# Genetic Assembly, Error-Correction and a High-Throughput Screening Strategy for Protein Expression Optimization

by

Jiayuan Quan

Department of Biomedical Engineering
Duke University

Date:
Approved:
Jingdong Tian, Supervisor
Lingchong You
Uwe Ohler
Kent Weinhold

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering in the Graduate School of Duke University

#### **ABSTRACT**

Gene Synthesis, Error Correction and a High-Throughput Screening Strategy for Protein

**Expression Optimization** 

by

Jiayuan Quan

Department of Biomedical Engineering
Duke University

Date:
Approved:
11
Jingdong Tian, Supervisor
Lingchong You
Uwe Ohler
Kent Weinhold

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biomedical Engineering in the Graduate School of Duke University

Copyright by Jiayuan Quan 2012

#### **Abstract**

Various types of genetic constructs are widely used as diagnostic, prophylactic, and therapeutic tools for human diseases. They are also the workhorse in the biotech and pharmaceutical industry for production of therapeutic antibodies and proteins.

Since the majority of the genetic constructs encode protein products, it is therefore of tremendous value to human health and the society that we could find a way to fine-tune and optimize genetic constructs and hence protein expression for achieving maximal potency or long-lasting effects in therapeutics or for obtaining the highest yields in pharmaceutical protein production. However, for protein-coding genes to be expressed in a heterologous host, the coding sequences need to be optimized by using synonymous codons to achieve reasonable levels of expression, if at all. Since codon optimization is done in a protein-by-protein basis with respect to specific host organisms, tissue/cell types, even health conditions, and there is no set of standard rules to follow, this process is still very unpredictable and time-consuming.

This thesis presents the development of a feasible platform for solving the problem of optimizing regular and long DNA constructs for academic or industrial purposes through the development of a novel cloning method for complex gene libraries, and based on the library expression system constructed in such manner, a platform for high-throughput screening of codon-optimized and error-corrected proteins, and a novel protocol for screening long gene constructs which could be

extremely difficult to achieve by using regular screening methods. This multi-step platform has the potential for studying the natural systems: how codon bias correlates to protein expression efficiency, for generating improved pharmaceutical proteins and enhanced DNA vaccines and for constructing improved genome libraries.

Keywords: synthetic biology, molecular cloning, genetic assembly, gene design, synonymous codon, protein optimization, error-correction

# **Dedication**

I would like to dedicate this thesis to my husband, Rui Zhang and my parents, Huitao Quan and Yulan Jia.

# **Contents**

Abstractiv
List of Tablesxi
List of Figuresxii
Acknowledgementsxix
1. Introduction
1.1 Molecular cloning technologies
1.2 Error correction technologies
1.3 Genetic manipulation of protein expression5
1.4 Codon usage and protein expression regulation
2. A novel method for cloning of gene libraries and multi-DNA fragment assembly11
2.1 Introduction
2.2 Materials and methods
2.2.1 Reagents and equipments15
2.2.2 Reagent setup
2.2.3 Design of overlapping sequences between vector and insert(s)
2.2.4 Preparation of linear vector
2.2.5 Preparation of inserts21
2.2.6 CPEC cloning
2.2.7 Multi-cycle CPEC for library cloning
2.2.8 Combinatorial library cloning
2.2.0 Multi way CPEC

2.2.10 CPEC product gel electrophoresis	24
2.3 Results	24
2.3.1 CPEC cloning of a single gene	24
2.3.2 CPEC cloning of a gene library	26
2.3.3 CPEC cloning of CPEC cloning of combinatorial libraries	29
2.3.4 CPEC cloning of multi-component pathway	33
2.4 Discussions and conclusions	35
3. Error-correction in gene synthesis and optimization technology	40
3.1 Introduction	40
3.2 Materials and methods	40
3.2.1 Reagents	40
3.2.2 ECR of assembled genes	41
3.2.3 Cloning, sequencing and functional analysis of synthetic genes	42
3.3 Results	43
3.3.1 General design of the ECR using Surveyor nuclease	43
3.3.2 Determine error frequency of on-chip gene synthesis	45
3.3.3 Reduction of error frequencies after ECR	48
3.4 Discussions and conclusions	54
4. High-throughput protein expression screen strategy	56
4.1 Introduction	56
4.2 Materials and methods	60
4.2.1 Oligonucleotide synthesis	60

	ssembly	61
4.2	2.3 Enzymatic error correction	62
4.2	2.4 Construction of a synthetic $lacZlpha$ gene library	62
4.2	2.5 Evaluation of the expression levels of synthetic $lacZ\alpha$ gene library	63
4.2	2.6 Plasmid library construction using CPEC method	63
4.2	2.7 Protein expression screen	64
4.2	2.8 Cleavage and purification of transcription factor-GFP fusion proteins	65
4.3	Results	66
4.4	Discussions and conclusions	81
5. Scree	ening and optimization of synthetic long genes	82
5.1	Introduction	82
5.2	Materials and methods	83
5.2	2.1 Reagents and primers	83
5.2	2.2 Oligonucleotides and genes	84
5.2	2.3 Polymerase cycling assembly (PCA)	84
5.2	2.4 Polymerase chain reaction (PCR)	85
5.2	2.5 Plasmid library construction using CPEC method	85
5.2	2.6 Plasmid PCR	86
5.2	2.7 Sequencing preparation and sequencing	86
5.3	Results	87
5.4	Discussions and conclusions	106
Annen	dix A	111

Supplementary sequences	111
Chapter 2	111
Chapter 3	112
Chapter 4	118
1. Test gene constructs:	118
2. Drosophila transcription factor gene fragments (wild-type):	119
3. Drosophila transcription factor gene fragments (expression optimized):	134
4. Chip oligonucleotide sequences	148
References	152
Biography	158

# **List of Tables**

Table 1: Error analysis of synthetic gene sequences before and after ECR with Surveyor nuclease
Table 2: Error frequencies as determined by sequencing in chip-synthesized RFP genes with and without error correction using Surveyor nuclease. Clones were randomly selected from each population and sequenced from both directions
Table 3: Comparison of CAI values of 15 expression optimized TF sequences (CAI-opt) vs. wild-type non-expressing sequences. CAI value was calculated using CAIcal server at http://genomes.urv.es/CAIcal [84]
Table 4: Comparison of CAI values of the rest 59 expression optimized TF sequences (CAI-opt) vs. wild-type sequences

### **List of Figures**

Figure 2: Schematic diagram of CPEC cloning of combinatorial gene libraries. Two gene libraries are cloned in frame into a vector. The vector and the inserts share overlapping regions at the ends. In each CPEC cycle, after denaturation and annealing, the hybridized inserts and vector extend using each other as a template until they complete a full circle. The assembled plasmid library can be directly used for transformation into competent cells.

 Figure 8: Assembly of multi-component pathway using CPEC. (A) A schematic diagram of the multi-way CPEC. Any two neighboring fragments share an overlapping region with identical Tm. Multiple cycles are usually needed to drive the reaction to completion. The positions of the two nicks (arrow head) in the final completely assembled plasmid may vary depending on the number, lengths, and sequences of the fragments. (B) Gel electrophoresis analysis of the final assembly product after a 20-cycle CPEC. 5 ml of the reaction was separated on a 0.8% agarose gel and visualized after

Figure 11: Employing error correction increases the percentage of fluorescent RFP colonies. Images below are examples of the increased fluorescent population derived by employing ECR with Surveyor nuclease. Image A1/A show colonies derived from the uncorrected RFP synthesis product that only yields 50.2% fluorescing colonies. Iterations of correction, using 20 min incubations in this case, yield colonies with an increased fluorescent population as shown in B1/B and C1/C. Images A1-C1 are the same images as A-C with an added pseudo-colored red mask to highlight the brightly fluorescent colonies.

Figure 12: Predicted effects of ECR as a function of sequence length. (A) Purity of gene synthesis products (percentage of error-free clones) decreases exponentially with the length of the product synthesized. Employing ECR (1 error in 8701 bp, blue line) dramatically increases the probability of locating an error-free clone than the uncorrected population (1 error in 526 bp, red line). (B) Employing ECR significantly reduces the number of colonies that need to be screened to have a high (95%) probability

with Surveyor (blue line) could yield a correct 10 kb product by sequencing eight random clones. Plots are derived from the result of model calculations as describe the text.	ed in
Figure 13: The expression system map of pAcGFP1-TF. Inserts were subcloned into 5'MCS and fused in frame to the N-terminum of AcGFP1	
Figure 14: The integrated on-chip oligo array synthesis, amplification and gene as process. Small pools of oligos are synthesized in separate chambers on a plastic D microchip using an inkjet DNA microarray synthesizer. The chambers are then fill with a combined amplification and assembly reaction mixture and sealed. In a nic and strand displacement amplification reaction, a DNA polymerase (Bst large frag shown in yellow) extends and displaces the proceeding strand while a nicking endonuclease (Nt.BstNBI, shown in teal) separates the construction oligos from the universal primer (in red) and generates new 3' -ends for extension. After amplification free oligos in each chamber are assembled into gene products by polymerase cassembly.	NA led cking gment, ne cation, chain
Figure 15: Assembly of target genes by on-chip nSDA-PCA reactions. (a) Agarose image of the nSDA-PCA reaction product showing as a typical smear (left lane) at PCR amplified $lacZ\alpha$ gene product (right lane). The middle lane is 100-bp DNA la (b) Assembly of the Red Fluorescent Protein (RFP) gene by on-chip nSDA-PCA re followed by PCR amplification.	nd the adder.
Figure 16: Statistical evaluation of errors in synthetic RFP genes with and without Surveyor nuclease treatment. (a) Percentage of fluorescent RFP colonies with and without error correction using Surveyor nuclease. On-chip RFP gene synthesis wi error correction resulted in 50.2% fluorescent colonies while those treated with the Surveyor nuclease yielded a 84% fluorescent population. The total number of color each population was approximately 3,000. (b) Predicted correlation of the probability an error-free clone with product length [61] before and after error correction with Surveyor nuclease. Error correction using Surveyor nuclease (error frequency fc = per kb, blue line) increases the probability of locating an error-free clone than the uncorrected population (error frequency fu = 1.9 per kb, red line), thereby drastical reducing the number of colonies that need to be screened. The error frequencies we calculated from sequencing data in Table 2.	thout e onies in ility for : 0.19 illy vere
Figure 17: Expression of synthetic $lacZ\alpha$ codon variants in <i>E. coli</i> . (a) A set of 1,296	E. coli

colonies expressing distinct  $lacZ\alpha$  codon variants sorted by color intensity. Raw images

xv

C7180 Flatbed Scanner. (b) Bar graph and box plot showing distribution of color
intensities of a different set of 1,468 random colonies expressing distinct $lacZ\alpha$ codon
variants on an agar plate. Owing to the large size of the synthetic codon variant library,
the chance of having identical clones on a plate was extremely low, as confirmed by
sequencing several hundred blue colonies (data not shown). In the box plot, the
expression level of the WT $lacZ\alpha$ is marked with a dash line

Figure 21: A schematic diagram of the proposed long gene screening mechanism. In the first step, each gene fragment library constructed from degenerate oligo libraries was cloned into vector. In the second step, the cloning product was transformed into *E. coli* competent cells and screened for GFP signal representing highest protein expression. In the third step, 8-10 clones with the highest expression were collected for DNA plasmids. In the fourth step, the plasmid mixture for each highest expression gene fragment were amplified by plasmid PCR using one normal non-degenerate primer (in black) and one degenerate connector primer (the part that hybridized with the other connector was

shown in green; the part that hybridized with the gene library was shown in red/blue). In the fifth step, amplified enriched fragment library was used together with two end primers to assemble the full-length enriched gene library. In the sixth step, the amplified full-length enriched gene library was cloned into the vector. In the seventh step, the cloning product was transformed into <i>E. coli</i> competent cells for screening. In the eighth step, a few clones with the highest GFP signal and hence protein expression were selected.
Figure 22: PCA result for MTggps, SAggps, AFggps both fragment libraries. 100 ng, 200 ng, and 400 ng purified oligo mixture were used respectively in Lane 1, 2 and 3 for each gene fragment. 1 kb and 100 bp DNA ladder were used respectively93
Figure 23: PCR result for MTggps, SAggps, AFggps and cat2 both fragment libraries. Red arrows indicated the formation of correct product
Figure 24: CPEC cloning gel electrophoresis result for MTggps, SAggps, AFggps and cat2 both fragments. Arrows indicated the correct CPEC product. The band below the product was the exessive vector.
Figure 25: cat2-f1 plasmid PCR result using different plasmid mixture concentrations and degenerate connector primer concentrations. Row 1-3: 1 ng, 5 ng, 12.5 ng, 25 ng, 50 ng and 100 ng plasmid mixture with 0.5 $\mu$ M/ 2.5 $\mu$ M/ 5 $\mu$ M/ 10 $\mu$ M of degenerate connector cat2-16co3, respectively; 1 kb DNA ladder (NEB); 100 bp DNA ladder (NEB).
Figure 26: cat2-f2 plasmid PCR result using different plasmid mixture concentrations and degenerate connector concentrations. From top left to lower right lane: 1 ng , 5 ng, 12.5 ng and 25 ng plasmid mixture with 0.5 $\mu$ M/ 2.5 $\mu$ M/ 5 $\mu$ M/ 10 $\mu$ M of degenerate connector cat2-17co; 100 bp DNA ladder (NEB)
Figure 27: Plasmid PCR for MTggps, SAggps and AFggps using 0.5 µM connector primer. Slight different set of plasmid mixture concentrations were used for different genes. 100 bp DNA ladder in the middle (NEB).
Figure 28: Full gene assembly result for cat2, MTggps, SAggps and AFggps using different plasmid PCR concentrations. Gene assembly product was indicated by red arrows. 1 kb DNA ladder (NEB)
Figure 29: cat2, MTggps, SAggps and AFggps full gene library CPEC cloning result. Red arrows indicated the plasmid formed with gene library inserted. 1 kb DNA ladder

Figure 30: scPCR result for cat, MTggps, SAggpa and AFggps full gene libraries. Clon-	es
picked with the highest fluorescence were analyzed with scPCR to amplify the insert	
region. 1kb DNA ladder (NEB) was used for all rows.	106

#### Acknowledgements

I feel honored to be at this point to acknowledge all of the people: mentors, family and friends that made it possible for me to accomplish my goal. I could not have achieved such a high aspiration without your guidance, support and love. I am truly blessed to have so many wonderful people in my life.

I would like to express my sincerest thanks and gratitude to my advisor, Dr.

Jingdong Tian, for taking me in his lab, appreciating my strengths and personal characteristics and providing me with guidance and confidence to be a good, solid researcher and a better person. Your understanding, kindness and support were endless.

I know that I am a much better researcher from having worked with you and I am forever grateful for the opportunities you have given me.

I would next like to thank my thesis committee members, both past and present:

Drs. Ashutosh Chilkoti, Uwe Ohler, Kent Weinhold and Lingchong You. They have provided me guidance and very valuable advice over the years. I also want to extend my thanks to Dr. Zhongying Chen, who taught me many lab techniques and generated valuable reagents when I first joined the lab which I made good use in the years to follow.

I owe my deepest gratitude to my parents. Everything I have achieved was because of them. They have loved me and supported me so much and they have always encouraged me to give my best effort in whatever I do in life and enjoy the journey

along the way. And finally I want to say thank you to my fiancé, Rui Zhang and his parents. Rui has been such a supportive, patient and loving boyfriend and fiancé and I'm forever grateful to having met him. We appreciate each other, learn from each other and we are both better person when we are with each other. Rui's optimism, encouragement and sense of humor helped me so much in my last two years in graduate school. And Rui's parents are just wonderful parents-in-law. They took me in as a daughter since we first met and loved me no less than my parents since then on. I'm forever grateful for their genuine love, understanding and support.

#### 1. Introduction

#### 1.1 Molecular cloning technologies

Molecular cloning is a foundational technology for molecular biology and biotechnology. Pioneered by the restriction digestion and ligation based method [1-3], new cloning technologies have continuously been invented and evolved to suit various requirements and applications. Depending on whether specific sites or sequences are used in the insert and the vector for cloning, cloning methods can be broadly divided into two categories: sequence-dependent and sequence-independent. Sequencedependent cloning is based either on restriction digestion and ligation, or site-specific recombination, such as the Univector plasmid-fusion system [4] and Gateway [5, 6]. Sequence-independent cloning is largely based on homologous recombination and includes methods such as ligase-free [7] or ligation-independent cloning (LIC) [8], LIC with Uracil DNA glycosylase (UDG or USER cloning) [9, 10], MAGIC [11], SLIC [12], In-Fusion (Clontech) [13], and PIPE [14]. Although these methods all have their own special characteristics and advantages, new developments especially the emergence of synthetic biology have put ever increasing demands for more accurate, efficient, convenient and economical cloning technologies for purposes such as creating complex combinatorial synthetic gene libraries, gene circuits and metabolic pathways.

For synthetic biology applications involving high-complexity or multi-fragment cloning, sequence-dependent methods are generally inconvenient because they require

unique and specific sites in both the insert and the vector in order to generate the initial plasmids [4-6]. For this reason, the more flexible sequence-independent cloning methods are preferred. However, such methods usually require generating complementary single-stranded overhangs in both the insert and vector fragments, with or without RecA-mediation [8, 12, 14]. And some of these methods are not strictly sequenceindependent because they require the presence or absence of specific nucleotides at certain positions in the overlapping region [8, 15]. The generation of complementary single-stranded overhangs takes additional preparation steps and often uses expensive enzyme systems. These manipulations generally require large amounts of DNA at the beginning and tend to have insufficient efficiency for library cloning. Furthermore, the annealing step in these methods is normally performed at ambient temperature, which allows non-specific hybridization among single-stranded overhangs and leads to frequent assembly errors in multi-fragment cloning. Therefore, for the demanding tasks of assembling and cloning complex synthetic gene libraries and pathways, further improvements on accuracy and efficiency over existing methods would be highly desirable. For routine and high-throughput cloning, fewer steps and lower cost is always a significant improvement.

#### 1.2 Error correction technologies

Gene and genome syntheses are playing an increasingly important role in synthetic biology and biotechnology [32-36]. To increase throughput and reduce cost,

new gene synthesis methods that take advantage of DNA microarrays [37-39] and microfluidic devices [40-42] have recently been demonstrated. However, removing errors that arise from oligonucleotide synthesis and gene assembly remains a significant challenge, especially for gene synthesis using microarray-produced oligos, where error rates tend to be higher [37, 38]. Cloning and sequencing large numbers of synthetic constructs in order to identify correct clones has become a bottle neck for gene and genome syntheses.

A number of methods have been used to reduce synthesis errors. To improve the quality of gene-construction oligos, size exclusion purification using polyacrylamide gel electrophoresis (PAGE) [43] or high performance liquid chromatography (HPLC) [44] can be used to remove large insertions and deletions. An array hybridization method has also been developed to reduce errors in chip-generated oligo pools, which requires special microarrays of complementary oligos [37]. Using next-generation sequencing technology, it may also be feasible to sequence and select correct oligo sequences for gene construction, as a recent proof-of-concept experiment has demonstrated [45].

To eliminate errors in longer synthetic gene constructs, slow and labor-intensive cloning and sequencing methods are traditionally used. If the error rate is high or the sequence is long, large numbers of clones need to be sequenced in order to identify a correct sequence [46]. If a perfect clone cannot be isolated, site-directed mutagenesis

needs to be used to fix errors identified by sequencing [36, 47-50]. Multiple rounds of cloning, sequencing and site-directed mutagenesis can significantly increase the cost and turnaround time for gene synthesis.

In order to increase the chance of finding a correct clone, the overall error frequency in the synthetic gene pool needs to be significantly reduced. Methods of using mismatch-binding proteins (e.g. MutS) to remove error-containing DNA heteroduplexes have been developed [46, 51, 52]. However, MutS-based methods theoretically do not work well for error-rich sequences, because the correct sequences have to outnumber the erroneous sequences in order to avoid being depleted from the synthetic pool.

In comparison, methods using mismatch-cleaving enzymes show an advantage, as these enzymes can cleave the heteroduplexes at the vicinity of the mismatch sites, which allows the mutant bases to be subsequently removed by exonuclease activity present in the reaction mixture. A number of enzymes have been tested, including T7 endonuclease I, T4 endonuclease VII and *Escherichia coli* endonuclease V, which showed various effectiveness due to various specificities of the enzymes [53-55]. CEL endonuclease is a new member of the S1 nucleases isolated from celery, and prefers double-stranded mismatched DNA substrates [56, 57]. It is not inhibited by high GC content, and can cut mismatch-containing heteroduplexes efficiently at neutral pH whether the mismatches are base substitutions, insertions or deletions anywhere from 1 to 12 nt. CEL nuclease is able to act efficiently on molecules with multiple mismatches,

even with only 5 nt between mismatches. Additionally, it can handle substrates anywhere from 40 bp to ~30 kb. Its broad substrate specificity and low non-specific activity has made CEL nuclease one of the best tools for mismatch detection [56-60]. In a previous study, we first reported that Surveyor nuclease, a commercialized form of the CEL endonuclease, was effective in removing errors during chip-based gene synthesis [61]. Here, we describe detailed characterization of the molecular mechanism of the Surveyor-based error correction reaction (ECR) and the development of an optimized ECR protocol, which further reduced the error rate down to 1 error in ~8700 bp.

#### 1.3 Genetic manipulation of protein expression

Protein expression is a very complex multi-step process involving DNA replication, transcription, RNA turnover, translation, post-translational processing and protein stability. In addition to adjusting host-specific variables and environmental conditions, a number of regulatory elements, such as promoters [16, 17] and ribosomal binding sites [18, 19], have been used to modulate protein expression. However, if the protein-coding DNA sequence itself is poorly translatable in a given host, modifying these elements may have a limited effect. This suggests that recoding the sequence with synonymous codons may be required. However, we need to consider more than the sequences that will ensure enough mRNA and an adequate rate of translational initiation: the codon choices themselves must not limit expression under the anticipated conditions of use. When a gene is synthesized, it is generally modified from the natural

version: sometimes to eliminate or add restriction sites for manipulations using old-fashioned cloning methods, but more significantly to substitute synonymous codons in order for a natural gene to express in heterologous hosts. Moreover, it has not yet been possible to determine the full expression potential of a given protein in a given host or to use computer algorithms to reliably modify the coding sequence to achieve desired levels of protein expression [20, 21]. Existing methods of optimizing codon usage for protein production in a heterologous host are slow, costly and unreliable. This problem has become a bottleneck for biomedical research and pharmaceutical development and, if not addressed, could significantly hamper synthetic biology efforts to design and construct novel biomacromolecules, genetic networks and other synthetic biological systems [22-28].

Low-cost and high-throughput gene synthesis and precise modulation of protein expression are of critical importance to the development of synthetic biology and biotechnology. Recently our lab has developed an integrated on-chip gene synthesis technology, in which inkjet oligonucleotide array synthesis, amplification release, and parallel gene assembly were combined for the first time on a microchip. With this newly developed technology preferably in use, we plan to synthesize large quantities of oligo libraries and assemble them into gene libraries. Then we designed a high-throughput platform for assessing the expression potential of a protein and for reliably obtaining synthetic gene sequences with desired protein expression levels using these gene

libraries. The high-throughput protein expression strategy should meet the following criteria: a) target-independent: the strategy should be suitable for any target protein; b) efficient and high-throughput: the strategy should be able to screen a relatively large (>108) expression library in a short period of time (i.e. hours); c) stringent: the strategy should cleanly separate high-expression clones from low/none-expression ones; d) simple and economical: can be carried out in a standard academic research lab with standard equipments.

