purple/orange screen.TATA box from mutated but active promoters identified through a purple/orange screen.

The phenotypes should be classified into "negative", "weak", "moderate", and "strong" to attribute relative importance to conserved nucleotides within functional promoters.

Colorimetric analysis of mutants from the *bop* promoter mutagenesis allowed for a similar classification of promoter strengths.
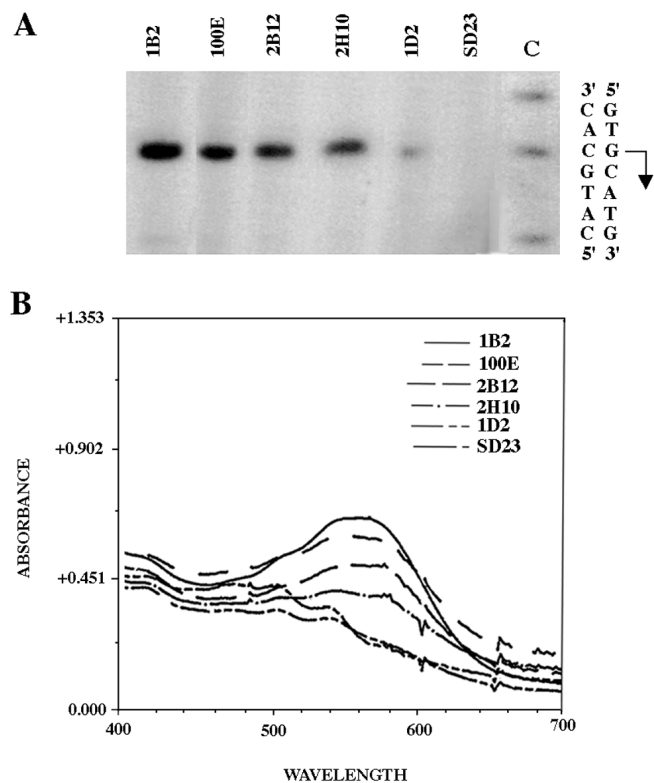

Fig. 5: Analysis of TATA box mutants at the transcriptional (A) and translational (B) levels. (A) Primer extension analysis of *bop* mRNA in 4 *bop* promoter TATA-box mutants using crude RNA is shown with strain designations labeled above. The controls used in this experiment were the Pum⁻ (*bop*⁻) host strain SD23, strain 100E (SD23 containing plasmid with unmutated promoter) and TATA box mutants (1B2, 2B12, 2H10 and 1D2). A sequencing reaction (C lane) performed with the same primer on pMS1 template (plasmid containing the cloned *bop* gene) is shown as is the double stranded sequence across the transcription start point (start site and direction indicated by arrow). (B) Spectra (absorbance versus wavelength from 400 to 700 nm) of purple membrane preparations for the 4 TATA-box mutant strains and the 2 controls strains (same as in part A) are shown. Relative BR content was quantified from comparison of absorbance at 568 nm to that at 280 nm.

The phenotypes in this screen were subject to an arbitrary classification of orange (Pum⁻), weak purple (Pum^{+/-}), purple (Pum⁺), and intense-purple (Pum^{++}) (1, 2).

Danner and Soppa (8) utilized trimethoprim resistance as a method for classification of promoter strength by culturing a randomly selected set of mutants in growth medium supplemented with various dilutions of the drug. The classification in this case was in terms of minimum inhibitory concentrations (MIC); MIC<1µgml-1 (sensitive), MIC 5-200 µgml-1 (partially resistant), MIC >400µgml-1 (very resistant).

Transcription analysis

Primer extension analysis is routinely used for expression profiling (Fig. 5) (1, 2). This method has advantages over other profiling methods like RT-PCR, and Northern blotting in that it is easy, quantitative, and allows for mapping of the transcription start site. It is essential to map the start site since the mutagenesis might lead to activation of secondary promoters which can interfere in the final analysis. Primer extension kits are commercially available from Promega Corporation. Primer extension with an oligonucleotide specific for a constitutively expressed gene, for e.g. 16S rRNA, should be included as an internal control in each primer extension reaction to normalize both priming efficiency and message levels. The message levels can be quantified by densitometry of autoradiograms or phosphoimager analysis.