For years, research efforts to characterize and employ sequences that locate in the regulating regions of the target proteins are underway in many synthetic biology laboratories [16-19]. However, a design process that needs more attention is the treatment of coding sequence, because if coding sequence itself is poorly translatable in a given host environment, recoding it with synonymous codons seems to be the last resort. Some efforts have been done to improve coding sequences, mainly by altering synonymous codons at specific regions by applying one or a few rules [20-22, 29-31]. However, none of the rules has reliably improved expression; the tendency has been to layer more and more rules on top of each other, greatly complicating the gene design task and creating problems of prioritization when applying several conflicting principles. A robust and generally applicable gene design method must instead be based on well-established relationships that are validated by thorough experimentation. The

ability to create synthetic gene sets with variations in synonymous gene coding will be essential to elucidate these relationships.

#### 1.4 Codon usage and protein expression regulation

There are all together 61 nucleotide triplets (codons) encoding 20 natural amino acids. An amino acid can be encoded by as few as one or as many as six codons and the variations occur in most cases at the third and sometimes second codon position which is called 'redundancy'. This phenomenon suggests that a protein consisting of 100 amino acids of average amino acid composition could be encoded by more than 1030 different gene sequences. Therefore, if the codon choice at each position is considered an independent variable, theoretically the possibility of codon combinations is enormous. However, in nature, the codon usage is regulated and not totally random. The pattern of codon usage differs between organisms. Some codons are used more frequently in one organism, but rarely in another. It is generally believed that synonymous codon usage biases result from co-adaptation between codon usage and tRNAs abundance, mainly adopted by cells as a mechanism to optimize the efficiency of protein synthesis [16-19]. Therefore, when the codon usage of the target protein differs significantly from the average codon usage of the expression host, problems arise during expression. Calculation of the codon adaptation index [20] can give a general indication of the relative divergence of the codon distribution in the sequence from the distribution of codons in highly expressed genes in the target organism.

So far, there is no comprehensive and effective strategy for codon optimization. The commonly followed recommendation is to replace codons that are rarely found in highly expressed host genes with more favorable codons throughout the target gene [21-24]. In practice, two approaches are used. In the first approach, the protein sequence is back-translated, assigning only the most frequently used codon for each amino acid from the codon table. Although this approach will generate a gene with a perfect codon adaptation index value, it will allow only part of the tRNA pool to be used during translation. For example, in highly expressed human genes, CGC is the most frequently used codon for arginine but represents less than 40% of the total arginine codon distribution. Therefore, a sequence using only CGC for arginine would not be able to utilize 60% of the available tRNA for arginine. In the second approach, the protein sequence is back-translated by assigning codon so as to reflect the natural distribution in highly expressed genes [25, 26]. Codon-optimization algorithms exist to automate the process, such as Syngene [27], UpGene (www.vectorcore.pitt.edu/upgene/upgene.html), and Backtranslation [26], etc.

In addition to the general replacement of rare codons with abundant ones, there are numerous other factors to consider on a case-by-case basis. For example, when expressing prokaryotic genes, in vertebrate hosts, removing prokaryotic sequences such as CpG's from coding sequences by alternative codon usage has been shown to dramatically increase the amount and duration of gene expression through the

avoidance of gene silencing mechanisms [28-30]. Over- or under-represented codon pairing or clustering can cause ribosome pausing for structural reasons involving the ribosome or the mRNA [31-33]. Other elements or sequences that have been found to affect protein expression include AU-rich elements (ARE), near-consensus splice sites (NCSS), and HIV rev dependence [34-37]. Codon changes also alter other properties of the coding sequence simultaneously, including G+C% content, mRNA secondary structure, internal regulatory sequences, which all could affect protein translation efficiency or immunogenicity [28, 34, 36, 38-42]. Due to the lack of systemic studies and a comprehensive understanding of the factors influencing codon usage and protein expression, a huge amount of time, effort, and money are spent on designing, synthesis and testing of codon "optimized" sequences without ever knowing the real potential. The fact that de novo gene synthesis is still very costly and is conducted in a lowthroughput fashion makes the codon-optimization process even more difficult and costly.

# 2. A novel method for cloning of gene libraries and multi-DNA fragment assembly

#### 2.1 Introduction

Circular polymerase extension cloning (CPEC) is a simple, efficient and economical circular DNA assembly and cloning method developed to meet the ever-increasing demand from high-throughput genomics, proteomics and synthetic biology. Compared with existing cloning strategies, either sequence dependent or independent, CPEC offers significant benefits by combining simplicity, efficiency, versatility and cost-effectiveness in one method [43]. In addition to routine single-gene cloning, CPEC is ideal for a wide variety of other applications, including complex gene library cloning, high-throughput expression cloning and multi-way assembly of genetic pathways [44].

CPEC is a single-tube, one-step reaction that normally takes 5–10 min to complete for everyday laboratory cloning. The method is directional, sequence independent and ligase free. It uses the polymerase extension mechanism [45] to join overlapping DNA fragments into a double-stranded circular form, such as a plasmid. In a typical CPEC reaction, linear double-stranded insert(s) and vector are first heat-denatured; the resulting single strands then anneal with their overlapping ends and extend using each other as a template to form double-stranded circular plasmids. In CPEC, all overlapping regions between insert(s) and the vector are unique and carefully designed to have very similar and high melting temperatures (Tm), which eliminates vector re-annealing and concatenation of inserts and makes CPEC very efficient and

accurate. The low concentrations of fragments in the reaction favor plasmid circularization and effectively prevent plasmid concatenation. After the CPEC reaction, the perfectly formed double-stranded circular plasmids, with one nick in each strand, can be directly transformed into competent host cells (**Figure 1**).

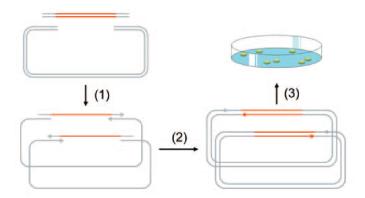


Figure 1: A schematic diagram of the proposed CPEC mechanism for cloning an individual gene. The vector and the insert share overlapping regions at the ends. After denaturation and annealing (Step 1), the hybridized insert and vector extend using each other as a template until they complete a full circle and reach their own 59-ends (Step 2). The final completely assembled plasmid has two nicks, one on each strand, at the positions marked by an arrow head. They can be used for transformation (Step 3) with or without further purification. For library cloning, the cycle maybe repeated in order to increase the yield of complete plasmids.

CPEC has its root in assembly PCR, also called polymerase cycling assembly (PCA), which has typically been used to assemble double-stranded linear DNA constructs [46-50]. An earlier attempt by Stemmer et al. [47] to assemble a circular plasmid using a pool of overlapping oligonucleotides resulted in the formation of linear concatemers of the plasmid unit, which had to be cut by a unique restriction enzyme and then circularized by ligation. Since then, there have been very few reports on using

overlap extension as a cloning method [51], until our laboratory recently developed and optimized the CPEC procedure not only for single-gene cloning but also for high-efficiency library cloning, multi-way cloning and combinatorial library cloning [44].

CPEC differs from PCA mainly in that CPEC forms abundant circular products of distinct lengths whereas PCA forms a smear of linear products, and the correct-sized product can only be obtained by further PCR amplification of PCA products.

Complex library cloning and multi-way pathway assembly require high cloning efficiency and accuracy. Although other relevant cloning methods only allow the overlapping fragments to anneal or recombine once, CPEC allows multiple annealing-extension cycles that not only increase the chance of hybridization but also permanently join the fragments through polymerase extension, thereby maximizing the cloning efficiency. Whereas the other relevant cloning methods perform the critical annealing/incubation step under ambient temperature, which tends to cause nonspecific hybridization and leads to compromised cloning efficiency and accuracy, CEPC designs the overlapping ends to have very similar Tm (±2 °C) and performs the annealing step at high, stringent temperatures (typically in the range of 55–65 °C) to ensure highest accuracy in multi-way assembly and complex library cloning. Unlike PCR, CPEC does not amplify sequences and therefore does not propagate errors with an increased number of thermal cycles.

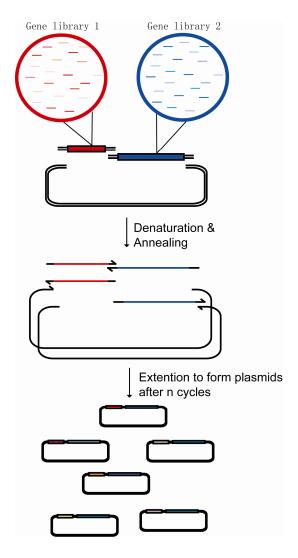


Figure 2: Schematic diagram of CPEC cloning of combinatorial gene libraries. Two gene libraries are cloned in frame into a vector. The vector and the inserts share overlapping regions at the ends. In each CPEC cycle, after denaturation and annealing, the hybridized inserts and vector extend using each other as a template until they complete a full circle. The assembled plasmid library can be directly used for transformation into competent cells.

We believe that CPEC has a real advantage in complex and combinatorial library cloning because of its exceptionally high efficiency [44]. The combinatorial library cloning strategy using CPEC is illustrated in Figure 2. In this example, two libraries are

cloned simultaneously into a single vector for expression or functional screens to identify the best combinatorial sequences. It is anticipated that such screens will be performed more and more frequently in synthetic biology applications to construct and identify the optimal macromolecular complexes or gene networks. So far, CPEC is the only *in vitro* method that works well in our hands for combinatorial library cloning [44]. The successful development of this cloning strategy will highly accelerate the process of protein expression screening using large quantity of library gene variants and hence and the development of synthetic biology applications as such screens will be performed more and more frequently to construct and identify the optimal macromolecular complexes or gene networks.

#### 2.2 Materials and methods

#### 2.2.1 Reagents and equipments

Cloning vectors was from Clontech and was modified. dNTP mix (dATP, dCTP, dGTP and dTTP), Phusion High-Fidelity DNA polymerase with 5× Phusion HF buffer, Taq DNA polymerase with ThermoPol reaction buffer, gel loading dye, blue (6×) and DNA ladder (1 kb amd 100 b) were from NEB. Custom DNA primer synthesis was available from IDT and a listed could be found in Supplementary Tables 1 and 2).

DNase/RNase-free water, Tris Acetate-EDTA buffer (10× TAE buffer), ethidium bromide solution, LB agar and glycerol were from Sigma-Aldrich. Agarose was from Denville Scientific. LB broth, Miller was from BD. GC5 competent cells were from Genesee

Scientific. S.O.C. medium was from Cellgro. GeneJET plasmid miniprep kit was from Fermentas. E.Z.N.A. gel extraction kit was from Omega Bio-Tek. ExoSAP-IT for PCR cleanup was from Affymetrix.

*E. coli* GC5 or DH5α competent cells were used for all CPEC clonings. LB (Miller) agar plates (Sigma) with appropriate antibiotics were used for culturing bacteria after transformation. Antibiotics were used at the following concentrations: carbenicillin (Cellgro) 100 μg/ml, kanamycin (Sigma) 30 μg/ml, and chloramphenicol (Sigma) 20 μg/ml. Phusion High-Fidelity DNA polymerase (Finnzymes) was used for CPEC reactions and Taq polymerase for single-colony PCR. The E.Z.N.A gel extraction kit (Omega Bio-Tek) was used for DNA purification.

A list of equipment used includes a thermal cycler, a microcentrifuge, an electrophoresis apparatus for agarose gels, a UV transilluminator, a gel imaging, a NanoDrop spectrophotometer, a shaker incubator at 37 °C, a cabinet incubator at 37 °C, a water bath at 42 °C, a microwave, glassware, razors, and parafilm.

#### 2.2.2 Reagent setup

Oligonucleotide primers. Prepare stock solutions of primers (e.g.,  $100 \mu M$ ) using DNase/RNase-free water. Prepare aliquots of  $10 \times$  working solution (e.g.,  $10 \mu M$ ) and store at –  $20 \, ^{\circ}$ C to prevent contamination of stock and repeat freeze–thaw cycles. Under proper storage conditions, active working solutions of primers that are subject to freeze–

thawing should be stable for several weeks, and reserve solutions of primers should be stable for at least a year.

1–1.5% (wt/vol) agarose gel. Place the gel trays and selected combs into the gel casting stand on a level surface. Weigh the required amount of agarose powder into an appropriate glass flask containing a measured amount of 1× TAE buffer. Heat in a microwave until the agarose is completely dissolved. Allow the melted agarose to cool down to 70 °C or lower before pouring. Add ethidium bromide stock (10 mg/ml) to the cooled agarose to a final concentration of 0.5  $\mu$ g/ml. Stir the solution to disperse the ethidium bromide and then pour it into the gel trays. Remove any air bubbles on the surface or inside the gel and allow the gel to solidify for ~20 min at room temperature (25 °C).

DNA ladders. Both 1-kb and 100-bp DNA ladders are diluted to 250  $\mu$ g/ml with DNase/RNase-free water and 6× gel loading dye (supplied with the DNA ladder). For a 5-mm-wide lane, 2  $\mu$ l of the mixture should be loaded onto the agarose gel. The amount of mixture should be scaled up or down, depending on the width of the lanes.

50% (vol/vol) sterile glycerol solution. Pour equal amounts of glycerol and ddH<sub>2</sub>O into a glass bottle. Mix well and autoclave. The glycerol solution can be stored at room temperature.

## 2.2.3 Design of overlapping sequences between vector and insert(s)

In order to make complex library cloning and multi-way cloning successful, it is very important to carefully choose and design the overlapping sequences between the vector and the insert(s) so that all the overlapping regions share similar melting temperatures (Tm). The Tm of the overlapping regions should be as high as possible (ideally between 60-70°C) in order to maximize hybridization specificity. The melting temperatures of all overlapping regions in the final CPEC assembly reaction should match each other as closely as possible, ideally with differences within ± 2-3°C. This helped eliminate mis-hybridization and ensure highest cloning efficiency and accuracy. The length of the overlapping region, typically between 15-35 bases, is of secondary consideration and is dictated by the Tm. Standard PCR primer selection rules and software can be applied to facilitate the design process [52]. If using PCR to introduce overlapping regions with the vector or with adjacent fragments, primers should be designed to include at least two parts, each hybridizing to one end of the two neighboring fragments to be joined. If an additional short sequence needs to be inserted between two existing fragments, it can be simply included in the primer design between the two overlapping regions. CPEC primer design examples for single and multiple insert cloning are given in Figure 3. In case synthetic genes or libraries are used, it would be more convenient to directly add overlapping regions during synthesis. No chemical modifications to the primers are required (e.g., phosphorylations).

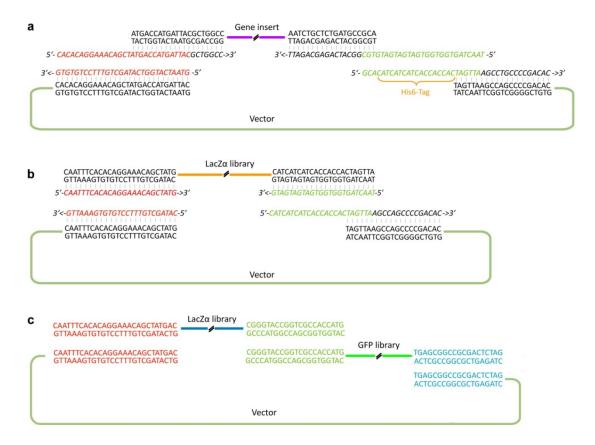


Figure 3: Schematic diagrams of primer design in CPEC. Fragment ends are shown as double stranded. Overlapping regions between the insert(s) and the vector are shown in colors. The 5'->3' direction of the PCR primers used to amplify the insert(s) or the vector is marked. (a) An example of a single insert cloning. The primers designed to amplify the gene insert introduce overlapping regions with the vector. Two primers are used to add an 18-base sequence encoding a His6-tag to the end of the gene. (b) Primer sequences used for  $lacZ\alpha$  gene library cloning as described in Anticipated Results. (c) Sequences of overlapping regions in combinatorial library cloning of the  $lacZ\alpha$  gene library and the GFP gene library as described in Anticipated Results. Overlapping regions have been added to the ends of the synthetic  $lacZ\alpha$  and GFP libraries. Therefore, there is no need to use bipartite PCR primers to add overlapping regions.

#### 2.2.4 Preparation of linear vector

The linear vector can be prepared most conveniently by PCR amplification using primers designed to introduce overlapping regions with the insert(s), as described below

in the Procedure; this approach offers the most flexibility in selecting cloning sites. If a convenient restriction site is available on the vector that does not introduce unwanted sequences, restriction digestion can also be used to linearize the vector. To prevent carryover of undigested or intact circular vector templates, we recommend gel purification of the linear vector after PCR amplification or restriction digestion. In addition, to eliminate the effect of any residual carryover vector, we recommend using an empty vector as the starting material for PCR amplification or restriction digestion; this way, any carryover of the empty vector will not interfere with downstream functional assays or screens.

The pUC19stop vector was constructed by first inserting a stop codon TAA after nucleotide position 425 of pUC19 plasmid (Invitrogen) and then moving the multiple cloning site (MCS, from nucleotide position 395 to 454) out of the open reading frame. The pUC19stop plasmid was linearized and amplified by PCR using primers SOSH6-L and SOSH6-R. pAcGFP1N1 vector was linearized and amplified by PCR using primers pAcGFP1N1Fw3 and pAcGFP1N1Rv3. pASK was amplified from pASK-IBA7C vector (Cayman Chemical) by PCR using primers pASKFw2 and pASKRv. PhaAB was amplified from a previously constructed plasmid phaCAB Topo 15 by PCR using primers phaABFw and phaABRv. The terminator was amplified from commercial vector by PCR using primers TermFw and TermRv. Cat2phaC was amplified from a previously

constructed plasmid pSOS-cat2phaC by PCR using primers cat2phaCFw and cat2phaCRv2.

# 2.2.5 Preparation of inserts

The inserts can be a single gene, a gene library, multiple genes or even multiple libraries. They can be isolated from natural sources or synthesized on the basis of in silico designs. Irrespective of whether they are single sequences or libraries, ensure that they share overlapping regions with the vector or neighboring fragments, as described above. If PCR is used to prepare the inserts, as described in the Procedure below, a high-fidelity DNA polymerase (e.g., Phusion DNA polymerase) is preferred in order to minimize the introduction of mutations or addition of an extra nucleotide at the ends of amplified products. Gel purification is sometimes necessary to ensure purity of the products.

The inserts of the  $lacZ\alpha$  and the HIV gp120 gene libraries were obtained from other projects in the lab, which were made by substituting original codons with selected synonymous codons used in  $E.\ coli$ . Primers LacZH-L and LacZH-R were used for amplifying the  $lacZ\alpha$  library. The HIV gp140 gene was split into two fragments, VacF1 and VacF2. Primer pairs GP140-R/GP140-28L and GP140L/GP140-29R were used to amplify VacF1 and VacF2, respectively. GP140-28L and GP140-29R were complementary with each other.

## 2.2.6 CPEC cloning

In the final CPEC assembly and cloning reaction, prepared linear vector and inserts are mixed together with the reaction cocktail, which includes dNTPs,  $1 \times HF$  Buffer and Phusion DNA polymerase. The composition of the CPEC reaction cocktail is almost identical to that of a standard PCR, except that no primers are added. The final vector concentration was normally in the range of 5–10 ng/ $\mu$ l and the insert-to-vector molar ratio was in the range of 1:1 to 2:1. The insert and vector mixture was denatured at 98 °C for 30 s, annealed at 55 °C for 30 s, and extended for 15 s per kb according to the length of the longest piece. In the end an extra extension period of 5 m was added. For average-sized vectors and inserts, the total reaction time was less than 5 m. We transformed 1–4 ml of the mixture into 50 ml of chemically competent GC5 $\alpha$  cells and plated a fraction of them on carbenicillin plates with 2% X-gal.

# 2.2.7 Multi-cycle CPEC for library cloning

We set up the cloning reaction exactly the same way as in single-cycle CPEC. After the initial 30 s denaturation step, we performed multiple cycles each consisted of 10 s denaturation at 98 °C, 30 s annealing at 55 °C, and extension at 72 °C for 20–30 s per kb according to the length of the longest piece. We ended the reaction with an extra 5 m of extension. We transformed a fraction of the reaction mixture into cells and plated an aliquot of the cells on a carbenicillin plate with 2% X-gal.

## 2.2.8 Combinatorial library cloning

For the strategy combining PCA with CPEC, we set up the PCA reaction by mixing 100 ng of each sub-library (VacF1 and VacF2) with the Phusion reaction mixture in a 50 ml volume. After an initial 30 s denaturation at 98 °C, we performed 5 cycles of PCA which consisted of 10 s denaturation at 98 °C, 30 s annealing at 52 °C, and extension at 72 °C for 20 s, and completed with an extension step at 72 °C for 5 m. We used 0.5 ml of the PCA reaction as a template and performed 30 cycles of PCR amplification of the full-length library using GP140R and GP140L primers (**Appendix A**) and the Phusion enzyme. We performed multi-cycle CPEC using identical conditions as described except that during the annealing step, we applied slow ramping at 0.1 °C/s from 70 °C to 55 °C before annealing at 55 °C for 30 s.

# 2.2.9 Multi-way CPEC

We mixed equal molar concentrations of the insert and vector fragments for multi-way CPEC. We used extension time which was sufficient to cover the full-length of the plasmid. Otherwise the reaction condition was identical to the multi-cycle CPEC. We transformed 1.25 ml of the reaction into 50 ml chemically competent DH5 $\alpha$  cells and plated 100–200 ml aliquots from 1 ml culture on 2% agar plates containing 20 mg/ml chloramphenicol and 0.5 mg/ml Nile Red.

# 2.2.10 CPEC product gel electrophoresis

To quickly assess whether a CPEC reaction is successful before proceeding to the transformation step, a small aliquot of the reaction can be separated by electrophoresis with an agarose gel. An appropriate molecular marker and the amount of insert/vector DNA present in the initial CPEC reaction should be loaded side by side onto the gel as controls. If the reaction has been successful, a new high-molecular-weight band corresponding to the total linear length of the final construct should be visible on the gel. The presence of intermediate assembly products or unincorporated vector/insert(s) in the same lane is not a problem for downstream experiments and need not be removed as long as a band representing the full-length product is visible on the gel.

#### 2.3 Results

# 2.3.1 CPEC cloning of a single gene

To confirm the validity of this mechanism, we first attempted cloning of a simple test gene,  $lacZ\alpha$ , into a modified pUC19 expression vector (Appendix A). The vector was linearized using either restriction digestion or PCR method. We added sequences on both ends of the  $lacZ\alpha$  gene to overlap with the ends of the linearized vector. The overlapping regions between the inset and the vector were designed to have similar melting temperatures (Tm), which were typically between 60–70 °C (Appendix A). We mixed the linearized vector with the  $lacZ\alpha$  gene without adding any PCR primers in an otherwise typical PCR reaction mixture.

We performed CPEC cloning as we would perform one cycle of PCR using a high-fidelity DNA polymerase. The reaction involved a brief denaturation step to denature the double-stranded insert and the linear vector, an annealing step for the overlapping ends of the insert and the vector to hybridize, and an extension step to form a complete plasmid (see Methods). We examined a small aliquot of the reaction mixture using DNA agarose gel electrophoresis and used another small amount for transformation.

The gel electrophoresis results showed formation of a significant amount of vector-insert merging product (**Figure 4**, lane 1, upper band) after only one reaction cycle with equal molar concentrations of the vector and the insert. The amount of this product increased proportionally after 2 and 5 reaction cycles (**Figure 4**, lanes 2 and 3, upper band). This band appeared to migrate at the same position as the purified full-length plasmid in its nicked relaxed conformation (Figure 4, lane 4, upper band). An examination of the transform results found that approximately 100% of the colonies showed blue color, indicating minimal cloning error or carry-over of intact vectors. Sequencing results of randomly picked colonies confirmed that the cloning reaction happened exactly as expected, with no mutations at the cloning junctions.

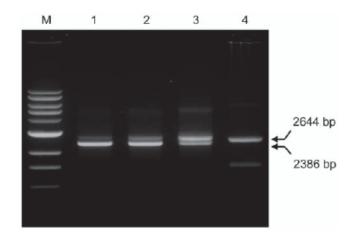


Figure 4: CPEC cloning of the *lacZα* gene. The image shows gel electrophoresis analysis of the CPEC reaction product after 1, 2 and 5 cycles (lanes 1–3). 5 ml of the reaction was separated on a 0.8% agarose gel and visualized after ethidium bromide staining. The assembled full-length plasmid was 2644 bp; the empty vector, 2386 bp. A sequence verified, full-length plasmid purified from a bacteria colony was used as a positive control (lane 4). The upper band (2644 bp) represented the relaxed circular form and the fastmigrating lower band, the closed circular form of the plasmid. The molecular weight marker used in this figure was NEB 1 kb DNA ladder (lane M).

# 2.3.2 CPEC cloning of a gene library

For individual gene cloning, we determined that one cycle of CPEC reaction was optimal. For complex gene library cloning where sufficient numbers of clones need to be obtained in order to maintain the complexity of the library, more cycles of CPEC reaction might be needed. To determine the optimal library cloning conditions with CPEC, we examined the cloning a synthetic library containing codon variants of the  $lacZ\alpha$  gene that was designed and synthesized for studying the effects of synonymous codon usage on protein expression. We selected the  $lacZ\alpha$  gene because we could use the blue or white color of the colonies to demonstrate the cloning and expression results. For

this complex gene library, no convenient restriction sites could be found for cloning into the modified pUC19 expression vector without cutting a fraction of the insert sequences. Therefore, a sequence-independent cloning method must be used.

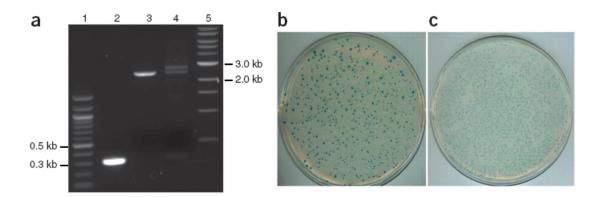


Figure 5: CPEC cloning of the  $lacZ\alpha$  gene library. (a) Agarose gel examination of the gene library inserts (lane 2), the linearized vector (lane 3) and the CPEC cloning products after five thermal cycles (lane 4). The upper band in lane 4 represents the full-length cloning product; the lower band represents the empty vector. Lanes 1 and 5 are 100-bp and 1-kb DNA ladders (NEB), respectively. Amount of DNA loaded: 300 ng for the insert, 200 ng for the vector and 15  $\mu$ l out of 25  $\mu$ l of the cloning product. (b) *E. coli* colonies expressing a small portion of the library of  $lacZ\alpha$  codon variants show a wide range of variation in the intensity of blue color. (c) Control *E. coli* colonies expressing wild-type  $lacZ\alpha$  show very little variation in color intensity.

We performed CPEC cloning of the library and determined the cloning efficiency at different cycle numbers (see Methods). Cloning result with 5 thermal cycles was shown below (**Figure 5**). The results indicated that 5,333 transformants were obtained from one nanogram of insert after only one cycle, which was sufficient for routine library cloning. The number of transformants obtained peaked at around 15 cycles and reached 56,000 colony forming units (c.f.u.) per nanogram of insert (**Figure 6A**). As a comparison, we typically achieved approximately 1,200~1,500 c.f.u per nanogram of

control insert using the ligation method. Approximately 100% of the colonies on the positive control plate transformed with wtlacZ $\alpha$  gene showed blue color, while the colonies with codon variations demonstrated, as expected, a wide range of blue color intensities. We isolated and sequenced plasmids from several hundred independent colonies which showed different intensities of the blue color. The sequencing results showed the presence of a distinct codon variant sequence of  $lacZ\alpha$  in every plasmid, which demonstrated correct and unbiased cloning of the complex library using CPEC. We further investigated the percentage of clones that might carry more than one insert after a different number of CPEC cycles by single-colony PCR using primers on the vector (Figure 6B). Sixteen colonies were randomly picked from plates culturing cells transformed with CPEC reaction product after 1, 2, 5 and 15 cycles. It was found that all 64 plasmids examined contained the correct, single-copy insert. This result suggested that the CPEC cloning mechanism was highly specific and did not favor carryover of concatemers, if any, into the final clones.

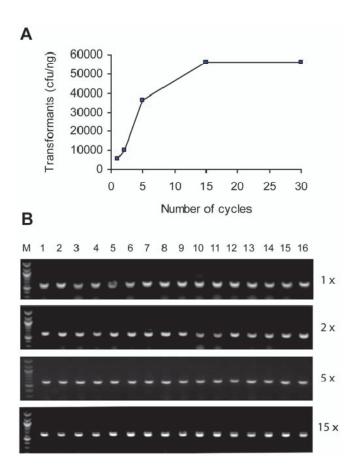


Figure 6: Gene library cloning using CPEC. (A) Cloning efficiency of CPEC at different cycles using the *lacZα* codon variants library. (B) Examination of the length of the insert from 64 independent colonies by single-colony PCR. The colonies came from cells transformed with CPEC reaction products after 1, 2, 5, and 15 cycles. The vector-insert ratio in the CPEC reactions was 1:1. The length of the amplicon with one insert was 592 bp. Single-copy inserts were found in all of the 64 colonies examined. The molecular weight marker used in this figure was NEB 100 bp DNA ladder (lane M).

# 2.3.3 CPEC cloning of CPEC cloning of combinatorial libraries

Construction of combinatorial library is extremely useful for synthetic biology and molecular evolution. We designed and tested two strategies of constructing complex combinatorial synthetic gene libraries using CPEC. The first strategy was to assemble the full-length inserts from shorter fragments first, followed by cloning the pre-assembled

full-length inserts into a vector by CPEC. The second strategy was to combine the assembly and cloning steps into one CPEC reaction (**Figure 7**).

For these tests, we selected a synthetic library which contained codon variants of the HIV envelope gene, gp120. The 1.7-kb full length codon variant library was divided into two fragments of approximately equal lengths, which were synthesized separately (Appendix A). The two fragments and the vector were designed to share overlapping sequences with similar melting temperatures. To test the first strategy, we preassembled the 1.7-kb combinatorial library using a two-step polymerase cycling assembly (PCA) reaction [49] (see Methods) and then mixed the insert with the linearized vector and performed a multi-cycle CPEC reaction. An aliquot of the reaction was taken after 5, 10 and 20 cycles, respectively, and the reaction products were analyzed by gel electrophoresis (Figure 7A, lanes 1–3). The results indicated that after 5 cycles of CPEC, a significant amount of the full-length 6.4-kb plasmid had formed. By 10 cycles, approximately 80% of the 1.7-kb inserts had merged with the vector. After 20 cycles, all free inserts and vector DNA had merged to form the complete plasmid.

Next, we tested one-step combinatorial assembly and cloning of the library from two sub-libraries using CPEC. We mixed the two insert libraries with the linearized vector in equal molar concentrations and performed 25 cycles of CPEC. The annealing step was carefully controlled in terms of annealing temperature and cooling rate in order to achieve highest hybridization efficiency and accuracy. A single band

representing the 6.4-kb complete plasmid was clearly visible in gel electrophoresis analysis (**Figure 7B**). We then transformed an aliquot of the reaction mixture directly into competent cells. Approximately 2.436105 colonies were obtained from one picomole of vector DNA. We randomly picked independent colonies on the plate and performed single colony PCR to verify the presence of the correct insert. The result showed that all 16 colonies examined contained the full-length insert, indicating a 100% cloning efficiency (**Figure 7C**).

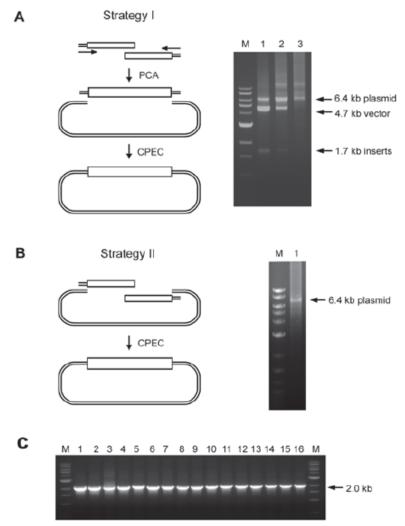


Figure 7: Combinatorial gene library cloning using CPEC. (A) A multi-step strategy combining PCA and CPEC for constructing combinatorial synthetic gene library. As shown in the schematic diagram on the left, in the first step, polymerase cycle assembly (PCA) is used to assemble two sub-libraries into a full-length library; in the second step, CPEC is employed to clone the full-length gp120 gene library inserts (1.7 kb) into the vector (4.7 kb). The gel electrophoresis picture on the right shows the analysis of CPEC progression after 5, 10 and 20 cycles (lanes 1–3). The molecular weight marker used in this figure was NEB 1 kb DNA ladder (lane M). (B) A one-step strategy of constructing combinatorial library using CPEC. The two sub-libraries and the linear vector were mixed in equal concentrations in the CPEC reaction. After 25 cycles, the reaction product was analyzed by 0.8% agarose gel electrophoresis. Arrow marked the 6.4-kb band representing the full-length plasmid. The molecular weight marker used in this figure was NEB 2-log DNA ladder. (C) Gel electrophoresis analysis of inserts amplified from 16 independent colonies from the

gp120 library cloned using the one-step CPEC strategy. The molecular weight marker used in this figure was NEB 1 kb DNA ladder.

## 2.3.4 CPEC cloning of multi-component pathway

We then tested if CPEC can be applied for efficient assembly and cloning of multi-component pathways in a single reaction. The proposed multi-way cloning mechanism is shown in **Figure 8A**. Similar to combinatorial library cloning, multi-way assembly using CPEC may also require longer annealing time by slow-ramping and increasing thermal cycle numbers. However, unlike PCR, multi-cycle CPEC is not an amplification process, therefore will not accumulate or propagate errors generated by the DNA polymerase.

Multi-way CPEC was tested on the construction of a metabolic pathway for synthesizing a biodegradable plastic, poly(3HB-co-4HB) in *E. coli*. The pathway consisted of four genes and additional regulatory elements. To construct the plasmid, four PCR fragments of various lengths: 3280, 2959, 2047, and 171 bp, each representing one or several functional parts in the pathway, were assembled. The total length of the ensemble was 8360 bp (Appendix A). These fragments were mixed in equal molar concentrations and different thermal cycles were tested for the cloning. Cloning result was analyzed on an agarose gel and transformed into competent cells and plated on chloramphenical plates containing Nile Red. The transformation efficiency was calculated from the plate and functional test were performed. Besides, plasmids from

randomly picked colonies will be used to perform restriction mapping to confirm the result of the CPEC multi-way cloning.

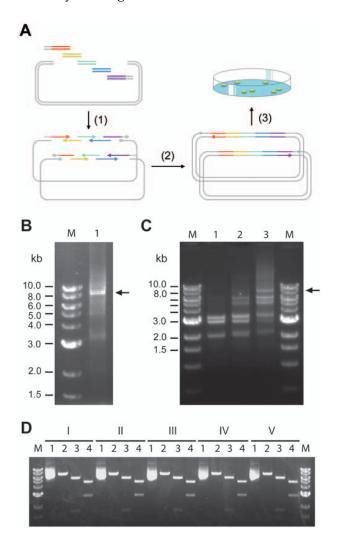


Figure 8: Assembly of multi-component pathway using CPEC. (A) A schematic diagram of the multi-way CPEC. Any two neighboring fragments share an overlapping region with identical Tm. Multiple cycles are usually needed to drive the reaction to completion. The positions of the two nicks (arrow head) in the final completely assembled plasmid may vary depending on the number, lengths, and sequences of the fragments. (B) Gel electrophoresis analysis of the final assembly product after a 20-cycle CPEC. 5 ml of the reaction was separated on a 0.8% agarose gel and visualized after ethidium bromide staining. The full-length plasmid was 8360 bp. (C) Gel electrophoresis analysis of the multi-way CPEC reaction. 5 ml was taken out

of the reaction after 2, 5, and 10 cycles and separated on a 0.8% agarose gel (lanes 1, 2 and 3). The starting lengths of the four fragments were 3280, 2959, 2040, and 171 bp, respectively. The 171-bp band was not visible. (D) Restriction mapping of the isolated plasmids derived from the CPEC reaction. Plasmid DNA from five independent colonies (I-V) were digested with BamHI (lane 2, 8.4 kb), BamHI-XhoI (lane 3, 6.6 kb and 1.8 kb), and NdeI (lane 4, 5.4 kb and 3 kb). Purified plasmids not subjected to restriction digestion are shown in lane 1. The molecular weight marker used in this figure was NEB 1 kb DNA ladder.

#### 2.4 Discussion and conclusions

Unlike any other cloning method, CPEC relies solely on the simple and robust polymerase extension mechanism to clone individual genes, libraries, or multiple fragments. In a single closed-tube reaction, the insert and vector fragments are first heat denatured, then annealed at elevated temperatures to ensure specific hybridization of overlapping regions, and finally extended to form complete plasmids, leaving only one nick in each strand. The fully-formed relaxed double-stranded plasmids are then efficiently introduced into *E. coli* cells where the nicks are sealed and covalently closed plasmids are formed. The most significant advantages of CPEC include accuracy, efficiency, convenience and cost-effectiveness in complex library and pathway assembly.

For routine single-gene cloning, one denature-annealing extension cycle is sufficient and optimal, which can be completed in five minutes. For library cloning where sufficient numbers of clones need to be obtained in order to maintain the complexity of the library, CPEC offers the unique advantage of being able to perform multiple cycles to maximize the total number of clones constructed without using excessive amounts of vector DNA. We recommend 2–5 cycles depending on library

complexity. For multi-fragment cloning, 5–25 cycles may be used depending on the number of fragments.

Unlike PCR, CPEC is not an amplification process and therefore will not accumulate mutations. However, excessive numbers of cycles should be avoided in order to minimize possible concatemer formation. In cases where concatemers may form, the cloning efficiency will not be significantly affected because concatemers usually do not have the correct complementary ends for efficient circularization and therefore will not form covalently closed plasmids in the cells.

Only PCR polymerases with no strand displacement activity should be used in CPEC reactions to avoid long concatemer formation or other cloning artifacts. Many of the commercially available PCR polymerases belong to this category. The Phusion DNA polymerase was used in this study due to its robustness, speed and accuracy. The reaction conditions may need to be adjusted if other polymerases are used, especially the extension time. PCR polymerases with low efficiency or low fidelity should be avoided for demanding cloning tasks using CPEC. Compared to sequence-dependent cloning methods, such as those mediated by restriction-ligation or site-specific recombination, CPEC has the advantage of complete flexibility with respect to sequence junctions.

Compared to other sequence-independent cloning methods, in addition to enjoying all of their benefits, CPEC offers other significant advantages. First, CPEC eliminates the extra steps or enzymes required by other sequence-independent cloning

methods to generate single-stranded regions for annealing. For example, in LIC, overlapping sequences lacking a particular dNTP are added to the insert by PCR and complementary 12-nt single-stranded regions in both the insert and the vector are generated by T4 DNA polymerase treatment in the presence of that particular dNTP. In UDG-based methods, a ribonucleotide U replaces a T in the PCR primers used to add overlapping sequences to the insert and subsequent treatment with UDG enzyme generates single-stranded ends in both the insert and the vector for annealing. In SLIC, T4 DNA polymerase treatment or incomplete PCR with two pairs of primers are used to generate mixed products containing ss-overlapping regions. In PIPE, a different version of incomplete PCR is used so that some PCR products are not fully extended and therefore leave heterogeneous single-stranded regions toward the ends. These extra preparation steps take more time, more DNA, and many require extra expensive enzymes or reagents. In contrast, CPEC uses double-stranded overlapping inserts and vector directly without any treatment. The whole single-cycle CPEC reaction can be completed in 5minutes and uses only a PCR polymerase, making CPEC one of the most convenient, economical and versatile cloning methods, which can also be easily adapted for high-throughput cloning.

Another notable advantage of CPEC is its high cloning accuracy and efficiency, which makes it uniquely suitable for complex, combinatorial, multi-fragment or multi-library cloning. In CPEC, all overlapping regions among fragments are designed to have

similar high melting temperatures (typically 55–70 °C) so annealing between fragments can be very specific. This is most desirable for complex, combinatorial or multi-fragment cloning where non-specific annealing can cause cloning errors. It is our experience that typically 95–100% of CPEC-generated colonies contain the correct inserts, including multi-way assembly. All other sequence-independent cloning methods use ambient annealing temperatures and, as a result, the specificity and success rate of multi-way cloning can be significantly compromised.

The high cloning efficiency of CPEC, especially for multi-way or complex library cloning, comes from a combination of two special features. First, CPEC forms covalently joined complete plasmids *in vitro*. Secondly, multiple CPEC cycles can drive the reaction into near completion. In contrast, all other sequence-independent cloning methods either loosely anneal fragments without covalent bonding or allow only a small fraction of the fragments to form plasmids due to the low efficiency of multi-fragment hybridization.

With increasing demands for complex or combinatorial library cloning and multi-fragment gene pathway and network assembly, we expect CPEC to play a significant role in various applications of synthetic biology. It will enable rapid and high-throughput construction of combinatorial libraries, gene circuits and pathways. It will also liberate researchers from tedious and time-consuming everyday cloning tasks.

In conclusion, the CPEC cloning strategy can accomplish all the tasks that other *in vitro* cloning methods can perform and compares favorably in almost every aspect, including accuracy, efficiency, speed and cost [44, 53]. However, because of the intrinsic limitations of DNA polymerase and the overlap extension method, we expect that there will be limits to the size of the plasmid and the total number of fragments that can be assembled by CPEC, although no exhaustive experiments have been done to determine those limits. So far, we have comfortably assembled an 8.4 kb plasmid using four PCR fragments of 3,280, 2,959, 2,047 and 171 bp [44]. With careful optimization, linear products as long as 20 kb have been constructed using overlap extension PCR [48], which could indicate an approximate upper limit for CPEC. To assemble bigger constructs with more fragments, *in vivo* homologous recombination methods using yeast as a host organism may be attempted [54, 55].

# 3. Error-correction in gene synthesis and optimization technology

#### 3.1 Introduction

The development of economical and high-throughput gene synthesis technology has been hampered by the high occurrence of errors in the synthesized products, which requires extensive labor and time to correct. Here, we describe an error correction reaction (ECR), which employs Surveyor, a mismatch-specific DNA endonuclease, to remove errors from synthetic genes. In ECR reactions, errors are revealed as mismatches by re-annealing of the synthetic gene products. Mismatches are recognized and excised by a combination of mismatch-specific endonuclease and 3'-> 5' exonuclease activities in the reaction mixture. Finally, overlap extension polymerase chain reaction (OE-PCR) reassembles the resulting fragments into intact genes. The process can be iterated for increased fidelity. With two iterations, we were able to reduce errors in synthetic genes by >16-fold, yielding a final error rate of ~1 in 8700 bp.

#### 3.2 Materials and methods

### 3.2.1 Reagents

Chemicals were purchased either from Sigma-Aldrich or VWR. Enzymes were from New England Biolabs. The Surveyor nuclease was purchased from Transgenomic as part of the Surveyor Mutation Detection Kit. GC5 chemical competent cells were purchased from Invitrogen. Oligonucleotides were synthesized in house on a plastic

chip using a custom-made inkjet DNA microarray synthesizer [56]. The exact oligos synthesized are listed in **Appendix A**. On-chip oligo amplification and gene assembly using combined nicking strand displacement and polymerase cycle assembly (nSDA–PCA) reaction was performed as described with minor modifications [57].

# 3.2.2 ECR of assembled genes

Once PCR amplification of the on-chip assembled gene was completed, the gene products were purified by agarose gel electrophoresis and extracted to yield a concentration of >100 ng/ml (measured using a Nanodrop analyzer). These PCR products were then diluted with either 1X Taq buffer or 1X Phusion HF buffer to yield a final concentration of 50 ng/ml. This was then melted by heating at 95°C for 10 min, cooled to 85°C at 2°C/s and held for 1 min. It was then cooled down to 25°C at a rate of 0.3°C, holding for 1 min at every 10°C interval.

For ECR using a 20 min Surveyor cleavage incubation, 4 ml (200 ng) of the reannealed gene product was mixed with 0.5 ml of Surveyor nuclease and 0.5 ml enhancer [which is known to be DNA ligase in nature and enhances the reaction [57-59]] and incubated at 42°C for 20 min. Two microliter of the reaction mixture was used for subsequent overlap extension–PCR (OE-PCR) using the same reaction conditions as the PCR above. The OE-PCR product was cloned and sequenced to serve as the result from the first iteration of error correction. For the second iteration of error correction, the OE-PCR product band was diluted to 50 ng/ml using 1X Taq buffer and reannealed as

before. Similar to the first iteration, a 5 ml reaction consisting of 4 ml re-annealed product, 0.5 ml of Surveyor nuclease and 0.5 ml enhancer was incubated at 42°C for 20 min. Two microliter of the product was subjected to OE-PCR, cloned and sequenced to serve as the result from the second iteration of error correction.

For ECR using a 60 min Surveyor cleavage incubation, 8 ml of the re-annealed gene product in 1X Phusion buffer (final DNA concentration of 50 ng/ml) was added to 2 ml of Surveyor nuclease and 1 ml enhancer to yield a total of 11 ml that was then incubated at 42°C for 60 min. Two microliter of the reaction mixture was then subjected to OE-PCR, and the resulting PCR product was cloned and sequenced to serve as the result from the first iteration of error correction. For the second iteration, the product from the first iteration was diluted to 50 ng/ml using 1X Phusion buffer and re-annealed as before. Similar to the first iteration, an 11-ml reaction consisting of 8 ml of re-annealed product, 2 ml of Surveyor nuclease and 1 ml of enhancer was incubated at 42°C for 60 min. Two microliter of the product was used for OE-PCR and the PCR product was cloned and sequenced to serve as the result from the second iteration of error correction.

# 3.2.3 Cloning, sequencing and functional analysis of synthetic genes

Synthetic gene products, before or after ECR, were cloned into the pAcGFP1 vector using circular polymerase extension method (CPEC) [44, 53]. Briefly, 250 ng of the linear vector was mixed with the synthetic gene products at 1:2 molar ratios in a 25-ml CPEC reaction using Phusion polymerase. The reaction involved 10 cycles of

denaturation at 98°C for 10 s, annealing at 55–60°C for 30 s and extension at 72°C for 15 s, and finished with an extended elongation step at 72°C for 5 min.

Two microliter of the cloning product was transformed into GC5 chemically competent cells (Invitrogen) according to the manufacturer's instructions. Cells were grown on agar plates with 100 mg/ml carbenicillin for 16 h and then kept at room temperature for 48 h before been imaged in an AlphaImage gel documentation system. The percentage of fluorescent colonies was automatically determined using CellC program (http://sites.google.com/site/cellcsoftware/download). The results were verified by thresholding the UV images using Adobe Photoshop and counting using ImageJ. Sequence analysis was done by extracting plasmids from randomly selected colonies using a miniprep kit (Qiagen) and sequencing at the Duke University Sequencing Facility.

# 3.3 Results

# 3.3.1 General design of the ECR using Surveyor nuclease

In this study, I collaborated with Ishtiaq Saaem, Siying Ma and aimed to develop a simple and convenient method to effectively remove errors from synthetic genes. The general strategy of using the Surveyor endonuclease to correct errors in synthetic genes is illustrated in **Figure 9**. After gene synthesis, the products are denatured and reannealed to form mismatch-containing heteroduplexes (left panel). The subsequent ECR, right panel involves incubation of the re-annealed product with the Surveyor nuclease,

followed by OE-PCR using a proofreading DNA polymerase. The 3'->5' exonuclease activity of the DNA polymerase removes 30 overhangs that contain the mismatch base(s) and allows OE to proceed efficiently. Mismatch structures formed at the deletion, insertion and substitution sites in the heteroduplexes are recognized by the Surveyor mismatch-specific endonuclease, which cuts each strand at the phosphodiester bond at the 30 side of the mismatch site [58]. During the subsequent OE-PCR reaction, the 3'->5' exonuclease activity of the proof-reading DNA polymerase chews away any 30 overhangs that contain the mismatch base(s) (substitutions and insertions). Finally, the error-free fragments are extended and amplified into full-length gene constructs by the DNA polymerase. One round of ECR may not completely remove all errors and we are interested to determine whether more iterations of the ECR can be used to further remove any remaining errors.