Effects of mutagenesis on transcription.

A 1:1 (or direct) correspondence of phenotype to transcript levels confirms that the effect of the mutagenesis is at the level of transcription (Fig. 5) (1, 2). An absence of correlation may be observed if additional mutations exist within the coding sequence or if the targeted region is within the transcript (translational effect). Therefore it is essential to map the transcription start site of the gene and identify the minimal promoter prior to designing the oligonucleotide for mutagenesis. Deletion or insertion mutations were rarely identified during saturation mutagenesis of the bop promoter. The rare instance when a single base deletion was identified resulted in a corresponding shift in the transcription start site (Baliga and DasSarma, unpublished). Therefore, I strongly recommend using primer extension analysis as means to quantify expression, since it also maps the transcription start site. Clearly, such mutants should be excluded from the analysis. Moreover, though the location of the TATA box is relatively flexible between promoters, for a given promoter it is centered at a relatively fixed position from the transcription initiation site. Therefore, saturation mutagenesis has been successfully used to characterize TATA box consensus sequences (1, 8).

DNA sequence analysis to identify mutations

The DNA sequence in the promoter regions is sequenced and aligned with the wild-type promoter sequence to identify the nature and position(s) of mutations. Sequencing of both strands is essential to confirm the result. Sequencing reads of 400-500bp on both strands should be analyzed to confirm lack of mutations at positions external to the mutagenized region.

Methods to derive the consensus sequence

The consensus sequence represents conserved nucleotides in active promoter sequences. The importance of a nucleotide in the function of the promoters is reflected in the extent to which mutations are tolerated at the nucleotide position.

A

Strain	\-31	\-30	\-29	\-28	\-27	\-26	\-25	Phenotype Pum
	Nucleotide Position							
S9	A	G	T	T	A	C	A	++
100E	A	G	T	T	A	C	A	++
1B2	G	T	.	.	T	T	.	++
2H1	G	C	.	.	T	T	C	++
2B2	.	T	C	.	T	T	.	++
2B12	A	T	A	.	T	T	C	++
ZA1	T	A	A	.	.	T	G	++
1A7	A	A	C	.	T	T	.	++
1B11	T	C	C	.	T	T	.	++
1H7	G	T	.	.	T	T	T	++
2A3	C	T	C	.	.	A	.	++
2H10	.	T	C	.	T	T	C	++
1D1	G	C	.	.	G	A	.	+
1D2	G	A	C	.	T	A	.	+
1H10	C	C	.	A	T	T	C	+
1A1	C	G	.	G	G	T	T	-
2D6	T	A	A	A	C	A	.	-
ZF6	.	T	G	G	C	G	G	-
ZF8	T	G	G	.	G	T	.	-

B

	\-31	\-30	\-29	\-28	\-27	\-26	\-25
G	5	1	0	0	1	0	2
A	5	3	2	1	3	3	8
T	2	6	6	13	10	10	1
C	2	4	6	0	1	1	3
	r	t	Y	T	T$_{/a}$	T	a

Fig. 6: Tabulation of strain designations, promoter sequences, phenotypes, and bacteriorhodopsin content of TATA-box mutants (A) and analysis of the promoter sequences (B). (A) The strain designations are shown in the first column, and sequence in the –31 to –25 nucleotide region (identity to wild type base denoted by a dot), Pum phenotype (-/+/++). (B) The consensus sequence for the TATA-box was determined by counting the appearance of individual nucleotides at each position in Pum[+] strains. The consensus sequence is indicated at the bottom with numbers in subscript referring to the percentage of times the most common base(s) were found at that position.