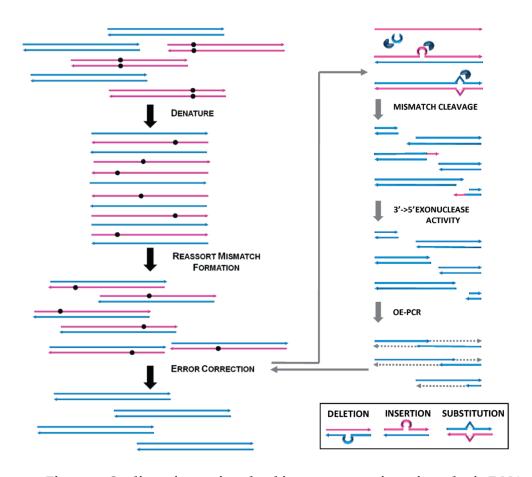


Figure 9: Outline of steps involved in error correction of synthetic DNA constructs. Gene synthesis products are heat denatured and then slowly cooled down to form heteroduplexes containing mismatches at the error sites (left panel). Heteroduplexes are cleaved by the Surveyor nuclease at the sites flanking the mismatch bulges. The resulting single-stranded overhangs, where mismatch bases are located, are removed by the proofreading exonuclease activity of Phusion polymerase used in the OE-PCR. The resulting fragments with mismatch bases removed are efficiently assembled back into full-length gene constructs during OE-PCR (right panel).

# 3.3.2 Determine error frequency of on-chip gene synthesis

Integrating oligo synthesis with gene assembly on a microchip can significantly reduce synthesis cost and increase throughput. As described in the Materials and Methods section, we synthesized DNA microarrays using a custom inkjet DNA

synthesizer and used a combined nSDA–PCA reaction for on-chip oligo amplification and gene assembly. To determine error frequency of on-chip gene synthesis without error correction, we chose red fluorescent protein (rfp) as a test gene for convenient screen of functionally correct genes, which served as a good approximation of sequence correct genes. After transforming the pAcGFP1-rfp plasmid construct into bacteria, the colonies produced were either non-fluorescent, dimly or brightly fluorescent. A rough approximation of synthesis quality without error correction could be made using colony counts on agar plates. Using automated colony counting, it was found that 50.2% of the rfp colonies formed from uncorrected product fluoresced brightly (Figure 10A).

DNA sequencing was performed on 42 randomly picked rfp colonies from both directions. The sequencing results indicate an error rate of ~1.9/kb (**Table 1**). Deletions were found to be the dominant form of errors (75.4%), which was similar to column DNA synthesis where monomers are not successfully added to all of the growing polymer chains.

4

Table 1: Error analysis of synthetic gene sequences before and after ECR with Surveyor nuclease

	Without	ECR1	ECR1	ECR2	ECR2
Error type	ECR	(20 min)	(60 min)	(20 min)	(60 min)
Deletion (total)	43	3	0	0	0
Single-base deletion	30	2	0	0	0
Multi-base deletion	13	1	0	0	0
Insertion (total)	4	0	0	0	0
Single-base insertion	3	0	0	0	0
Multi-base insertion	1	0	0	0	0
Substitution (total)	10	7	11	5	6
Transition					
G/C to A/T	3	2	3	1	2
A/T to G/C	3	4	0	2	1
Transversion					
G/C to C/G	0	0	2	0	0
G/C to T/A	1	1	4	1	1
A/T to C/G	2	0	1	0	2
A/T to T/A	1	0	1	1	0
Total errors	57	10	11	5	6
Bases sequenced	29958	31866	42714	27798	52206
Error frequency (errors per kb)	1.9	0.31	0.26	0.18	0.11
Error frequency (bases per error)	526	3187	3883	5560	8701

Random clones of synthetic genes before (without ECR) or after one or two ECR iterations (ECR1, ECR2) were sequenced in both directions. Surveyor incubation time (20 min or 60 min) was indicated. The occurrence of different type of errors was counted.

# 3.3.3 Reduction of error frequencies after ECR

Both functional colony counting and DNA sequencing were performed to estimate error frequencies of chip-synthesized genes after ECR with 20-min or 60-min Surveyor treatment. It was reasoned that in one round of ECR, error sequences could form homodimers by chance during annealing and thus escape detection and cleavage. We, therefore, tested whether an additional round of ECR could eliminate more errors. Two iterations of ECR were performed with both 20 min and 60 min incubations as outlined in the Materials and Methods section. Full-length gene products were cloned and used for functional assays and Sanger sequencing in order to estimate error frequencies.

As shown in **Figure 10A**, increasing Surveyor cleavage time and number of iterations led to increases in the number of brightly fluorescent colonies. Using 20-min Surveyor treatment, the fluorescent population increased from 50.2% (untreated) to 74% and 84% in the first and second iteration, respectively. Using 60-min Surveyor treatment resulted in 78.4% and 94% fluorescent colonies after the first and second iteration. Example images showing the fluorescent colonies can be found in **Figure 11**.

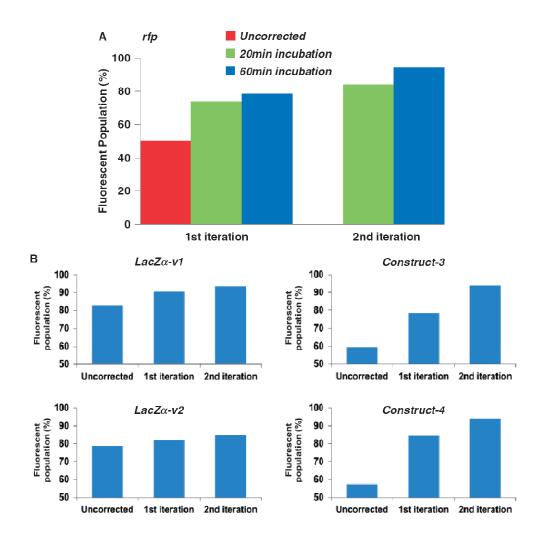


Figure 10: ECR results as measured by gene function or reporter assays. Percentage of functional or fluorescent clones was measured before and after one or two iterations of ECR for five different gene constructs. (A) The effects of Surveyor incubation time (20 and 60 min) and number of ECR iterations on the synthesis of rfp gene by counting fluorescent colonies. (B) The percentage of blue ( $lacZ\alpha$ -v1&2) or fluorescent colonies (constructs 3 and 4) after 1 or 2 ECR iterations.

To investigate the reproducibility and robustness of the method, we applied it to synthesis of four additional gene constructs and measured its effectiveness using functional or reporter assays. Of the four constructs, two were codon variants of the  $lacZ\alpha$  gene, the expression of which cause the colony to turn blue in the presence of X-

gal. The other two constructs could not be screened by their own functions and therefore were fused to the N-terminus of the green fluorescent protein (GFP) (**Figure 10B**). Blue or fluorescent colonies indicated that there were no frame shifts or mutations in the gene constructs that could abolish the function or expression of the genes. Therefore, the percentage of positive colonies could be used as an approximate indicator of the quality of the sequences. The results from the four additional constructs showed iterative increases in positive populations after each round of ECR (**Figure 10B**). As expected from the model predictions shown in **Figure 12A**, the small  $lacZ\alpha$  genes had a large fluorescent population even before error correction (~80% positive) as it had fewer errors to begin with due to their short length (174 bp). In comparison, the longer constructs (#3 and 4) had lower percentages of correct colonies to begin with (~500 bp, ~55–60% positive), but the effect of ECR was more obvious, reaching >90% positive after two iterations (**Figure 10B**).

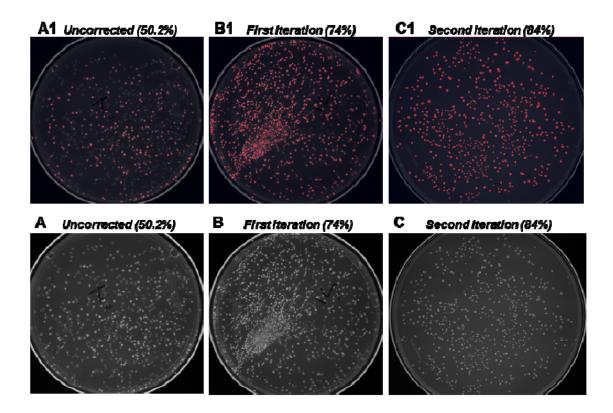


Figure 11: Employing error correction increases the percentage of fluorescent RFP colonies. Images below are examples of the increased fluorescent population derived by employing ECR with Surveyor nuclease. Image A1/A show colonies derived from the uncorrected RFP synthesis product that only yields 50.2% fluorescing colonies. Iterations of correction, using 20 min incubations in this case, yield colonies with an increased fluorescent population as shown in B1/B and C1/C. Images A1-C1 are the same images as A-C with an added pseudo-colored red mask to highlight the brightly fluorescent colonies.

Results from DNA sequencing analysis of randomly selected colonies correlated with the observations made with the colony counting experiments and revealed more details on the correction efficiency of different types of errors. The results in **Table 1** showed that ECR with Surveyor was very efficient in reducing errors arising from deletion and insertion events. Most deletion and insertion type of errors could be eliminated after one round of 60-min treatment or two rounds of 20-min treatment.

Surveyor treatment was also effective toward substitutions albeit with reduced efficiency. Substitution types of errors were still present after two rounds of 60-min incubations.

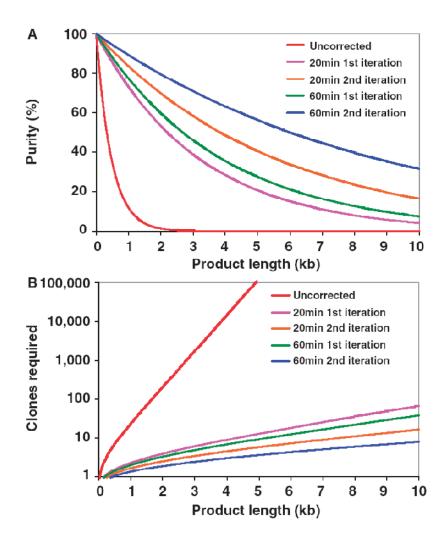


Figure 12: Predicted effects of ECR as a function of sequence length. (A) Purity of gene synthesis products (percentage of error-free clones) decreases exponentially with the length of the product synthesized. Employing ECR (1 error in 8701 bp, blue line) dramatically increases the probability of locating an error-free clone than the uncorrected population (1 error in 526 bp, red line). (B) Employing ECR significantly reduces the number of colonies that need to be screened to have a high (95%) probability of obtaining at least one error-free clone. Two iterations of 60 min cleavage incubations with Surveyor (blue line) could yield a correct 10 kb product by

# sequencing eight random clones. Plots are derived from the result of model calculations as described in the text.

For the purpose of developing the most efficient ECR procedure, data in **Table 1** indicated that increasing incubation time from 20 to 60 min reduced error frequency from 0.31 to 0.26 error/kb (~16% reduction); while adding another round reduced error frequency from 0.31 to 0.18 error/kb for 20-min incubations (~42% reduction) and from 0.26 to 0.11 error/kb for 60-min incubations (~58% reduction). It appeared that adding a second round of ECR was more effective than increasing the Surveyor incubation time with only one round of ECR, although the accumulative effects of more iterations and longer Surveyor incubation was most dramatic.

Following the model predictions of Carr et al. [60] and Furhmann et al. [61], we performed statistical analysis to better understand the implications of our results. As can be seen in **Figure 12A**, the percentage of gene synthesis products that yield error-free clones decreases exponentially with the length of the product synthesized. Employing ECR for error correction (1 error in 8701 bp for two iterations of 60-min ECR, blue line) significantly increases the probability of locating an error-free clone than without error correction (1 error in 526 bp, red line). From the practitioner's perspective, this means that dramatically fewer clones need to be sequenced (**Figure 12B**). For example, as predicated in **Figure 12B** with ECR, one will have to screen, on average, only 8–10 clones of a 10 kb treated or a single 1 kb clone in order to locate a correct one. The model prediction correlated well with our sequencing analysis results. Analyzing sequencing

data of 77 random colonies from the second iteration of the 60-min ECR, we found 72 of the colonies contained the correct rfp gene. The determined error rate of 0.11/kb meant a >16-fold reduction of errors present in the synthetic pool. With such an improvement, larger DNA targets can be conveniently synthesized and corrected within 2–3 h without resorting to additional cloning or excessive sequencing.

## 3.4 Discussion and conclusions

The method presented here was a collaboration work between me and my labmates Ishtiaq Saaem and Siying Ma. It performs enzymatic error correction on synthetic genes using Surveyor nuclease, which has the broadest substrate specificity toward all types of mismatches as compared to other known mismatch specific binding proteins or endonucleases. The method utilizes the mismatch-specific endonuclease activity of the Surveyor enzyme to cut heteroduplex sequences at the mismatch sites and uses the exonuclease activity of the proof-reading DNA polymerase to remove the mismatch bases, followed by an OE-PCR reaction to re-assemble the cleaved fragments into full-length gene constructs. The results from the current study demonstrate that the optimized ECR procedure is robust and effective for all error types, especially insertions and deletions, yielding superior results than previous methods. The ECR method is probably more suitable for long and error-rich synthetic products and can be performed in less time than MutS-based procedures, which require gel-shift assay and DNA extraction from PAGE. Additionally, in comparison to the commercial ErrASE kit [62],

the ECR reaction mitigates the need for tittering and excessive enzyme usage. Using the protocol developed in the current study, two ECR iterations could be completed in <5 h and reduces the error frequency by >16-fold. Future research to improve ECR may involve increasing its efficiency toward substitution types of errors.

# 4. High-throughput protein expression screen strategy

## 4.1 Introduction

Low-cost, high-throughput gene synthesis and precise control of protein expression are of critical importance to synthetic biology and biotechnology [63-65]. Here we describe the development of an on-chip gene synthesis technology, in which inkjet oligonucleotide array synthesis, isothermal oligonucleotide amplification and parallel gene assembly were combined for the first time on a microchip. With this newly developed technology preferably in use, we synthesized large quantities of oligo libraries, assembled them into gene libraries, and designed a high-throughput platform for assessing the expression potential of a protein and for reliably obtaining synthetic gene sequences with desired protein expression levels using these gene libraries. The high-throughput protein expression strategy should meet the following criteria: a) target-independent: the strategy should be suitable for any target protein; b) efficient and high-throughput: the strategy should be able to screen a relatively large (>108) expression library in a short period of time (i.e. hours); c) stringent: the strategy should cleanly separate high-expression clones from low/none-expression ones; d) simple and economical: can be carried out in a standard academic research lab with standard equipment.

For years, researchers have tried to characterize and employ sequences that locate in the regulating regions of the target proteins are underway in many synthetic

biology laboratories [66-69]. However, the design process that needs more attention is the treatment of coding sequence, because if coding sequence itself is poorly translatable in a given host environment, recoding it with synonymous codons seems to be the last resort. Some efforts have been done to improve coding sequences, mainly by altering synonymous codons at specific regions by applying one or a few rules [64, 70-74]. However, none of the rules has reliably improved expression; the tendency has been to layer more and more rules on top of each other, greatly complicating the gene design task and creating problems of prioritization when applying several conflicting principles. A robust and generally applicable gene design method must instead be based on well-established relationships that are validated by thorough experimentation. The ability to create synthetic gene sets with variations in synonymous gene coding will be essential to elucidate these relationships.

In order to better design genes, optimize protein expression and systematically study the relationship between codon usage and gene expression, we developed a technology platform for synthesizing a cDNA library that contains large numbers of designed codon combinations for any target protein. The codon choice used in guiding a design of such a 'codon-optimization' library can be totally unbiased or biased toward a specific host organism. Such a codon-optimization library was used for screening cDNA sequences that give high expression levels under specific conditions. In order to separate high-expression clones from low or none-expression ones in the codon-optimization

expression library, we first studied the distribution of protein expression levels from large numbers of synthetic genes, all encoding the same protein, or "codon variants". A convenient target for this study is the  $lacZ\alpha$  gene, a well-established reporter gene whose expression makes the host  $E.\ coli$  cells turn blue in the presence of isopropyl- $\beta$ -D-thiogalactopyranoside (IPTG). The synthetic codon variants were designed using an unbiased codon usage table, in which all codons representing an amino acid are present with equal frequency. For each amino acid residue along the polypeptide sequence, a synonymous codon was randomly assigned. The assembled gene library was cloned into pUC19 vector using CPEC cloning method developed in Chapter 1. The expression profile of all the colonies produced from this library was studied to see if we could achieve various expression levels among this library and then extend this platform to more meaningful applications.

In addition, to make this approach generally applicable to most proteins, the target genes need to be tagged with a reporter gene, such as the green fluorescent protein (GFP), which allows for direct measurement of their expression levels. We plan to apply this strategy to optimizing the expression of a batch of 74 Drosophila transcription factor (TF) protein domains that are going to be used for generating antibodies for high-throughput genomics studies [75]. To determine the effectiveness of this approach, the first 15 candidates that had been found not to express in *E. coli* were tested. A library of synthetic codon variants was constructed for each candidate protein.

The synthetic genes were fused to the N-terminus of GFP and cloned into the pAcGFP1 expression vector using CPEC [44]. The plasmid libraries were transformed into *E. coli* cells and cultured on agar plates. GFP fluorescence from all colonies was monitored during a 20-hour period at 37°C. We selected a small number of highly fluorescent colonies for sequencing and then examined the characteristics of these gene variants. For comparison, the wild-type controls were cloned in the same vector and cultured under the same conditions to compare with their 'codon-optimized' counterparts. After 15 non-expressing candidates were tested, the same experimental codon optimization procedure was performed for the rest of the 74 candidates in a streamlined fashion. Furthermore, we designed a way to free the GFP fusion partner from the highly-expressed TF moiety if achieved after screening, because removing the GFP fusion partner was desirable for obtaining unique and pure antigen proteins.

In addition, the use of a mismatch-specific endonuclease for error correction resulted in an error rate of ~0.19 errors per kb. We applied this approach to synthesize pools of thousands of codon-usage variants of  $lacZ\alpha$  and 74 challenging Drosophila protein antigens that were then screened for expression in *Escherichia coli*. In one round of synthesis and screening, we obtained DNA sequences that were expressed at a wide range of levels, from zero to almost 60% of the total cell protein mass. This technology may facilitate systematic investigation of the molecular mechanisms of protein

translation and the design, construction and evolution of macromolecular machines, metabolic networks and synthetic cells.

A key technical barrier to optimizing codon usage is the inability to synthesize genes at sufficiently low cost and high throughput. Such a capability would enable many gene and genome variants to be synthesized to explore the vast protein coding space [64]. High-throughput gene synthesis technology has been driven by recent advances in DNA microarrays that can produce pools of up to a million oligonucleotides for gene assembly [50, 63, 76-78], albeit in minute quantities (~105–106 molecules per sequence). The presence of too many oligo sequences in a pool makes it difficult to effectively use the entire oligo pool for gene assembly, as similar sequences can cross-hybridize. Practical solutions include more efficient assembly strategies [62, 78], selective amplification of oligos [62] or, as we do here, physical division of the oligo pool.

#### 4.2 Materials and methods

# 4.2.1 Oligonucleotide synthesis

Chip oligos were synthesized using a custom-made inkjet DNA microarray synthesizer on embossed cyclic olefin copolymer (COC) chips [56]. Gene construction oligos were designed to be 48 or 60 bases long with a 25-base adaptor at the 3' end, which provided a nicking site and anchored the oligo to the chip surface. In the current designs, COC slides were pre-patterned to form 8 or 30 subarrays of silica thin-film spots 150 µm in diameter and 300 µm in inter-feature spacing (center to center). Each

chamber in the 30-chamber design could print 361 spots and was used to synthesize only one gene or gene library up to 0.5–1 kb in length. Multiple spots were used to synthesize one oligo sequence.

# 4.2.2 Combined nSDA-PCA reaction for on-chip oligo amplification and gene assembly

The chambers on the printed COC slides were filled with the nSDA-PCA reaction cocktail containing 0.4 mM dNTP, 0.2 mg/ml BSA, Nt.BstNBI, Bst large fragment and Phusion polymerase in optimized Thermopol II buffer. The slides with sealed chambers were placed on the slide adaptor of a Mastercycler Gradient thermocycler (Eppendorf) to carry out combined nSDA-PCA reactions. nSDA involved incubation at 50 °C for 2 h followed by 80 °C for 20 min; the subsequent PCA reaction involved an initial denaturation at 98 °C for 30 s, followed by 40 cycles of denaturation at 98 °C for 7 s, annealing at 60 °C for 60 s and elongation at 72 °C for 15 s/kb, and finished with an extended elongation step at 72 °C for 5 min.

After the nSDA-PCA reaction, 1–2  $\mu$ l of the reaction from each chamber was used for PCR amplification with Phusion polymerase. The PCR reaction involved an initial denaturation at 98 °C for 30 s, followed by 30 cycles of denaturation at 98 °C for 10 s, annealing at 60 °C for 60 s and elongation at 72 °C for 15 s/kb, and finished with a final elongation at 72 °C for 5 min.

# 4.2.3 Enzymatic error correction

Chip-synthesized genes diluted in 1× Taq buffer were denatured and reannealed by incubating at 95 °C for 2 min before cooling down first to 85 °C at a rate of 2 °C per second and then to 25 °C at 0.1 °C per second. The reaction (4  $\mu$ l) was mixed with 1  $\mu$ l of the Surveyor nuclease reagents (Transgenomic) and incubated at 42 °C for 20 min. The product (2  $\mu$ l) was PCR amplified, cloned and sequenced.

## 4.2.4 Construction of a synthetic *lacZα* gene library

 $lacZ\alpha$  gene library was assembled from the oligo libraries synthesized on Dr. Oligo DNA synthesizer in house which have codon variants designed using an unbiased codon usage table, in which all codons representing an amino acid were present with equal frequency. The  $lacZ\alpha$  library was cloned into pUC19 linear vector using CPEC and the cloning product was transformed into GC5 competent cells. The transformants were spread evenly on 15-cm agar plates supplemented with carbenicillin antibiotics. To compare the expression profile of this gene library with an un-modified control  $lacZ\alpha$  gene, a control plasmid pUC19ctl was made by deleting MCS from the pUC19 vector using one of the established mutagenesis methods. The control plasmid was cloned into the same vector, transformed and plated under the same conditions with the library plasmids.

# 4.2.5 Evaluation of the expression levels of synthetic $lacZ\alpha$ gene library

We spread 150-mm LB agar plates evenly with transformed *E. coli* cells and incubated overnight at 37 °C. One of the plates from both the library and control group was placed side by side in a scanner and scanned every 15 min for 48 hours in the 37°C incubator. The shade of blue of each colony on both plates was recorded in this way and 192 scanned pictures were used for  $lacZ\alpha$  expression analysis since the deepness of blue is a fine indicator of the expression level of the lacZ $\alpha$  gene. The scanned pictures of both library and control lacZ $\alpha$  plates were subjected to statistical analysis. Cell Profiler software was used to identify and record the location of each colony on both plates as well as its color intensity on every time point. Bacterial colonies were then identified as a set of objects ranging from 2 to 30 pixels in diameter on scanned images. An automatic thresholding method using a mixture of Gaussians was used to identify local maxima [79]. The images were converted to grayscale and pixel intensities were inverted. From the set of pixels located in each colony, ten pixels with the maximum intensities were selected and averaged to give an estimate of colony color intensity.

# 4.2.6 Plasmid library construction using CPEC method

The commercial vector pAcGFP1 was modified by inserting a His6-tag immediately after the start codon and a TVMV cleavage site (ETVRFQS) in front of the GFP gene (**Figure 13**). The modified vector was linearized by PCR to add overlapping end sequences with the insert. Transcription factor open reading frames were cloned

into the vector using the CPEC cloning method [44, 53]. Briefly, 250 ng of the linear vector was mixed with inserts at 1:2 molar ratio in a 25  $\mu$ l CPEC reaction using Phusion polymerase. The reaction involved ten cycles of denaturation at 98 °C for 10 s, annealing at 55 °C for 30 s and extension at 72 °C for 15 s, and finished with an extended elongation step at 72 °C for 5 min. We used 4  $\mu$ l of the cloning product for direct transformation of *E. coli*.

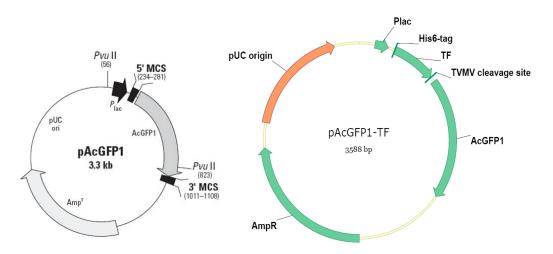


Figure 13: The expression system map of pAcGFP1-TF. Inserts were subcloned into the 5'MCS and fused in frame to the N-terminum of AcGFP1.

## 4.2.7 Protein expression screen

 $E.\ coli$  libraries of codon variants were cultured on LB agar plates containing 100 µg/ml carbenicillin. On each plate with 1,000–1,500 colonies, 1–10 colonies with the highest GFP signals were selected and cultured overnight in Luria Broth at 37 °C with shaking. The saturated culture was diluted 1:50 in the same media and grown at 37 °C until mid-log phase (A600 = 0.5), when the temperature was shifted to 30 °C and 1 mM final concentration of IPTG was added. After another 4 h, 10 ml of each culture was

centrifuged and the cell pellet was resuspended in 1× NuPAGE LDS Sample Buffer (Invitrogen). After the samples were heated at 90 °C for 5 min and centrifuged at 14,000g for 10 min, aliquots of the supernatant were analyzed by SDS-PAGE using a NuPage 4–12% gradient gel (Invitrogen) and stained with EZBlue Gel Staining Reagent (Sigma).

# 4.2.8 Cleavage and purification of transcription factor-GFP fusion proteins

For intracellular processing of transcription factor–GFP fusion proteins, *E. coli* cells co-transformed with an optimized transcription factor–GFP plasmid and the pRK1037 vector containing the TVMV protease gene were grown in 2 ml of Luria Broth with 100 µg/ml carbenicillin and 30 µg/ml kanamycin at 37 °C overnight. The saturated culture (1 ml) was added into 500 ml of the same medium and grown at 37 °C to mid-log phase ( $A_{600} = 0.5$ ), when the temperature was shifted to 30 °C and IPTG was added to a final concentration of 1 mM. After another 4 h, the cells were harvested by centrifugation.

To purify His6-tagged transcription factor proteins, the cell paste was resuspended in 1×LEW Buffer (USB) and lysed by mixing with 1 mg/ml lysozyme for 30 min followed by sonication. The cell lysate was centrifuged at 10,000 g for 30 min at 4 °C to pellet the insoluble material. The supernatant was transferred to a clean tube for loading on PrepEase Ni-IDA column (USB) under native condition. The insoluble material was resuspended in 1× LEW buffer and centrifuged at 10,000 g for 30 min at 4 °C. The cell pellet was then resuspended in 1× LEW denaturing buffer (USB) and kept on

ice for 1 h with occasional stirring to dissolve the inclusion bodies. The suspension was then centrifuged at 10,000 g for 30 min at 4 °C to remove any remaining insoluble material. The supernatant was transferred to a clean tube for loading on PrepEase Ni-IDA column (USB) under denaturing condition following kit instructions.

#### 4.3 Results

To effectively use all the oligos synthesized on a microarray, we divided the whole microarray into subarrays, each containing only the oligos that are needed to assemble a longer DNA molecule of about 0.5–1 kb in total length. Subarrays were physically isolated from the rest of the chip by being located in individual wells, eliminating the need for post-synthesis partitioning of the oligo pool. Oligos were synthesized on an embossed plastic microchip using a custom-made inkjet DNA microchip synthesizer [56]. The printing area in each subarray was patterned with 150-µm spots of silica thin film to reduce 'edge-effects', which could lead to poor oligo synthesis [80]. Our design allowed a standard 1" × 3" chip surface to be divided into as many as 30 subarrays, each containing 361 silica spots for synthesizing a unique DNA oligonucleotide sequence. With the setup used in this study, 10,830 different 85-mer oligo sequences could be synthesized on a single chip, providing a capacity to produce up to ~30 kb of assembled DNA.

We next sought to achieve additional increases in throughput by integrating oligo synthesis with amplification and gene assembly on the same chip. In previous

work, chemical methods such as NH4OH treatment have been used to cleave oligos from the chip for subsequent off-chip gene assembly reactions [50]. Progress toward automating and miniaturizing the subsequent gene assembly reactions has been reported using microfluidics, resulting in reduced costs and reagent consumption [81]. Here we first use isothermal nicking and a strand displacement amplification reaction (nSDA) to amplify oligos from the microarray surface, followed by polymerase cycling assembly (PCA) reaction in the same chamber (Figure 14). Briefly, 60-mer gene construction oligo sequences are synthesized with a 25-mer universal adaptor added at the 3' end, which is anchored on the chip surface. This adaptor contains a nicking endonuclease recognition site (Appendix A). After array synthesis, a universal primer (Appendix A) hybridizes to the adaptor and initiates continuous elongation and nicking on the extending strand. This is catalyzed by a combination of a strand-displacing polymerase (e.g., Bst large fragment) and a nicking endonuclease (e.g., Nt.BstNBI). The amplification is linear so as to keep the ratios constant among amplified oligos. The extent of the amplification is adjusted by controlling the reaction time. We estimate that a ~2 h reaction time results in an approximately fourfold amplification.

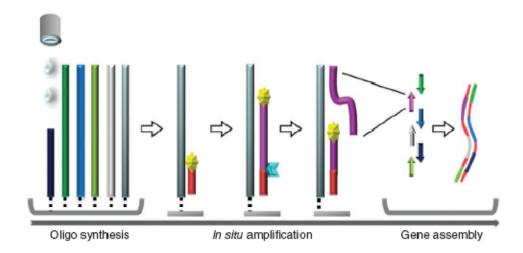


Figure 14: The integrated on-chip oligo array synthesis, amplification and gene assembly process. Small pools of oligos are synthesized in separate chambers on a plastic DNA microchip using an inkjet DNA microarray synthesizer. The chambers are then filled with a combined amplification and assembly reaction mixture and sealed. In a nicking and strand displacement amplification reaction, a DNA polymerase (Bst large fragment, shown in yellow) extends and displaces the proceeding strand while a nicking endonuclease (Nt.BstNBI, shown in teal) separates the construction oligos from the universal primer (in red) and generates new 3'-ends for extension. After amplification, the free oligos in each chamber are assembled into gene products by polymerase chain assembly.

To avoid complex microfluidic manipulations that would otherwise be required to collect and purify the amplified oligos for downstream gene assembly reactions, we designed the gene-assembly reaction cocktail to allow the PCA reaction to take place immediately after nSDA without a buffer change. After appropriate concentrations of the amplified oligos were accumulated by nSDA, the reaction mode was switched from isothermal amplification to thermal cycling, which results in assembly of the amplified oligos into gene fragments in the same reaction chamber. The gene products were further amplified off-chip by PCR (**Figure 15**). The size range of the combined nSDA-

PCA reaction products is currently set at 0.5–1 kb for overall throughput and assembly efficiency considerations. Longer sequences can be hierarchically assembled from these 0.5–1 kb building blocks.

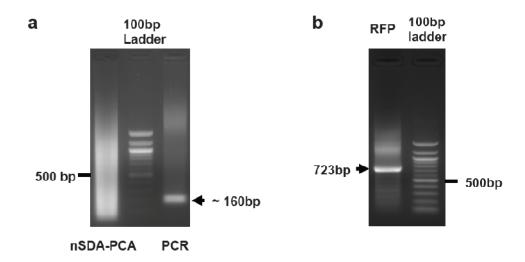


Figure 15: Assembly of target genes by on-chip nSDA-PCA reactions. (a) Agarose gel image of the nSDA-PCA reaction product showing as a typical smear (left lane) and the PCR amplified *lacZα* gene product (right lane). The middle lane is 100-bp DNA ladder. (b) Assembly of the Red Fluorescent Protein (RFP) gene by on-chip nSDA-PCA reaction followed by PCR amplification.

To reduce gene synthesis errors, we developed a simple yet effective errorcorrection method using the plant CEL family of mismatch-specific endonucleases,
which have been shown to recognize and cleave all types of mismatches arising from
base substitutions or from small insertions or deletions. A commercial source of a
subtype of the CEL enzymes was the Surveyor nuclease, which has been used primarily
for mutation detection [59]. To use it for error correction, we first denature by heat and
reanneal the synthetic genes, and then treat them with Surveyor nuclease to cleave error-

containing heteroduplexes at the mismatch sites. The error-free DNA duplexes remain intact and are amplified by overlap-extension PCR.

Table 2: Error frequencies as determined by sequencing in chip-synthesized RFP genes with and without error correction using Surveyor nuclease. Clones were randomly selected from each population and sequenced from both directions.

				Total	Bases	Error Frequency
	Deletions	Insertions	Substitutions	Errors	Sequenced	f (per kb)
Before						
correction	43	4	10	57	29,958	1.9
Surveyor						
correction	6	0	48	54	291,180	0.19

To test the effectiveness of this approach, we cloned chip-synthesized genes encoding RFP into an expression vector with and without Surveyor nuclease treatment. We performed sequencing and automated fluorescent colony-counting experiments to determine and compare error frequencies. By Sanger sequencing 470 randomly selected clones, we observed error frequencies of 1/526 bp (or ~1.9 errors per kb) and 1/5,392 bp (or ~0.19 errors per kb) before and after Surveyor nuclease treatment, respectively (**Table 2**). Automated counting of thousands of colonies showed that ~50% and 84% of the RFP colonies were fluorescent in untreated and Surveyor nuclease—treated populations (**Figure 16A**). The results of the sequencing and the colony counting experiments correlated well according to statistical analysis (**Figure 16B**). Another study published while this manuscript was being revised reported comparable error frequencies using the commercial ErrASE kit [62].

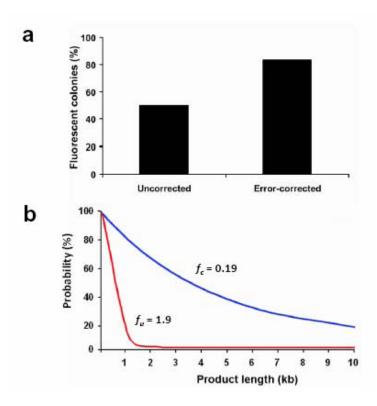


Figure 16: Statistical evaluation of errors in synthetic RFP genes with and without Surveyor nuclease treatment. (a) Percentage of fluorescent RFP colonies with and without error correction using Surveyor nuclease. On-chip RFP gene synthesis without error correction resulted in 50.2% fluorescent colonies while those treated with the Surveyor nuclease yielded a 84% fluorescent population. The total number of colonies in each population was approximately 3,000. (b) Predicted correlation of the probability for an error-free clone with product length [61] before and after error correction with Surveyor nuclease. Error correction using Surveyor nuclease (error frequency fc = 0.19 per kb, blue line) increases the probability of locating an error-free clone than the uncorrected population (error frequency fu = 1.9 per kb, red line), thereby drastically reducing the number of colonies that need to be screened. The error frequencies were calculated from sequencing data in Table 2.

To apply high-throughput gene synthesis to optimize protein expression, we studied the distribution of protein expression levels of a large number of synthetic genes that all encode the same protein, called 'codon variants'.  $LacZ\alpha$  was used as an example in this study. Expression of  $lacZ\alpha$  makes the host  $E.\ coli$  cells turn blue in the presence of

isopropyl-β-d-thiogalactopyranoside (IPTG). First, we designed synthetic codon variants using an unbiased codon usage table, in which codons representing an amino acid were used with equal frequency (**Appendix A**). Then, we constructed a library of  $lacZ\alpha$  codon variants and transformed the variants into *E. coli* competent cells. We plated a small fraction of the library on solid agar and measured the blue color intensity of the individual colonies in real time by automated image analysis. Clones representing a full spectrum of protein translation levels could be readily identified with fine shades of differences in protein expression (Figure 17A). Notably, we observed a bell-shaped distribution of the maximum protein expression levels of random codon variants growing on the plate (**Figure 17B**). Approximately one-third of the variants showed higher expression levels than wild-type  $lacZ\alpha$ . The expression level of the wild-type gene was slightly above the median level of all the clones with measurable expressions. Although understanding the causes and implications of this distribution requires further study, the distribution allowed us to estimate the translational potential of the  $lacZ\alpha$ gene in E. coli, which is indicated by the upper boundary in the quantile box plot (**Figure 17B**). These observations suggest the feasibility of an experimental approach to reliably obtain gene sequences with the desired protein expression levels in a given expression system.

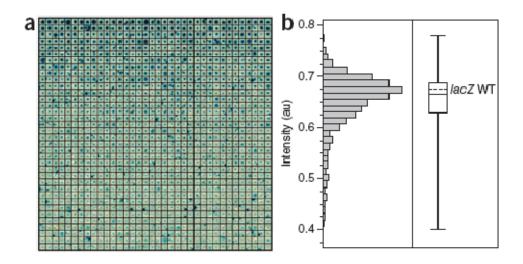


Figure 17: Expression of synthetic  $lacZ\alpha$  codon variants in E.~coli. (a) A set of 1,296 E.~coli colonies expressing distinct  $lacZ\alpha$  codon variants sorted by color intensity. Raw images were acquired by scanning an agar plate on the scanning window of a HP Photosmart C7180 Flatbed Scanner. (b) Bar graph and box plot showing distribution of color intensities of a different set of 1,468 random colonies expressing distinct  $lacZ\alpha$  codon variants on an agar plate. Owing to the large size of the synthetic codon variant library, the chance of having identical clones on a plate was extremely low, as confirmed by sequencing several hundred blue colonies (data not shown). In the box plot, the expression level of the WT  $lacZ\alpha$  is marked with a dash line.

Next we describe the successful development of such an optimization approach in *E. coli*, which has been a workhorse for expressing a variety of proteins for research and industrial applications. To allow direct measurement of protein expression levels, we tag each target gene with a GFP reporter gene. Proteins expressed at higher levels should result in colonies with brighter fluorescence.

We applied this strategy to optimize the expression of 74 Drosophila transcription factor protein domains to be used for generating antibodies for the ENCODE (ENCyclopedia Of DNA Elements) Project [82]. We first tested the approach

on 15 candidates that were not expressed in *E. coli* (N.N. & K.P.W., unpublished data). Libraries of synthetic codon variants were designed based on an *E. coli* codon-usage table [83] and constructed using our on-chip gene synthesis technology (**Figure 18**). The enzymatic error correction procedure was not performed here because heteroduplexes might form between closely related codon variants. The synthetic genes were fused to the N terminus of GFP and cloned into the pAcGFP expression vector using the sequence-independent circular polymerase extension cloning method (CPEC) [44]. *E. coli* cells were transformed by the plasmid libraries and cultured on agar plates. GFP fluorescence from all colonies was monitored continuously and a small number of highly fluorescent colonies were selected from each pool for sequencing. All colonies contained plasmids with different codon usages throughout the sequence of the candidate proteins.

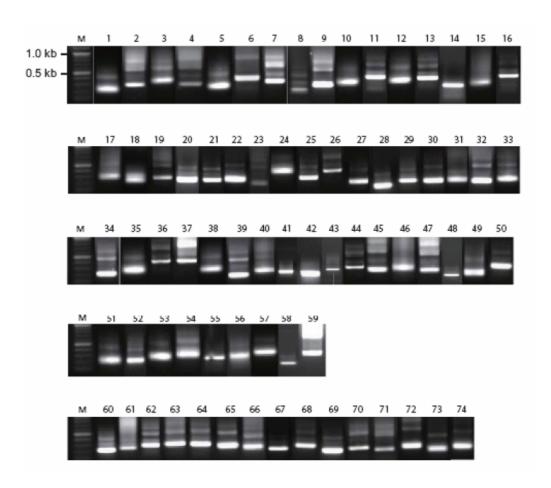


Figure 18: Compiled images of agarose gel electrophoresis of PCA assembled and PCR amplified codon libraries of 74 Drosophila transcription factor gene fragments. The lengths of the gene fragments fall in the range of ~ 0.3-0.5 kb. Lane M is 100-bp DNA ladder. The wild-type gene sequences are listed in Appendix A.

The sequence-confirmed, highly fluorescent colonies were cultured individually in liquid media and the expression of the protein domains was measured by running the total protein extracts on polyacrylamide gels. High-expression clones were identified for all 15 candidates using this strategy (**Figure 19** and **Appendix A**). In comparison, the wild-type controls cloned into the same vector and cultured under the same conditions showed undetectable protein expression. This result indicates that this method has the

capability to reliably increase protein expression from an undetectable level to as high as representing ~50–60% of the total cell protein mass.

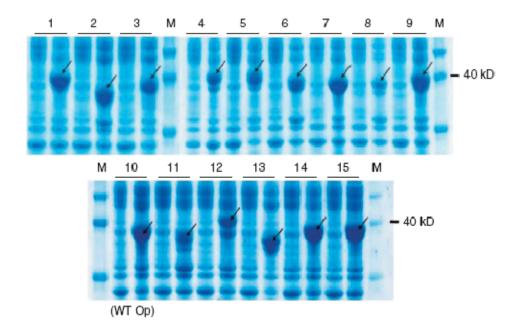


Figure 19: Optimization of protein expression. Seventy-four Drosophila transcription factor gene fragments were optimized for production in *E. coli* by synthesizing ~1,000–1,500 codon variants of each, cloning them in frame with GFP and screening for the colonies with the highest fluorescence. Data for 15 proteins shown here (see Figure 20 for remaining 59). Each pair of lanes shows total cell protein extract of *E. coli* expressing the wild-type (left lane, WT) and optimized (right lane, Op) clones. The broad bands marked by an arrow represent highly expressed transcription factor–GFP fusion proteins. There was no detectable expression of wild-type transcription factor–GFP fusion proteins as shown in the wild-type lanes. Equal amounts of the total cell protein extracts were separated on NuPage 4–12% gradient gels and stained with EZBlue Gel Staining Reagent. M lanes are Novex Prestained protein standards (Invitrogen).

Encouraged by the high success rate, we performed the same experimental codon optimization procedure for the remaining 59 proteins. Sequencing and protein gel results confirmed that we were able to predictably obtain high-expression clones for all

candidates tested (**Figure 18** and **Figure 20A**). Wild-type and optimized gene fragment sequences are listed in Supplementary Sequences. Calculation of codon adaptation index (CAI) [20], which measures synonymous codon usage bias, for each sequence indicates that the average index of the selected, highly expressed synthetic sequences is slightly higher ( $0.756 \pm 0.041$ ) than that of the non-expressing wild-type sequences ( $0.663 \pm 0.047$ ) (**Table 3** and **Table 4**), suggesting a certain degree of correlation between CAI and protein expression level. The highly expressed transcription factor moiety could be freed from the GFP fusion partner by in vivo cleavage with a co-expressed TVMV protease and purified with an Ni-IDA column (**Figure 20B**). Removing the GFP fusion partner is desirable for obtaining unique and pure antigen proteins.

Table 3: Comparison of CAI values of 15 expression optimized TF sequences (CAI-opt) vs. wild-type non-expressing sequences. CAI value was calculated using CAIcal server at <a href="http://genomes.urv.es/CAIcal">http://genomes.urv.es/CAIcal</a> [84].

Name	CAI-opt	CAI-wt
AB1	0.706	0.583
AB11	0.795	0.709
AC12	0.681	0.681
AF4	0.833	0.635
AG3	0.738	0.707
AR1	0.753	0.661
B11	0.729	0.601
D5	0.741	0.699
F9	0.774	0.602
K1	0.821	0.639
К3	0.777	0.711
K4	0.717	0.618
K5	0.739	0.683
L5	0.769	0.688
M6	0.762	0.735

Average	0.756	0.663	
SD	0.041	0.047	

Table 4: Comparison of CAI values of the rest 59 expression optimized TF sequences (CAI-opt) vs. wild-type sequences.

Name	CAI-opt	CAI-wt
bcd_d2	0.684	0.63
cad_d1	0.766	0.651
hb_d2	0.753	0.673
lab_d1	0.805	0.715
lab_d2	0.782	0.643
pb_d2	0.726	0.673
Dfd_d1	0.806	0.689
Scr_d2	0.74	0.598
Antp_d1	0.736	0.738
lid_d2	0.793	0.618
lilli_d1	0.742	0.597
lilli_d2	0.786	0.668
E75_d1	0.797	0.698
E78_d1	0.761	0.683
E78_d2	0.793	0.625
DHR3_d1	0.76	0.626
DHR3_d2	0.774	0.704
EcR_d1	0.743	0.555
EcR_d2	0.802	0.613
DHR78_d1	0.801	0.636
Dis_d1	0.721	0.658
Dis_d2	0.692	0.548
ERR_d1	0.706	0.701
DHR38_d2	0.8	0.699
ftz-f1_d1	0.776	0.621
DHR39_d1	0.771	0.614
DHR39_d2	0.704	0.546
DHR4_d1	0.707	0.55
DHR4_d2	0.704	0.553
BRC_d1	0.771	0.688

Name	CAI-opt	CAI-wt
BRC_d2	0.82	0.679
E74_d1	0.798	0.601
E74_d2	0.804	0.674
E93_d1	0.738	0.695
E93_d2	0.807	0.731
mld_d1	0.816	0.64
salm/salr_d1	0.748	0.65
salm/salr_d2	0.772	0.693
ac_d1	0.785	0.593
ac_d2	0.715	0.683
sc_d1	0.803	0.589
l(1)sc_d1	0.817	0.601
l(1)sc_d2	0.726	0.669
ase_d1	0.78	0.578
Dsx_d1	0.751	0.637
Dsx_d2	0.746	0.596
Ovo/Svb_d1	0.758	0.697
Ovo/Svb_d2	0.771	0.749
dFOXO_d2	0.703	0.631
ey_d1	0.736	0.639
ey_d2	0.701	0.652
toy_d1	0.768	0.576
toy_d2	0.693	0.636
Stat92E_d2	0.802	0.645
Rx_d1	0.746	0.658
hbn_d1	0.756	0.688
otp_d1	0.742	0.683
dwg_d1	0.799	0.659
dwg_d2	0.789	0.7
Average	0.757	0.64
7110.480		

The integration of oligo synthesis and gene assembly on the same microchip facilitates automation and miniaturization, which leads to cost reduction and increases in throughput. On our current chip, each of the 30 chambers was used to synthesize one gene fragment up to 1 kb in length with a 9× redundancy in oligo usage (9 spots in one subarray were used to synthesize each oligo sequence). The estimated cost of chipoligonucleotide synthesis for this ~30 kb of sequence was < \$0.001/bp of final synthesized sequences, which is one-tenth of the lowest reported cost [62]. The cost estimation was calculated like this: since in the current design, a regular sized chip (1' x 3') with 30 chambers is capable of synthesizing 10,830 different oligo sequences and a single chip is capable of producing 30,000-bp of double-stranded gene sequences in one round of synthesis using each gene synthesis chamber to assemble only one gene fragment up to 1 kb in length (a 9x redundancy in oligo usage), the total cost of oligo synthesis per chip, including phosphoramidites, other synthesis chemicals, organic solvents, gases, and the COC chip, is under \$30. Therefore, the estimated cost of chip-oligonucleotide synthesis would be approximately < \$30/30,000bp = \$0.001/bp of final synthesized sequences.