Algorithms for generating consensus sequences are commercially available, for e.g. the program CONSENSUS of the GCG sequence analysis suite, accepts aligned sequences as input and derives a consensus sequence at a confidence level defined by the user. The consensus sequence can also be derived manually by tabulating a tally of nucleotides at each position (Fig. 6). If the desired confidence level for the consensus is 75%, i.e. a match with the derived consensus has a 3 in 4 chance of being a functional promoter, then at any given position the nucleotide represented in 75% or greater active promoters is given the consensus status. Nucleotides highly conserved in all promoters are required for transcription, whereas those loosely-conserved in strong promoters but not in the weak promoters are preferred but not absolutely essential for a functional promoter. Nucleotides conserved in most strong promoters are displayed in upper case and those loosely conserved are displayed in lower case.

REFERENCES

1. Baliga NS, DasSarma S. Saturation mutagenesis of the TATA box and upstream activator sequence in the haloarchaeal *bop* gene promoter. *J Bacteriol.* 1999;181(8):2513-2518.
2. Baliga NS, Dassarma S. Saturation mutagenesis of the haloarchaeal *bop* gene promoter: identification of DNA supercoiling sensitivity sites and absence of TFB recognition element and UAS enhancer activity. *Mol Microbiol.* 2000;36(5):1175-1183.
3. Barik S, Galinski MS. "Megaprimer" method of PCR: increased template concentration improves yield. *Biotechniques.* 1991;10(4):489-490.
4. Blattner FR. Direct amplification of the entire ITS region from poorly preserved plant material using recombinant PCR. *Biotechniques.* 1999;27(6):1180-1186.
5. Chrzanowska-Lightowlers ZM, Temperley RJ, McGregor A, Bindoff LA, Lightowlers RN. Conversion of a reporter gene for mitochondrial gene expression using iterative mega-prime PCR. *Gene.* 1999;230(2):241-247.
6. Colgan J, Manley JL. Cooperation between core promoter elements influences transcriptional activity in vivo. *Proc Natl Acad Sci U S A.* 1995;92(6):1955-1959.
7. Corona V, Aracri B, Kosturkova G, et al. Regulation of a carotenoid biosynthesis gene promoter during plant development. *Plant J.* 1996;9(4):505-512.
8. Danner S, Soppa J. Characterization of the distal promoter element of halobacteria in vivo using saturation mutagenesis and selection. *Mol Microbiol.* 1996;19(6):1265-1276.
9. Hain J, Reiter WD, Hudepohl U, Zillig W. Elements of an archaeal promoter defined by mutational analysis. *Nucleic Acids Res.* 1992;20(20):5423-5428.
10. Patenge N, Haase A, Bolhuis H, Oesterhelt D. The gene for a halophilic beta-galactosidase (*bgaH*) of *Haloferax alicantei* as a reporter gene for promoter analyses in *Halobacterium* salinarum. *Mol Microbiol.* 2000;36(1):105-113.
11. Palmer JR, Daniels CJ. In vivo definition of an archaeal promoter. *J Bacteriol.* 1995;177(7):1844-1849.
12. Reiter WD, Hudepohl U, Zillig W. Mutational analysis of an archaebacterial promoter: essential role of a TATA box for transcription efficiency and start-site selection in vitro. *Proc Natl Acad Sci U S A.* 1990;87(24):9509-9513.

13. Yang CF, Kim JM, Molinari E, DasSarma S. Genetic and topological analyses of the *bop* promoter of *Halobacterium halobium*: stimulation by DNA supercoiling and non-B-DNA structure. *J Bacteriol.* 1996;178(3):840-845.

ACKNOWLEDGEMENTS

PROTOCOL

1. Splice promoter to reporter gene
2. Conduct nested deletions to identify minimal promoter
3. Design oligonucleotide: replace sequence for targeted region with "N"s; flank mutations with at least 15 bp clamps.
4. Amplify reporter gene and promoter with mutagenic primer and downstream primer.
5. Clone PCR product into plasmid vector to construct a library.
6. Sequence a random set of plasmids in the mutated region to rule out sequence bias in promoter mutagenesis.
7. Screen library for non-functional, weak, moderate, and strong promoters.
8. Isolate plasmids and retransform to confirm phenotype.
9. Profile expression by primer extension.
10. Sequence and align promoter sequences.
11. Tally number of nucleotides (G, A, T, and C) at each position.
12. Derive consensus.