Furthermore, including enzymatic processing and error correction, the average cost of integrated gene synthesis on a chip is < \$0.005/bp of final synthesized gene sequences with an error frequency of <0.2 error/kb. Additional cost after oligo microarray synthesis comes from enzymatic reactions for oligo amplification (nSDA), gene assembly (PCA), gene amplification (PCR) and error correction. The total reaction

volume on a chip containing 30 subarrays is  $4\mu l \times 30 = 120 \mu l$ . The estimated cost of all reagents needed for nSDA, PCA and PCR, including enzymes, dNTPs, and buffers, is approximately \$70 per chip. The estimated cost of error correction using Surveyor nuclease is approximately \$50. Therefore, the estimated total cost of chip-gene synthesis would be approximately under (30+70+50)/30,000bp = 0.005/bp of final synthesized sequences. With multiplexing and more advanced chip design, greater throughput and lower costs are potentially achievable.

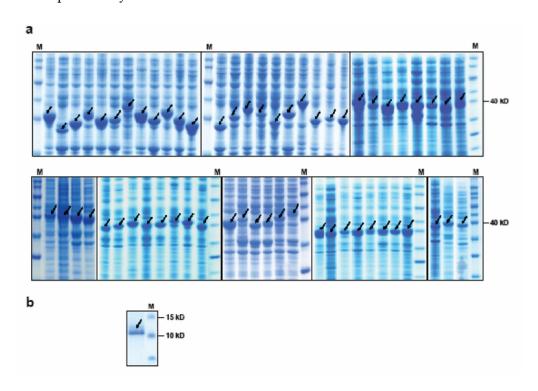


Figure 20: Protein expression of optimized TF genes. (a) SDS-PAGE showing protein expression level of 59 TF genes after codon optimization. Highly expressed TF-GFP fusion protein bands were marked by arrows. Total cell protein extracts were separated on NuPage 4-12% gradient gels (Invitrogen) and stained with EZBlue™ Gel Staining Reagent (Sigma). M lanes are Novex Prestained protein standards (Invitrogen). (b) Intracellular processing of TF-GFP fusion protein and purification of

His-tagged TF antigen. Arrow marked a purified His6-tagged TF protein with GFP fusion partner removed.

#### 4.4 Discussion and conclusions

Protein expression optimization using high-throughput gene synthesis and screening demonstrates a number of advantages over other codon optimization methods, such as testing one design at a time based on unproven design rules. First, our results indicate that a synthetic gene sequence with a desired protein expression level can be selected through one round of synthesis and screening with high confidence. To efficiently identify high-expression clones for a target protein in E. coli, we find that for most of our target gene libraries, screening 1,000–1,500 synthetic codon variants for a target protein seems to be sufficient. The capability to achieve not only the maximum but also intermediate levels of protein expression will be valuable for future synthetic biology applications. Second, the screening-based method does not rely on knowing all the rules of codon usage, which are still not completely known. Incomplete knowledge often leads to wrong predictions using other methods. Third, the screening-based method is faster and cheaper and can be performed on a large scale with highthroughput gene synthesis technology. Unpredictability and repeated trial and error using other methods often leads to substantially increased costs, longer production times and lower throughput. Combining high-throughput on-chip gene synthesis and screening may pave the way for systematic investigation of the molecular mechanisms of protein translation.

# 5. Screening and optimization of synthetic long genes 5.1 Introduction

In the last Chapter, we established a platform for protein expression optimization for theoretically any genes by low-cost and high-throughput gene synthesis and screening. It was identified that for those target genes of 300~500 bp, in order to identify high-expression clones in *E. coli*, screening 1,000~1,500 synthetic codon variants seemed to be sufficient. However, normal gene sizes are usually longer than just 500 bp. According to Xu, et. 2006 [85], the means of the Mean Length of Genic Coding Sequence (MLGCS) for prokaryotes and eukaryotes were 924 bp and 1,346 bp, respectively. For such genes of normal length or even longer genes, optimizing them would require a much larger synthetic codon variant library in order to identify the high-expression clones. And here the library sizes are increasing at an exponential scale as well as synthesis errors. In order to be able to screen and optimize a normal sized gene, the strategy for screening short genes (<500 bp) should be adjusted. For long gene and genome screening, there are fundamentally three obstacles. The first one is the assembly of the long gene library. The second one is to work around the high error rate contained in the synthesized gene libraries in order to find a correct and high-expression clone. The last obstacle is to sequence for the long gene clone to confirm its full identity which can be costly and time-consuming. In order to overcome these obstacles, it seems the key is to reduce the library size to a manageable scale while maintaining the

screening power so the high-expression clones won't be missed. Therefore, splitting long genes into smaller fragments and screening for each one separately and then screening the full-length gene library assembled from high-expression clones for each fragment could be a possible solution.

#### 5.2 Materials and methods

## 5.2.1 Reagents and primers

The reagents used in this chapter include Phusion DNA Polymerase with HF Reaction Buffer (NEB), Taq DNA Polymerase with ThermoPol Buffer (NEB), ExoSAP-It PCR Product Cleanup (Affymetrix) and dNTPs (40mM, NEB). Kits include Qiaquick Nucleotide Removal Kit (QIAGEN), Qiaquick Gel Extraction Kit (QIAGEN) and QIAGEN Plasmid Mini Kit (QIAGEN). Equipments include Dr. Oligo DNA synthesizer (Azco Biotech), vacuum concentrator (Eppendorf Vacufuge), plate reader (Tecan GENios Pro Plate Reader), thermo cycler (MyCycler, Bio-Rad), and a fluorescent light source for visualizing GFP flurescence.

For PCR of cat2, cat2-L (5'- AAC AAT TTC ACA CAG GAA ACA GCT) and cat2-R (5'- CGG TAC CCG GGG ATC C) were used. For PCR of all other genes tested for long gene screening purposes in this chapter, AcGFP-LS (5'- CAA TTT CAC ACA GGA AAC AGC TAT G) and AcGFPinsert-R (5'- CGG TAC CCG GGG ATC CTC) were used. For PCR of the vector pAcGFP1, pAcGFP Fw (5'- GAG GAT CCC CGG GTA CCG) and pAcGFP1st Rv (5'- CAT AGC TGT TTC CTG TGT GAA ATT G) were used. For colony

PCR, GFPSeqF (5'-AGC GCC CAA TAC GCA AAC CG) and pAcGFPSeqRv (5'-CCG TAG GTG GCA TCG CCC) were used. For sequencing, pAcGFPSeqFw2 (5'-ATG CTT CCG GCT CGT ATG) and pAcGFPSeqRv2 (5'-TGC CGG TGA ACA GCT CGG) were used.

# 5.2.2 Oligonucleotides and genes

The genes that were tested for long gene library screening and optimization were MTggps, SAggps, AFggps and cat2. The lengths of these genes were 972 bp, 987 bp, 948 bp and 1290 bp respectively. These genes were chosen with an interest to optimize them for collaboration. The oligos designed for each gene were of 60-mers. The design of the oligo sequences was a library of silent mutations for achieving expression optimization. The average size of these gene libraries was ~10<sup>120</sup>. All oligos were synthesized in house using Dr. Oligo DNA synthesizer. Oligos, after a final cleavage step in ammonia solution, were dried out using a vacuum concentrator. Dried oligos were re-suspended in ddH<sub>2</sub>O and concentrations were measured using a plate reader. Equal amounts of each oligo library from each well were mixed for each of the 6 gene libraries. Oligo mixtures were purified to get rid of chemicals and nucleotides using Qiaquick Nucleotide Removal Kit. The purified oligo mixtures were used as the starting blocks for the following gene library assembly and screening.

# 5.2.3 Polymerase cycling assembly (PCA)

An appropriate amount of oligo mixtures for a gene or gene fragment was mixed together with 0.5 µl Phusion DNA Polymerase, 10 µl Phusion HF Reaction Buffer, 1 µl

dNTPs and ddH<sub>2</sub>O. In a PCR thermo cycler, after an initial 30 s of denaturation at 98°C, 40 cycles which consisted of 10 s of denaturation at 98°C, slow ramping at 0.1°C/s from 70°C to 50°C before annealing at 50°C for 2 m, and extension at 72°C for 15 s were performed. The reaction was ended with an extra 5 m of extension at 72°C.

# 5.2.4 Polymerase chain reaction (PCR)

In a total of 50 µl of volume, 2 µl of each PCA product was mixed with 10 µl Phusion HF Buffer, 2.5 µl of each of two overlapping primers at each end, 1 µl dNTPs, 0.5 µl Phusion DNA Polymerase and ddH<sub>2</sub>O. In a PCR thermo cycler, after an initial 30 s of denaturation at 98°C, 30 cycles which consisted of 10 s of denaturation at 98°C, annealing at 56°C for 2 m, and extension at 72°C for 15 s per kb were performed. The reaction was ended with an extra 5 m of extension at 72°C.

# 5.2.5 Plasmid library construction using CPEC method

Linearized vector pAcGFP was obtained by PCR from a modified commercial vector pAcGFP1 which was inserted with a stop codon in the MCS region in front of the GFP gene to make it a negative control. The primers for linearizing vector pAcGFP1 had overlapping end sequences with the insert libraries. Target gene open reading frames were cloned into the vector using the CPEC cloning method [44, 53]. Briefly, 200 ng of the linear vector was mixed with inserts at 1:2 molar ratio in a 50  $\mu$ l CPEC reaction using Phusion polymerase. The reaction involved 20 cycles of denaturation at 98 °C for 10 s, annealing at 56 °C for 2 min and extension at 72 °C for 15 s (MTggps, SAggps and

AFggps) or 30 s (cat2), and finished with an extended elongation step at 72 °C for 5 min.  $4 \mu l$  of the cloning product was used for direct transformation of *E. coli*.

#### 5.2.6 Plasmid PCR

In a total 50  $\mu$ l volume, the plasmid mixture was mixed with 10  $\mu$ l Phusion HF Buffer, connector primer, 2.5  $\mu$ l of the non-degenerate primer at the other end, 1  $\mu$ l dNTPs, 0.5  $\mu$ l Phusion DNA Polymerase and ddH<sub>2</sub>O. In a PCR thermo cycler, after an initial 30 s of denaturation at 98°C, 30 cycles which consisted of 10 s of denaturation at 98°C, slow ramping at 0.1°C/s from 72°C to 53~56°C before annealing at 53~56°C for 2 min, and extension at 72°C for 15 s per kb were performed. The reaction was ended with an extra 5 min of extension at 72°C.

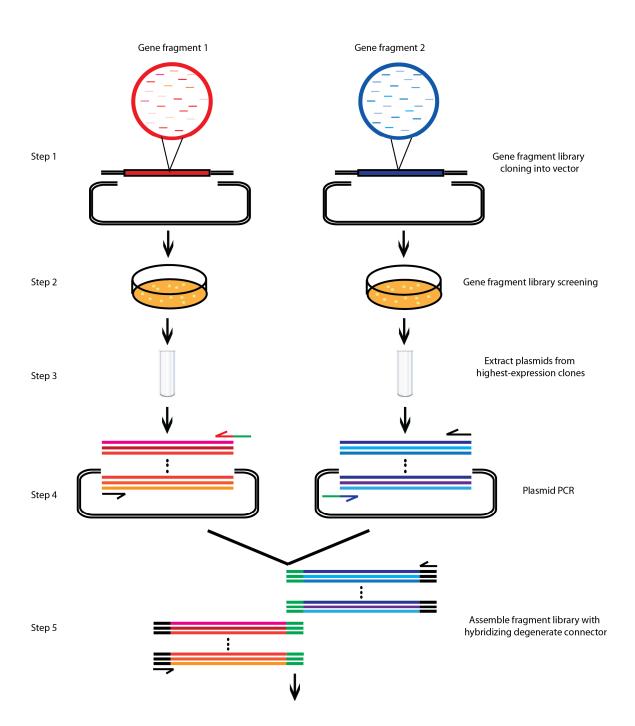
# 5.2.7 Sequencing preparation and sequencing

 $5~\mu l$  of single colony PCR product was mixed with  $2~\mu l$  of ExoSAP-IT reagent for a combined  $7~\mu l$  reaction volume. This was incubated at 37~deg. C for 15~min to degrade remaining primers and nucleotides, and then at 80~deg. C for 15~min to inactivate ExoSAP-IT reagent. The purified scPCR product was then ready for sequencing. For sequencing purposes,  $1~\mu l$  of purified scPCR product was mixed with  $0.5~\mu l$  of one  $10~\mu M$  sequencing primer and  $ddH_2O$  to bring to volume to  $12~\mu l$ . Sequencing was performed by Duke DNA Analysis Facilities using a Perkin Elmer Dye Terminator Cycle Sequencing system.

#### 5.3 Results

Unless screening for short gene libraries of less than 500 bp, which had been demonstrated to be quite easy to assembly and screening after one round of attempt, the bottleneck of screening for long gene libraries, even a normal gene of size ~ 1 kb, would require extensive effort for gene library assembly, error correction and subsequent screenings. In this study, we aimed to develop a simple and convenient method to effectively screen for high-expression clones from a gene library of much longer length for academic research or industrial purposes. Here we selected 4 genes of ~1 kb (MTggps, SAggps and AFggps) and 1 gene of ~1.3 kb (cat2) which resembled the average length of both prokaryotic and eukaryotic gene sizes. The general strategy of screening and optimizing long gene libraries is illustrated in Figure 21. We proposed and tested this multi-step platform as the following: in the first step, each gene fragment library constructed from degenerate oligo libraries was cloned into the same pAcGFP1 vector using CPEC. In the second step, the cloning product was transformed into E. coli competent cells and screened for GFP signal representing highest protein expression. In the third step, 8-10 clones with the highest expression were collected for DNA plasmids. In the fourth step, the plasmid mixture for each highest expression gene fragment were amplified by plasmid PCR using one normal non-degenerate primer (in black) and one degenerate connector primer (the part that hybridized with the other connector was shown in green; the part that hybridized with the gene library was shown in red/blue).

In the fifth step, amplified enriched fragment library was used together with two end primers to assemble the full-length enriched gene library. In the sixth step, the amplified full-length enriched gene library was cloned into the same pAcGFP1 vector. In the seventh step, the cloning product was transformed into *E. coli* competent cells for screening. In the eighth step, a few clones with the highest GFP signal and hence protein expression were selected. To achieve this goal, construction of gene libraries using 60-mer oligos was attempted in order to test the ease and efficiency of the strategy. 60-mer oligos which were used for screening of short gene libraries in Chapter 4 were used in this study for gene library assemblies including MTggps, Saggps, AFggps and cat2.



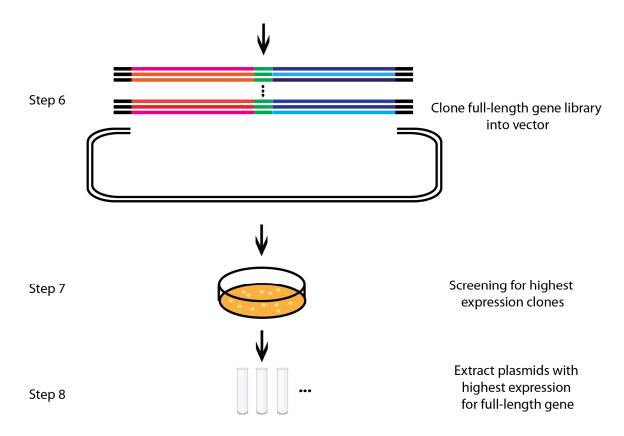


Figure 21: A schematic diagram of the proposed long gene screening mechanism. In the first step, each gene fragment library constructed from degenerate oligo libraries was cloned into vector. In the second step, the cloning product was transformed into *E. coli* competent cells and screened for GFP signal representing highest protein expression. In the third step, 8-10 clones with the highest expression were collected for DNA plasmids. In the fourth step, the plasmid mixture for each highest expression gene fragment were amplified by plasmid PCR using one normal non-degenerate primer (in black) and one degenerate connector primer (the part that hybridized with the other connector was shown in green; the part that hybridized with the gene library was shown in red/blue). In the fifth step, amplified enriched fragment library was used together with two end primers to assemble the full-length enriched gene library. In the sixth step, the amplified full-length enriched gene library was cloned into the vector. In the seventh step, the cloning product was transformed into *E. coli* competent cells for screening. In the eighth step, a few clones with the highest GFP signal and hence protein expression were selected.

In order to be able to screen for high-expression clones from long gene libraries and maybe genome libraries, each long gene library was split into shorter fragments.

Regular screening strategies as described in Chapter 4 were used for the screening of these shorter gene fragments. After each gene fragment was screened, a number of highexpression clones were selected. After confirming the sizes of these clones by colony PCR, the clone cultures with the correct size (the number of the clones was  $n_i$  and  $n_i$  = 10~16) were mixed and a mixture of plasmids from these clones was extracted to form a much smaller fragment library. Fragment libraries for all the fragments of a gene were then assembled together to form a full length gene library using degenerate sequences in between. The size of such a gene library was m and  $m = \prod n_i$ . It was indeed possible that the plasmids selected for each fragment library might contain errors such as substitutions or very small additions and deletions. However, compared the original library, such a selected small library had already eliminated the clones with additions or deletions that could be differentiated by colony PCR. Concerning the errors that could not be avoided and hence included in the fragment mixture, it was more of an issue of improving oligo synthesis quality since it was a limiting issue for whatever screening method that was used. If the synthesis quality was high, substitution errors should be minimal (see results below). The assembled full gene library was then cloned into plasmid and screened. Since such an assembled full gene library was intended to be of a much smaller size with each fragment optimized already, screening for high-expression clones for the full gene was highly possible.

In order to assess the feasibility of such a strategy, firstly 60-mer oligos was used for 4 long genes since it worked well for all the short gene library assemblies before. Here we split the 4 genes into 2 fragments, each of ~500 bp (for MTggps, SAggps and AFggps) or ~700 bp (for cat2). The first fragment (the latter half of the gene) was annotated as f1 (for example, MTggps-f1) and the second fragment (the first half of the gene) was annotated as f2 and so on. As for short gene libraries, 60-mer degenerate oligo mixture for each gene fragment was assembled by PCA (see Methods). A set of purified oligo mixture concentrations of 100 ng/200 ng/400 ng or 50 ng/100 ng/200 ng/400 ng was used in the reaction mixture of 50 µl. A low concentration of 50 ng per 50 ul of was used because cat2-f1 and cat2-f2 were longer, and lower concentrations of oligo mixtures seemed to favor gene assembly [86]. After PCA, part of the reaction product was resolved on a 1% agarose gel (Figure 22). It showed that some of the lanes had smears around the target band size of either ~500 bp or ~700 bp but not others which meant a set of PCA conditions should always be used in order to achieve the best assembly product after the next step of PCR.

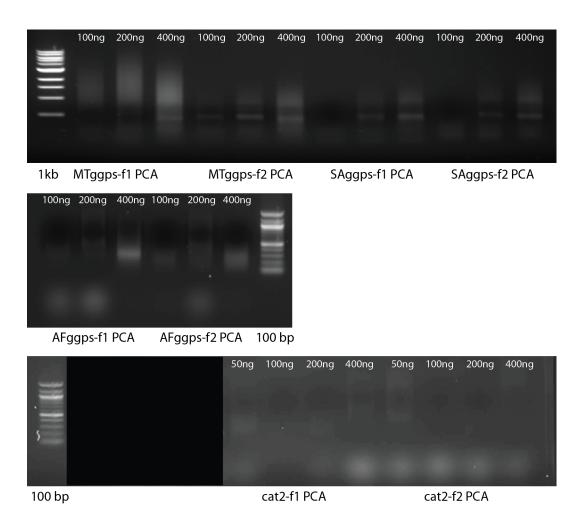


Figure 22: PCA result for MTggps, SAggps, AFggps both fragment libraries. 100 ng, 200 ng, and 400 ng purified oligo mixture were used respectively in Lane 1, 2 and 3 for each gene fragment. 1 kb and 100 bp DNA ladder were used respectively.

Next, PCR was performed using the assembled PCA crude product for the amplification of each gene fragment library (see Methods). After PCR, the entire reaction product was resolved on a 1% agarose gel (**Figure 23**). The results showed that as indicated by PCA gel electrophoresis result, some oligo mixture concentrations yielded a band at the correct size, indicating the formation of the correct gene fragment library product. It was interesting that the lower concentration of 50 ng and 100 ng for cat2-f1

and cat2-f2 didn't yield the correct product at ~700 bp as higher concentrations of 200 ng and 400 ng did. Again, since oligos synthesized for gene library assembly were a mixture of both correct and incorrect oligos, it was difficult to assess the actual concentrations of the correct oligo mixture without PAGE or HPLC purification of the synthesized oligos. Under such conditions, it was necessary to use a set of oligo mixture concentrations (e.g. from 100 ng to 400 ng) to try to get library PCR product. A successful library PCR result was indicated by a visible, preferably unique, band at the target size comparing with the DNA ladder as indicated in the gel picture by red arrows (Figure 23).

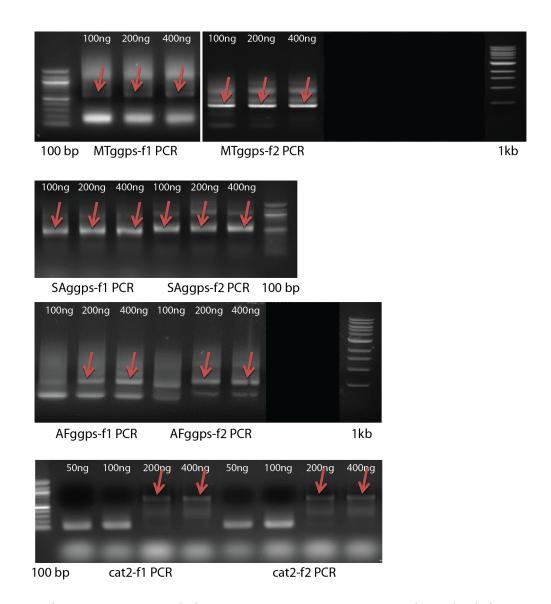


Figure 23: PCR result for MTggps, SAggps, AFggps and cat2 both fragment libraries. Red arrows indicated the formation of correct product.

PCR product for each gene fragment was purified from agarose gel, and concentrations were measured to prepare for the next cloning step. In this step, each gene fragment library PCR product was cloned into pAcGFP1 vector (see Methods) using CPEC cloning method. Since the gene fragment insert and the linearized vector had overlapping ends, they hybridized with each other during the annealing step and

formed a double-stranded circular plasmid after CPEC. In a total 50  $\mu$ l of volume, 200 ng of linearized vector and 1x or 2x vector molar amount of gene fragment library PCR was mixed with 10  $\mu$ l Phusion HF Buffer, 1  $\mu$ l dNTPs, 0.5  $\mu$ l Phusion DNA Polymerase and ddH<sub>2</sub>O. 20 thermal cycles were used for CPEC cloning of these fragment genes and annealing temperature was at 56 deg. C. The extension time was 15 s for all fragments. After CPEC, the product was run on a 1% agarose gel to confirm the CPEC result. A successful CPEC was indicated by a band at the size of linearized vector size plus the insert size (**Figure 24**).

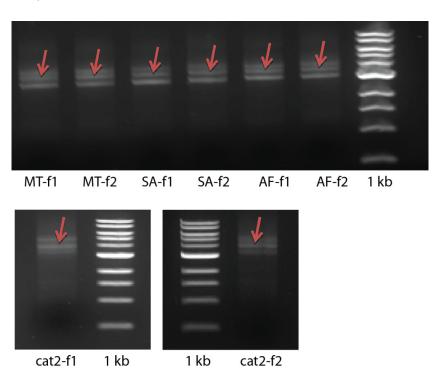


Figure 24: CPEC cloning gel electrophoresis result for MTggps, SAggps, AFggps and cat2 both fragments. Arrows indicated the correct CPEC product. The band below the product was the exessive vector.

CPEC product for each gene fragment library was then transformed into GC5 competent cells in order to screen for each gene fragment. After transformation, the transformants were spread on a large petri dish with LB agar and cultured in a 37 deg. C incubator for 16-18 hours. After that, the most fluorescent 8-16 colonies were selected for each gene fragment library from the petri dish and cultured in 3 ml LB for another 16-18 hours. Bacterial culture at this moment was used for colony PCR and then miniprep to extract the plasmids. scPCR was performed to confirm the size of the insert. Since the primers used for scPCR were ~360 bp in total from the insert ends, the scPCR product size should be the insert size plus 360 bp flanking sequence. The clones with the insert of the correct scPCR size were used for miniprep, and DNA plasmids were extracted for the following full gene assembly procedure.

Now that the highest expressing clone DNA for each gene fragment was prepared, means had to be discovered to link the two fragment libraries together into full gene library and screen for the highest expressing full gene clone for the ligated full library. Here the selected plasmids for each gene fragment were mixed into a plasmid mixture in equal molar amount. Each plasmid has an insert sequence that was different from the rest of the library. In order to link the two plasmid libraries together, two degenerate oligos which partly hybridize with the adjacent ends of both fragment libraries and partly hybridize with each other were used as connectors (Figure 21).

primers to amplify the fragment library. On the other side of each fragment library, the other one of the two primers was the one that hybridized with the non-degenerate sequence of the vector. PCR was performed to amplify each fragment library from the plasmid library using one non-degenerate primer and one degenerate primer/connector, which was called 'plasmid PCR'.

This method was first attempted on cat2 two fragment plasmid libraries. For cat2-f1 plasmid library, 1 ng, 5 ng, 12.5 ng, 25 ng, 50 ng or 100 ng of plasmid mixture was paired separately with 0.5 μM, 2.5 μM, 5 μM or 10 μM of the connector primer along with the rest of reaction reagents (see Methods). After 30 cycles of plasmid PCR, the entire reaction product was resolved on a 1% agarose gel (Figure 25). The result showed that after testing a combination of 6 plasmid mixture concentrations and 4 connector concentrations, only 1 ng and 5 ng plasmid mixture with 0.5 µM connector cat2-16co generated a clear band of ~700 bp which was the correct size of cat2-f1. This indicated that 0.5 µM was an appropriate concentration for connector as a primer in the plasmid PCR. Since the other non-degenerate primer also used a normal PCR concentration of 0.5 μM, it indicated that comparable concentrations of both primers with appropriate plasmid concentrations tended to yield the best result. The increase in plasmid concentrations tended to yield longer product than expected. For 50 ng and 100 ng plasmid concentrations used, barely any correct product was seen on the gel picture and

the major product was shown to shift to 2.5 kb to 5 kb which was highly undesirable and these concentrations should not be used for the following experiment conditions.

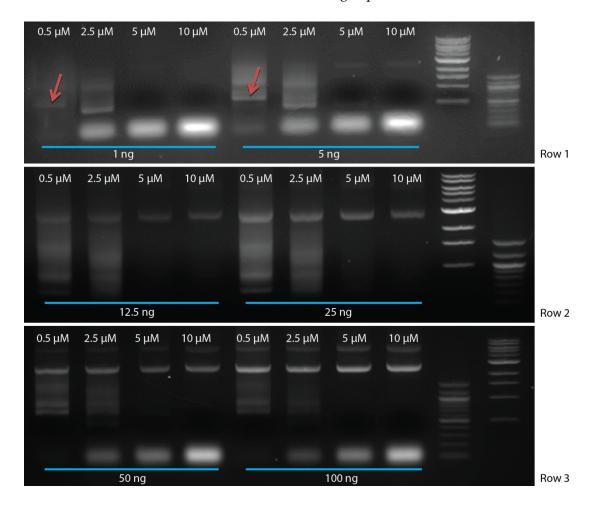


Figure 25: cat2-f1 plasmid PCR result using different plasmid mixture concentrations and degenerate connector primer concentrations. Row 1-3: 1 ng, 5 ng, 12.5 ng, 25 ng, 50 ng and 100 ng plasmid mixture with 0.5  $\mu$ M/ 2.5  $\mu$ M/ 5  $\mu$ M/ 10  $\mu$ M of degenerate connector cat2-16co3, respectively; 1 kb DNA ladder (NEB); 100 bp DNA ladder (NEB).

After finding out that higher plasmid mixture concentrations didn't work well for plasmid PCR, next, for cat2-f2 plasmid library, the same set of plasmid mixture concentrations except for 50 ng and 100 ng were used in combination with 0.5  $\mu$ M/ 2.5

 $\mu M/$  5  $\mu M/$  10  $\mu M$  of connector primer, respectively (Figure 26). The result showed that a clear band for cat2-f2 plasmid PCR was present for all plasmid mixture concentrations when the connector primer concentration is the range of 0.5  $\mu M\sim$  2.5  $\mu M$ . When the connector primer concentration went up and was higher than 5  $\mu M$ , the band of cat2-f2 plasmid PCR product disappeared. On one hand, it was confirmed that 0.5  $\mu M$  connector primer concentration was appropriate for plasmid PCR. On the other hand, since 1 ng and 5 ng plasmid concentrations yielded the clearest band, it showed that lower concentrations of plasmid mixture favored the formation of plasmid PCR product.

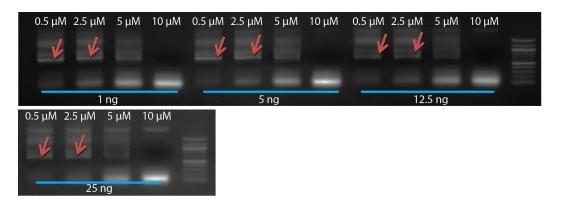


Figure 26: cat2-f2 plasmid PCR result using different plasmid mixture concentrations and degenerate connector concentrations. From top left to lower right lane: 1 ng , 5 ng, 12.5 ng and 25 ng plasmid mixture with 0.5  $\mu$ M/ 2.5  $\mu$ M/ 5  $\mu$ M/ 10  $\mu$ M of degenerate connector cat2-17co; 100 bp DNA ladder (NEB).

We confirmed from cat2 fragment plasmid PCR that  $0.5~\mu M$  was the lowest connector concentration that had always worked. Therefore, the same method to get fragment library PCR was used for MTggps, SAggps and AFggps fragments with  $0.5~\mu M$  connector primer concentration alone. Slightly different sets of plasmid mixture concentrations were used for different genes. The gel results for these 6 fragment

plasmid PCR were shown in **Figure 27** and the target bands were indicated by red arrow. It was noticed that except for the target bands, there were one or more non-specific amplifications that were mostly longer and above the target bands. The non-specific bands could be removed later by careful gel excision and extraction using the gel extraction kit.

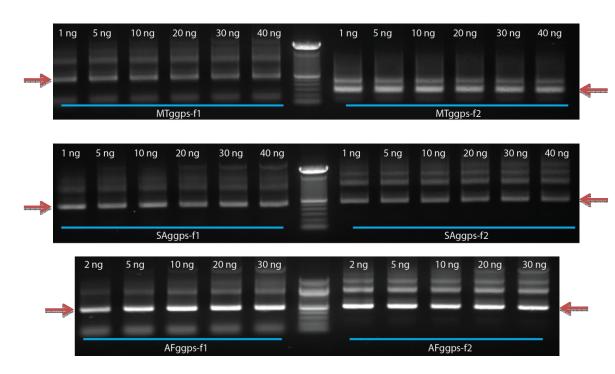
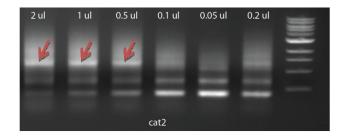
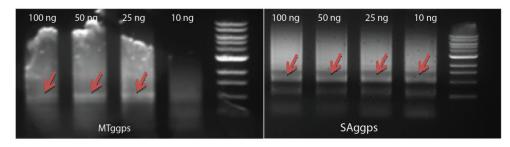


Figure 27: Plasmid PCR for MTggps, SAggps and AFggps using 0.5  $\mu$ M connector primer. Slight different set of plasmid mixture concentrations were used for different genes. 100 bp DNA ladder in the middle (NEB).

After each fragment library was amplified from the plasmid mixture, these fragments had degenerate ends that hybridized with each other. Then the full-length gene library was assembled by another PCR reaction with the two fragment library hybridizing with each other and the two non-degenerate primers used in the plasmid PCR at both ends. In a total 50  $\mu$ l of volume, cat2-f1 and cat-f2 plasmid PCR product

each in the amount of 2  $\mu$ l/ 1  $\mu$ l/ 0.5  $\mu$ l/ 0.2  $\mu$ l/ 0.1  $\mu$ l/ 0.05  $\mu$ l were mixed 10  $\mu$ l Phusion HF Buffer, 2.5  $\mu$ l each end primer (cat2-L and cat2-R), 1  $\mu$ l dNTPs, 0.5  $\mu$ l Phusion DNA Polymerase and ddH<sub>2</sub>O. In the case of MTggps, SAggps and AFggps, the plasmid PCR product was purified from the agarose gel and 100 ng/ 50 ng/ 25 ng/ 10 ng were used for each fragment library. In a PCR thermo cycler, after an initial 30 s of denaturation at 98°C, 30 cycles which consisted of 10 s of denaturation at 98°C, slow ramping at 0.1°C/s from 72°C to 45°C before annealing at 45°C for 2 m, and extension at 72°C for 20 s (15 s per kb) were performed. The reaction was ended with an extra 5 m of extension at 72°C. After full gene assembly reaction, the entire product was resolved on a 1% agarose gel. Using 2  $\mu$ l, 1  $\mu$ l and 0.5  $\mu$ l of each fragment plasmid PCR product yielded a clear band at ~1.4 kb but not the rest of conditions (**Figure 28**). It showed that higher concentrations of fragment library PCR favored the formation of the full-length gene library.





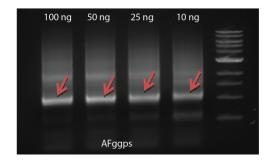


Figure 28: Full gene assembly result for cat2, MTggps, SAggps and AFggps using different plasmid PCR concentrations. Gene assembly product was indicated by red arrows. 1 kb DNA ladder (NEB).

Full gene assembly PCR product was purified from the agarose gel, and the concentration of the purified DNA was measured to prepare for the next cloning step. In this step, full gene assembly PCR product was cloned into pAcGFP1 vector using CPEC cloning method. Since the gene library and the linearized vector had overlapping ends, they hybridized with each other during the annealing step and formed a double-stranded circular plasmid after CPEC. In a total 50  $\mu$ l of volume, 200 ng of linearized vector and 1x or 2x vector molar amount of full gene library were mixed with 10  $\mu$ l

Phusion HF Buffer, 1 µl dNTPs, 0.5 µl Phusion DNA Polymerase and ddH<sub>2</sub>O. 20 thermal cycles were used for CPEC cloning, and the annealing temperature was at 56 deg. C for 2 m. The extension time was 45 s to cover the entire plasmid formed. After CPEC, the product was resolved on a 1% agarose gel to confirm the CPEC result. Both 1x and 2x full gene library used had a band at the size of vector plus full gene library (~4.7 kb) with 2x gene library used tended to have less unused vector (**Figure 29**). Therefore, for MTggps, SAggps and AFggps, a vector-to-insert ratio of 1:2 was used and yielded successful cloning product as shown by a clear band above the vector band indicated by the red arrows.

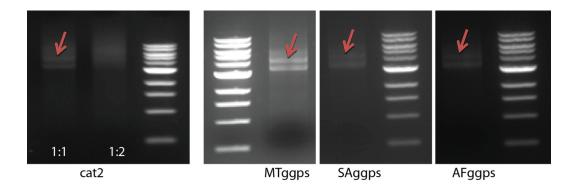


Figure 29: cat2, MTggps, SAggps and AFggps full gene library CPEC cloning result. Red arrows indicated the plasmid formed with gene library inserted. 1 kb DNA ladder (NEB).

CPEC product for the assembled full gene library was then transformed into GC5 competent cells to screen for the highest expressing clones. After transformation, the transformants were spread on a large petri dish with LB agar and cultured in a 37 deg. C incubator for 16-18 hours. After that, the most fluorescent 8-16 colonies were selected from the petri dish and cultured in 3 ml LB for another 16-18 hours. Bacterial cultures at

this moment were used for both scPCR and miniprep. scPCR was performed to confirm the size of the insert before sending for sequencing. Since the primers used for scPCR were ~360 bp in total from the insert ends, the scPCR product size for cat2, MTggps, SAggps and AFggps should be the insert plus 360 bp flanking sequence, ~1.7 kb for cat2 and ~1.3 kb for the rest of the genes. The result showed a number of clones out of all fluorescent clones analyzed by scPCR had the correct sized insert (Figure 30). The clones with insert of the correct scPCR size were used for miniprep to extract DNA plasmids and sequencing. From the gel electrophoresis of scPCR result, we could see that more than 87% of the colonies selected yielded correct sized insert for cat2, SAggps and AFggps. The percentage for MTggps was 50% which was probably due to a close second band from plasmid PCR (~400 bp) which inevitably got mixed into the gel extraction product (Figure 27). Since scPCR confirmed the size of the insert, the correct sized clones were used for miniprep to extract DNA plasmid and then sent for sequencing from both sides. The sequencing result showed that all the clones for each gene contained different and correct DNA sequence which encoded the same protein sequence (data not shown).

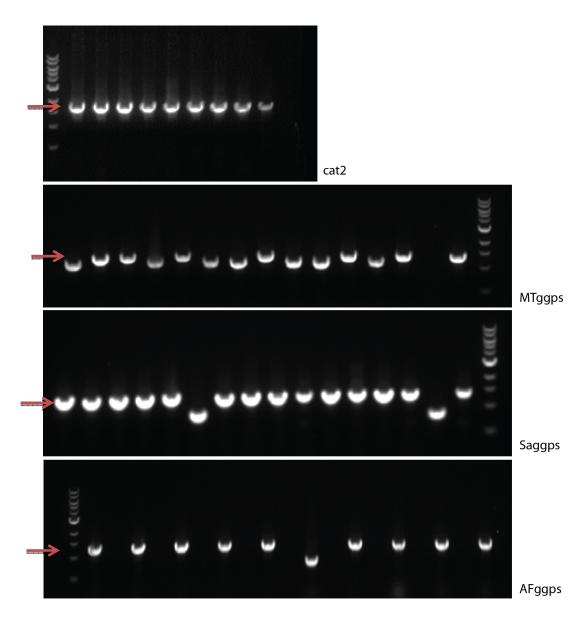


Figure 30: scPCR result for cat, MTggps, SAggpa and AFggps full gene libraries. Clones picked with the highest fluorescence were analyzed with scPCR to amplify the insert region. 1kb DNA ladder (NEB) was used for all rows.

# 5.4 Discussion and conclusions

In this study, a new method by using gene division and sub-optimization was created which enabled the screening of genes of average prokaryotic and eukaryotic gene

lengths (~1 kb and ~1.4 kb, respectively). Due to exponentially increased library size for genes of increasing sizes, a simple one-step screening method would not be feasible. This new method took advantage of the widely accepted assumption that by combining optimized gene fragments, optimized whole gene could be achieved. In another way, it is highly impossible that a combination of low-expression gene fragments could optimize the whole gene. This way the long genes could be first split into fragments and have each fragment library reduced to a much smaller optimized library. And then the full gene library could be assembled from fragment libraries which potentially contain a high-expression optimized gene sequence. The limit of this long gene optimization method has yet to be determined. So far, ~1.5 kb genes have been proven to be optimizable with this method. Longer oligos than 60-mers should be tested with the improvement of in vivo DNA synthesis technologies. However, it is still possible that an open reading frame cannot be optimized to a satisfactory expression level by changing synthetic synonymous codons. For such ORFs, no matter whether screening by one piece or screening by fragment, this would not work and other means such as changing the promoter or ribosomal binding site could be attempted.

# 6. Discussion and Conclusions

In this thesis, several projects were carried out for the final goal of the ultimate optimization of genes. Before, people have been trying to use rules summarized from a rather small gene variant dataset and computation tools to design the 'best' form of genes. However, these rules never worked reliably and efficiently for every gene in every organism. We proposed that instead of trying to force some random rules out of a small dataset or computation simulation, one could combine the ever-improving gene synthesis technology and high-throughput protein screening and be able to utilize the natural organism's expression system in a defined environment to screen out the best expressing gene variant among a significantly larger library. (For a gene of 300 bp, this library could easily go up to 10<sup>60</sup> variants.) At the beginning, we were obstructed by the cloning abilities available to us then. There was no convenient and efficient cloning method for large libraries. The CPEC cloning method we proposed and tested was the first and effective library and combinatorial library cloning method. The fact that this method does not require any expensive enzymes or take extra steps other than PCR thermal cycles made it an efficient workhorse for our following experiments and projects. After the reliable and efficient CPEC cloning method was developed, the next goal was to correct errors in the oligos synthesized in house which were used to assemble genes and gene libraries. We achieved satisfactory results that by performing two ECR iterations, the error frequency could be reduced by >16-fold and the whole process could

be completed in less than 5 hours. After we could prepare oligos with much reduced errors and an efficient library cloning method, my next goal was to establish a platform to screen and optimize synthetic genes in a high-throughput fashion by using these oligos synthesized in house and the CPEC cloning method. We demonstrated in this matter that a synthetic gene sequence with a desired protein expression level can be selected through one round of synthesis and screening with high confidence after screening for 1,000 to 1,500 colonies in *E. coli*. The results indicated that although a full set of codon usage rules were not being fully achieved or understood, optimized gene variant could still be selected from the library containing synonymous mutations. After the method was demonstrated on genes of 300~500 bp, the idea of applying this platform on longer genes were conceived and tested. In the fourth goal, we found that by using only one round of screening and optimization for a long gene library would not be effective since the errors contained in the library increased exponentially and the assembly and sequencing became very difficult. We resorted to gene division and suboptimization for each long gene and achieved good result so far. After testing optimizing genes between 1 kb and 2 kb successfully, we found that by dividing the gene with each optimized separately and assembling the located optimized gene fragments together into full gene libraries for final optimization, the protocol could be extended to even longer genes than 2 kb. Although the full capacity of the method hasn't been tested yet, the potential for this method to optimize normal length genes has been

confirmed. In conclusion, by combining an unique and efficient CPEC cloning method, and oligo error-correction protocol, a high-throughput screening and optimizing strategy and a gene division and sub-optimization way, we are currently on our way to achieving our final goal of being able to optimizing any given genes in a given organism which could greatly improve the science of gene design and protein engineering. The results we have achieved also contributed to the development of synthetic biology.

# **Appendix A**

# Supplementary sequences

# Chapter 2

# List of PCR templates and primers:

PCR products	Templates	Primers
2386 bp pUC19stop PCR	pUC19stop	SOSH6-L, SOSH6-R
product		
4746 bp pAcGFP1N1 PCR	pAcGFP1N1	pAcGFP1N1Fw3,
product		pAcGFP1N1Rv3
2959 bp pASK PCR product	pASK-IBA7C	pASKFw2, pASKRv
2040 bp phaAB	phaCAB Topo 15	phaABFw, phaABRv
171bp terminator	J04450	TermFw, TermRv
3280 bp Cat2phaC	pSOS-cat2phaC	cat2phaCFw, cat2phaCRv2
307 bp LacZhis6	Assembled LacZhis6	LacZH-L, LacZH-R
	library	
889 bp VacF1	Assembled VacF1 library	GP140-R, GP140-28L
882 bp VacF2	Assembled VacF2 library	GP140-L, GP140-29R
592 bp LacZhis6 single-	pUC19stop-LacZhis6	pUC19seqFw, pUC19seqRv
colony PCR product		
2029 bp gp140 single-colony	pAcGFP1N1-gp140	pAcGFP1N1seqFw, GFPRv
PCR product		

# List of PCR primer sequences (also the overlapping regions between insert and vector fragments):

Primer name	Primer sequences
SOSH6-L	CAA TTT CAC ACA GGA AAC AGC TAT G
SOSH6-R	TAA CTA GTG GTG GTG ATG ATG TGC
pAcGFP1N1Fw3	GCT GTG GTA TGT TAA CTA TCG TAC GCG GGA TCC ACC
	GGT CTT G
pAcGFP1N1Rv3	GGG CCC ACC GAA CGC CAT GAA TTC GAA GCT TGA GCT
pASKFw2	CAA AGC cAA ggc atg aCC CTC GAG GTC GAC CTG CAG

pASKRv	CTT TCA ATG GTT GCC CTC GTT ATC TAG ATT TTT GTC G
phaABFw	AGA TAA CGA GGG CAA CCA TTG AAA GGA CTA CAC AAT
	GAC TGA CG
phaABRv	GCT CTA GTA TCA GCC CAT ATG CAG GCC GCC
TermFw	CTG CAT ATG GGC TGA TAC TAG AGC CAG GCA TCA AAT
	AAA ACG
TermRv	GCT CAC TGC CCG CTT TCC ATA TAA ACG CAG AAA G
cat2phaCFw	TGG AAA GCG GGC AGT GAG CGC
cat2phaCRv2	CGA GGG TCA TGC CTT GGC TTT GAC GTA TCG C
LacZH-L	CAA TTT CAC ACA GGA AAC AGC TAT G
LacZH-R	TAA CTA GTG GTG GTG ATG ATG TGC
GP140-R	CGT ACG CTA GTT AAC CTA CCA SAG C
GP140-28L	GAC ATC ATC GGC GAC ATC C
GP140-L	GCG TTC GGT GGG CCC AAC
GP140-29R	GGA TGT CGC CGA TGA TG
pUC19seqFw	GCA GCT GGC ACG ACA GGT TTC
pUC19seqRv	CGT CAT CAC CGA AAC GCG CGA
pAcGFP1N1seqFw	CAT TGA CGC AAA TGG GCG GTA GG
GFPRv	TTG CCG GTG GTG CAG ATG AAC

# **Chapter 3**

>rfp (723-bp)

AATTTCACACAGGAAACAGCTATGATGGCTTCCTCCGAAGACGTTATCAAAGAG
TTCATGCGTTTCAAAGTTCGTATGGAAGGTTCCGTTAACGGTCACGAGTTCGAAA
TCGAAGGTGAAGGTGAAGGTCGTCCGTACGAAGGTACCCAGACCGCTAAACTG
AAAGTTACCAAAGGTGGTCCGCTGCCGTTCGCTTGGGACATCCTGTCCCCGCAGT
TCCAGTACGGTTCCAAAGCTTACGTTAAACACCCGGCTGACATCCCGGACTACC
TGAAACTGTCCTTCCCGGAAGGTTTCAAATGGGAACGTGTTATGAACTTCGAAG
ACGGTGGTGTTGTTACCGTTACCCAGGACTCCTCCCTGCAAGACGGTGAGTTCAT
CTACAAAGTTAAACTGCGTGGTACCAACTTCCCGTCCGACGGTCCGGTTATGCA
GAAAAAAACCATGGGTTGGGAAGCTTCCACCGAACGTATGTACCCGGAAGACG
GTGCTCTGAAAGGTGAAATCAAAATGCGTCTGAAACTGAAAGACGGTGGTCACT
ACGACGCTGAAGTTAAAACCACCTACATGGCTAAAAAACCGGTTCAGCTGCCG
GGTGCTTACAAAACCGACATCAAACTGGACATCACCTCCCACACAACGAAGACTA
CACCATCGTTGAACAGTACGAACGTGCTGAAGGTCGTCACTCCACCCGGTGCTTA
AGAAACCGTGCGTTTCCAGTCT

Chip oligos:

RFP-f1-1,

TTCAAATGGGAACGTGTTATGAACTTCGAAGACGGTGGTGTTGTTACCGTTACCC AGGACGCATGACTCGACCATCCGATTTTTT

RFP-f1-2.

TTCATAACACGTTCCCATTTGAAACCTTCCGGGAAGGACAGTTTCAGGTAGTCCG GGATGGCATGACTCGACCATCCGATTTTTT

RFP-f1-3,

CCAGTACGGTTCCAAAGCTTACGTTAAACACCCGGCTGACATCCCGGACTACCT GAAACTGCATGACTCGACCATCCGATTTTTT

RFP-f1-4,

CGTAAGCTTTGGAACCGTACTGGAACTGCGGGGACAGGATGTCCCAAGCGAAC GGCAGCGCATGACTCGACCATCCGATTTTTT

RFP-f1-5,

ACGAAGGTACCCAGACCGCTAAACTGAAAGTTACCAAAGGTGGTCCGCTGCCG TTCGCTTGCATGACTCGACCATCCGATTTTTT

RFP-f1-6,

GCGGTCTGGGTACCTTCGTACGGACGACCTTCACCTTCACCTTCGATTTCGAACTCGTGAGCATGACTCGACCATCCGATTTTTT

RFP-f1-7,

CATGCGTTTCAAAGTTCGTATGGAAGGTTCCGTTAACGGTCACGAGTTCGAAATC GAAGGCATGACTCGACCATCCGATTTTTT

CATACGAACTTTGAAACGCATGAACTCTTTGATAACGTCTTCGGAGGAAGCCAT CATAGCGCATGACTCGACCATCCGATTTTTT

RFP-f1-9,

RFP-f1-8.

AGCCTGGATGACGTTTTCATCAAAATTTCACACAGGAAACAGCTATGATGGCTTCCTCCGGCATGACTCGACCATCCGATTTTTT

RFP-f2-1,

CGGGAAGTTGGTACCACGCAGTTTAACTTTGTAGATGAACTCACCGTCTTGCAGG GAGGAGCATGACTCGACCATCCGATTTTTT

RFP-f2-2,

CGTGGTACCAACTTCCCGTCCGACGGTCCGGTTATGCAGAAAAAAACCATGGGT TGGGAAGCATGACTCGACCATCCGATTTTTT

RFP-f2-3,

TCAGAGCACCGTCTTCCGGGTACATACGTTCGGTGGAAGCTTCCCAACCCATGGT TTTTTGCATGACTCGACCATCCGATTTTTT

RFP-f2-4.

CGGAAGACGTGCTCTGAAAGGTGAAATCAAAATGCGTCTGAAACTGAAAGAC GGTGGTCGCATGACCCATCCGATTTTTT RFP-f2-5,

TTTAGCCATGTAGGTGGTTTTAACTTCAGCGTCGTAGTGACCACCGTCTTTCAGTT TCAGGCATGACTCGACCATCCGATTTTTT

RFP-f2-6,

GAAGTTAAAACCACCTACATGGCTAAAAAACCGGTTCAGCTGCCGGGTGCTTAC AAAACCGCATGACTCGACCATCCGATTTTTT

RFP-f2-7,

GTGTAGTCTTCGTTGTGGGAGGTGATGTCCAGTTTGATGTCGGTTTTGTAAGCACC CGGCGCATGACTCGACCATCCGATTTTTT

RFP-f2-8,

CCTCCCACAACGAAGACTACACCATCGTTGAACAGTACGAACGTGCTGAAGGTC GTCACTGCATGACTCGACCATCCGATTTTTT RFP-f2-9.

AAATTGAGACTGGAAACGCACGGTTTCTTAAGCACCGGTGGAGTGACGACCTTC AGCACGGCATGACTCGACCATCCGATTTTTT

#### PCR primers:

RFP-F, AATTTCACACAGGAAACAGCTATGA

RFP-R, AGACTGGAAACGCACGGTT

RFP-M, TGTTGTTACCGTTACCCAGGACTCCTCCCTGCAAGACGG

#### Vector primers for CPEC:

RFP\_VECTOR\_RV, ATGATGATGGTGGTGCATAGCTGTTTCCTGTGT RFP VECTOR FW,

ACCGTGCGTTTCCAGTCTGTCGCCACCGTGAGCAAGGGCGCCGAGCT

#### $> lacZ\alpha$ -v1 (174-bp)

AATTTCACACAGGAAACAGCTATGACCATGATTACGCTGGCCGTCGTTTTACAA CGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACAT CCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCCATCATCAC CACCACTAGTTA

#### PCR primers:

mini-lacZ\_L, AATTTCACACAGGAAACAGCTATGA mini-lacZ\_R, TAACTAGTGGTGGTGATGATGATGA

#### Vector primers for CPEC:

 $\label{eq:mini-lacZ} mini-lacZ\_R, CATCATCACCACCACCACTAGTTAAGCCAGCCCCGACAC \\ mini-lacZ\_L, CAATTTCACACAGGAAACAGCTATG$ 

# $> lacZ\alpha$ -v2 (174-bp)

AATTTCACACAGGAAACAGCTATGACCATGATTACGCTGGCCGTCGTTTTACAA CGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACAT CCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCCATCATCAC CACCACTAGTTA

## PCR primers:

mini-lacZ\_L, AATTTCACACAGGAAACAGCTATGA mini-lacZ\_R, TAACTAGTGGTGGTGATGATGATGG

# Vector primers for CPEC:

mini-lacZ\_R, CATCATCACCACCACCACTAGTTAAGCCAGCCCCGACAC mini-lacZ\_L, CAATTTCACACAGGAAACAGCTATG

## > Construct-3 (501-bp)

#### Chip oligos:

GFP-FRGA-1,

CAAATCAATACTATTACCGCCCATGTCGGAAGTAAATCCCACATGCATAGGTTG ATGAATGCATGACCCACCATCCGATTTTTT

GFP-FRGA-2,

TGAAGCTTTGCTCTTCACACTTCATGGGAGACATTCATCAACCTATGCATG TGGGGCATGACTCGACCATCCGATTTTTT GFP-FRGA-3,

GAAAGGAAGCAAAGCTTCAGTCATATTGTATCGCCTATCACTTGTACCGTGA CGGAAAGCATGACTCGACCATCCGATTTTTT GFP-FRGA-4, TGCCGCCGATTCAAAATTTCACCTCTCAATTGGGGCATTTCCGTCACGGTACA AGTGAGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-5,

ATCGCGCCGGCAACGCACGTTCCCACCATGGGGATCATGACAGTCACGTTGATAGGCATGACCATCCGATTTTTT

GFP-FRGA-6,

TTTCATTGATACCCCTGACCAAGCATGTTCCTTTGACTATCAACGTGACTGTCATG ATCCGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-7,

TTGGTCAGGGGTATCAATGAAATGTAAGGAGGAGGTCCAACGGTACTTGTACCA ATGACGGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-8,

AAACTTATCAAGCCTATGTGTTTGGCCAGATCAGATCCGTCATTGGTACAAGTAC CGTTGGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-9,

GCCAAACACATAGGCTTGATAAGTTTCCGTTGGCATAATCAGGCAACAACATCT TTACCGGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-10,

AGATTGCCCAGGACCTTCTTGAGCCAGAAGCCGCTCATGCGGTAAAGATGTTGT TGCCTGGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-11,

CAAGAAGGTCCTGGGCAATCTGACAGGTCATAACATGTCCTTTCACCAAG CTCTTAGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-12,

TTCTTTCTACTTCTCGCCCTTGTCGTGGAGCCCGGTGTAAGAGCTTGGTCAAAGG AAGGAGCATGACTCGACCATCCGATTTTTT GFP-FRGA-13,

ACAAGGCGAGAAGTAGAAAGAACACGGAGTACAGTCTGGTCTCGAGATCATC CTTGTCAGCATGACTCGACCATCCGATTTTTT

GFP-FRGA-14,
CCCCGATCCAAATACGCACTATAATCTAGATTTAACGCTGACAAGGATGATCTC
GAGACCGCATGACTCGACCATCCGATTTTTT

PCR primers:

GFP-FRGA-F,

CTTGCTCACGGTGGCGACCAAATCAATACTATTACCGCCCATGTCGGA

Vector primers for CPEC:

GFP-FRGA\_VECTOR\_RV,
GGAGTACAGTCTGGTCATAGCTGTTTCCTGTGTGAAATTGTTATC
GFP-FRGA\_VECTOR\_FW,
GGTAATAGTATTGATTTGGTCGCCACCGTGAGCAAGGGC

#### >Construct-4 (498-bp)

GFP-FRGB-1,

TATTTAATTACCTGCAGGGAATTCTTAATGATGATGATGATGATGTGCCAACGAA TGGTCGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-2,

GATTAAGCATGATTCTTAACCGTGTCCTGGGTTCGTCGACCATTCGTTGGC ACATCGCATGACTCGACCATCCGATTTTTT GFP-FRGB-3,

GACACGGTTAAGAATCATGCTTAATCTGATTCCTCCTTGGGCGATACGTTTCATA ACAATGCATGACTCGACCATCCGATTTTTT GFP-FRGB-4,

GACATTGTCGGATAAATATTTTAACACGCGAATGCCAATTGTTATGAAACGTATC GCCCAGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-5,

CGTGTTAAAATATTTATCCGACAATGTCTCTCCACTTTCTACATCCTTATACCCCC AGTTGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-6,

AAATATGCTAAGTAATCAATTAAGTTGGCTTGTAACTGGGGGTATAAGGATGTA GAAAGTGCATGACCCATCCGATTTTTT GFP-FRGB-7,

GCCAACTTAATTGATTACTTAGCATATTTGTTTGCACAAGTAGAGATATCATCGC ATTCTGCATGACTCGACCATCCGATTTTTT GFP-FRGB-8, GGGTTCTTGGTTGCAAGACGTGGAATCTTGGAAAGAATGCGATGATATCTCTACT TGTGCGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-9,

CGTCTTGCAACCAAGAACCCTCCGTAAAATTTCTCTGAATGTCTTGAAGCAAGG AATGCAGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-10,

TCATTCTAACTGCTGCTGATTATCACGGAAAGGATATGCATTCCTTGCTTCA AGACAGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-11,

TCAGCAGCAGCAGTTAGAATGATCTCTCGATCCCAGACGTGATGCAGGTTTGAT TTATGAGCATGACCACCATCCGATTTTTT

GFP-FRGB-12,

ACATGGGCGTAATAGTATTGATTTGAGGTGGTTTCGTCATAAATCAAACCTGCA TCACGGCATGACTCGACCATCCGATTTTTT

GFP-FRGB-13,

CAAATCAATACTATTACCGCCCATGTCGGAAGTAAATCCCACATGCATAGGTTG ATGAATGCATGACTCGACCATCCGATTTTTT GFP-FRGB-14,

TGAAGCTTTGCTCTTCACACTTCATGGGAGACATTCATCAACCTATGCATG TGGGGCATGACTCGACCATCCGATTTTTT

PCR primers:

GFP-FRGB-F,

TTCACACAGGAAACAGCTATGGAAGCTTTGCTCTTTCACACTTCAT GFP-FRGB-R,

CTTGCTCACGGTGGCGACATGATGATGATGATGTGCCAACGAATG

Vector primers for CPEC:

GFP-FRGB\_VECTOR\_RV,

AAGGAAGCAAAGCTTCCATAGCTGTTTCCTGTGTGAAATTGTTATC GFP-FRGB\_VECTOR\_FW,

GGTAATAGTATTGATTTGGTCGCCACCGTGAGCAAGGGC

# Chapter 4

# 1. Test gene constructs:

>Lac $Z\alpha$ 

ATGACCATGATTACGCCGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCT GGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTCGCCAGCTGGCGTA ATAGCGAAGAGGCCCGCACC

#### >RFP

ATGGCTTCCTCCGAAGACGTTATCAAAGAGTTCATGCGTTTCAAAGTTCGTATGG
AAGGTTCCGTTAACGGTCACGAGTTCGAAATCGAAGGTGAAGGTGAAGGTCGTC
CGTACGAAGGTACCCAGACCGCTAAACTGAAAGTTACCAAAGGTGGTCCGCTG
CCGTTCGCTTGGGACATCCTGTCCCCGCAGTTCCAGTACGGTTCCAAAGCTTACG
TTAAACACCCGGCTGACATCCCGGACTACCTGAAACTGTCCTTCCCGGAAGGTTT
CAAATGGGAACGTGTTATGAACTTCGAAGACGGTGGTGTTGTTACCGTTACCCA
GGACTCCTCCCTGCAAGACGGTGAGTTCATCTACAAAGTTAAACTGCGTGGTAC
CAACTTCCCGTCCGACGGTCCGGTTATGCAGAAAAAAACCATGGGTTGGGAAGC
TTCCACCGAACGTATGTACCCGGAAGACGGTGCTCTGAAAGGTGAAATCAAAAT
GCGTCTGAAACTGAAAGACGGTGGTCACTACGACGCTGAAGTTAAAACCACCTA
CATGGCTAAAAAAACCGGTTCAGCTGCCGGGTGCTTACAAAAACCGACATCAAACT
GGACATCACCTCCCCACAACGAAGACTACACCATCGTTGAACAGTACGAACGTG
CTGAAGGTCGTCACTCCACCGGTGCTTAA

# 2. Drosophila transcription factor gene fragments (wild-type):

#### >AB1

AGGAGTCCTATTAGCCTGGAACATGACCGCCAAAAGTCATCACCCGTCAGGGTC AAGAGCTTTTCCATTGCCGATATCCTGGGGCGAGGGCATGAAGAGGATCGTGTG GAGAAGAAGCCGGAGAGAATTGCCATTCCCAATGCTTTACCAGGAGTACCTGCT CCATTGGCTCTGATCAACGAAAGATTGCAAATACCCTTGGTGCCGAATTGCCCTC CGCCGGCTGCACTACATACATTTTTATCGCCCGCTTTGTGT

#### >AB11

GCCTATCACCTGGGCAGCCATCTCGGCCAGACATCGACGCCAACCACGACGCA GGCCACCCTCTCCGGTCAGATGGACAAGTTCGCCCTGGAGCGCAGCTCGTATCT GAGCAACTCCTCGCAGGCGAGTGCCTACAGCGCCTACTTGGACCCCAGTGTGCT GACCAAGGCGTATTTCGACTCCAAGATGTACCAGGATCGGGCGGCCAACTATGC CTTCGACATATCCAAGATTTATGGTGCCCAGCAACATGCAGCAGCCCACCACCA GTGT

#### >AC12

TTCCAGCAGACGACATCGGGCGCCCTACGGAGCGCAGCCCATTCAGAGTGGTTAC CTGCACTACGGCAACTATGGCGGATACGAGGGAATGGCCCAGAGTCAGCCACA GAACCAAGCGCCCGTCCACCATCAGCAGGCATCCCACTCGGCTCCGTGT

#### >AF4

#### >AG3

GGCCAGTGGACCGAGGACGATTACCAACGGGCGCAGTCGAGCAGCGTGAGTCC
ACGTGGTGGCTACTGCAGCAGCGCCTCCACGCCGCACAGCTCGGGAGGATCCTA
CGGTGCCACGGCGGCCAGTGCAGCGGTGGCAGCCACGGCCAATGGCTATGCAC
CTGCACCCAACATGGGCACACTCTCCTCGTCGCCCGGCAGCGTCTTCAATTCCAC
GTCAAGGGTCAGCAGCCTGAGCTTCAATCCCTTCGCCCTGCCCACCTGCAATAC
ACAGGGCTATAGCACCCAACTGGTGACGTCAACCAAATGT

#### >AR1

TTGTACCTCTGCCTCTCCCGCGACGACGCCATCTCCACCTGCATCTGCACCGAGT GCTGCTCGCAACTGGAGAGCTTCCACAACTTCTGGAAGCTGGAGCTAAAGC AGACGACGCTGTGCAGCCAGTTTTTGGCTATCGATTGTGATGTCAACTGGTCGGA GGATGGCAGCGAAACGCAGTTGGATGCGCAACCGCAATTGCTACTGGAGCCGG CGGAGGAGCCCAAAGTGGTGACTCCCACAACGGCGAATAAGTTTCCCTGCATGT TCTGCGAGAAGTCATTTTGT

#### >B11

ATGGTGAAGATCGAGGAAGGTCTGCCGTCCAGCGAGATTAGCGCTCACAGTCTC
CACTTTCAGCACCATCACCACCCACTGCCGCCCACCACTCACCATTCCGCGCTCC
AAAGTCCCCATCCCGTCGGGCTGAATCTCACCAATCTAATGAAAATGGCCAGGA
CACCACATTTGAAGTCCAGCTTTTCCATTAATTCCATTCTACCCGAGACTGTGGA
GCATCACGACGAGGATGAGGAAGAAGAAATCACCAGCGAAA
TTTCCGCCC

#### >D5

GAGAACGGGATGATGTACACCAACCTGGACTGCATGTATCCCACGGCCCAGGCT CAGGCTCCGGTTCACGGATATGCCGGCCAGATCGAGGAGAAATACGCCGCCGTC 

#### >F9

#### >K1

CATGACTTCACCCGGCCTATAGAATTCCCGGATACATGGAACAGTTGTATTCTC
TTCAACGAACTAATTCAGCTTCATCTTTTCACGATCCCTATGTCAATTGTGCATCA
GCATTCCATCTTGCCGGACTTGGTTTGGGATCAGCTGATTTTTTGGGCAGCCGAG
GTTTGAGCTCTTTGGGTGAACTGCATAATGCTGCTGTGGCAGCAGCATGT

#### >K3

GCATCGAACTCGTCGACGTCCAGCGAGGCTTCCAATTCATCGCAGCAGAATAAT GGATGGAGCTTTGAAGAGCAGTTTAAACAAGTCAGACAGCTCTATGAAATCAAT GATGATCCCAAGCGCAAAGAGTTCCTGGACGACTTGTTCTCGTTTATGCAAAAG CGCGGAACTCCGATCAATCGGCTGCCGATCATGGCCAAATCGGTGCTGGATCTC TACGAGCTGTACAATCTGGTGATAGCCCGCGGCGGCTTGGTGGATGTTTGT

#### >K4

TCGGCACACCTCGCCGCAGCAACGCGAAGCCCTGAACCTGTCCGACTCGCCT CCAAATCTCACAAATATCAAACGGGAACGCGAACGGGAACCCACACCAGAGCC CGTGGACCAGGATGACAAATTTGTGGACCAGCCACCTCCAGCGAAGCGCGTGG GCAGTGGCCTCCTTCCGCCCGGCTTTCCCGCCAACTTCTACCTGAATCCACACAA CATGGCCGCTGTGGCAGCATGT

#### >K5

AGTTTCGAAGAGCTCTATTGCGAGTGTGGTGCCGAAGTGATCTATCCACCAGTGC CGTGTGGCACCAAGAAGCCAATCTGCAAGCTTCCTTGCTCGCGCATTCATCCATG CGATCATCCGCCGCAGCACAACTGCCATTCGGGCCCCACCTGTCCGCCCTGCAT GATTTTCACGACGAAATTGTGTCACGGCAACCACGAGCTGCGCAAGACCATTCC CTGCTCGCAGCCCAACTTTAGCTGTGGCATGGCCTGTGGCAAGCCGTTGCCGTGT GGCGGGCACAAGTGTATAAAGCCGTGTCACGAGTGT

#### >L5

GGAGGAAACTCGCACCCAGCGATCGGGGTGATGAGCCTGTTCGATCCTCGATAC ATGGATTCGCCCGGCAATGTGCACTCGATGACGCGGCAGCGTTACATCGAAGCC ACCGGCGGCGGAGCAGGTGCGGGAACGGGCGCAGCCACCATCAACGT TAATCCGACCACCGCCATGAATCTACAGAGTATTAACTTCAACTCATGT

#### >M6

GGTGGAGCCGCAATGCCAATGCCGGTAATGCGGCCAATGCAAACGGACAGAA CAATCCGGCGGGCGTATGCCCGTTAGACCCTCCGCCTGCACCCCAGATTCCCG AGTGGGCGGCTATTTGGACACGTCGGGCGGCAGTCCCGTTAGCCATCGCGGCGG CAGTGCCGGCGGTAATGTGAGTGTCAGCGGCGGCAACGGCAACGCCGGAGGCG TACAGAGCGGCGTGTGT

#### >bcd d2

ATCTTGGAGCCTTTGAAGGGTCTGGACAAGAGCTGCGACGATGGCAGTAGCGACGACGACGACATGAGCACCGGAATAAGAGCCTTAGCAGGAACCGGAAATCGTGGAGCGCCATTTGCCAAATTTGGCAAGCCTTCGCCCCCACAAGGCCCTCAGCCGCCCCTCGGAATGGGGGGCGTGGCCATGGGCGAATCGAACCAATATCAATGCACGATGGATACGATAATGCAAGCGTATAATCCCCATCGGAACGCCGCGGGCAACTCGCAGTTTGCTTACTTCAAT

#### >cad d1

AGGACATCGCCATCGAAGCCGCCATACTTCGACTGGATGAAGAAGCCCGCCTAT CCAGCACAACCACAACCAGGCAAAACCCGCACCAAGGATAAGTACCGCGTGGT GTACACCGACTTCCAGCGCCTGGAGCTGGAGAAGGAGTACTGCACCTCCCGCTA CATCACCATCCGGCGGAAGAGCGAGCTCGCCCAGACGCTGTCGCTGTCGGAGC GCCAGGTTAAGATCTGGTTCCAGAACCGCCGCGCCAAGGAGCGCAAGCAGAAC AAGAAGGGTAGCGATCCGAACGTGATGGGCGTG

#### >hb d2

AAGCACAAGAACCAAAAGCCCTTCCAGTGCGACAAATGCAGCTACACGTGTGT
CAACAAATCCATGCTAAACTCGCACCGCAAGTCGCACAGTTCTGTGTATCAGTA
CCGTTGTGCGGATTGTGATTACGCCACCAAGTATTGCCACAGCTTCAAGCTGCAT
CTGCGCAAGTATGGTCACAAGCCCGGCATGGTTTTGGACGAGGATGGCACCCCG
AATCCCTCGTTGGTCATCGATGTTTACGGCACGCGTCGTGGTCCGAAGAGCAAG
AATGGTGGACCGATTGCCAGTGGAGGAAGTGGCAGCGGCAGCCGGAAGTCAAA
TGTTGCAGCTGTCGCTCCG

>lab\_d1

CCCAACGCCATGGGCTGGGTCTGGGCAGCGGATCCGGACTGAGCAGCTGCAG CCTCTCCAGCAACACCAACAACTCCGGCCGGACGAACTTCACCAACAAGCAGC TGACCGAGCTGGAAAAAGAGTTCCACTTCAATCGCTACTTGACGCGGGCGCCC GCATTGAAATCGCCAATACGTTGCAGCTTAATGAAACGCAGGTCAAAATCTGGT TCCAGAACCGCCGCATGAAGCAGAAGAAGCGCGTGAAGGAGGGCCTCATTCCG GCGGACATCCTGACGCAGCACTCCACGTCCGTGATCAGCGA GAAGCCG

## >lab\_d2

# >pb\_d2

AACTCCAATTCGAAGAAATCGTGTCAAGGCTGCGAACTGCCATCCGATGATATA
CCCGATTCCACGTCAAATTCCCGGGGTCACAACAACAACAACACGCCGAGCGCCAC
CAACAATAACCCGAGTGCTGGGAACCTGACTCCAAACTCGTCCCTGGAAACGG
GCATCTCCTCGAACCTGATGGGCAGCACCACCGTATCCGCCTCAAATGTGATCA
GCGCCGACTCCAGTGTGGCATCTAGTGTTAGCCTGGACGAGGATATTGAGGAGA
GCAGCCCCATAAAGGTGAAGAAGAAGACGATGGTCAGGTCATCAAAAAGGA
GGCGGTCTCCACCTCGTCGAAAGCCTCGCCTTTTGGTTATGAAAAACTCCACTCCC
AGTCTGGTCAGTTTTCGACGGGATTCAGATGCCTCC

## >Dfd d1

#### >Scr d2

# 

# >Antp\_d1

ACCTCGTACTTCACCAACTCGTACATGGGGGGCGGACATGCATCATGGGCACTAC
CCGGGCAACGGGGTCACCGACCTGGACGCCCAGCAGATGCACCACTACAGCCA
GAACGCGAATCACCAGGGCAACATGCCCTACCGCGCTTTCCACCCTACGACCG
CATGCCCTACTACAACGGCCAGGGGATGGACCAGCAGCAGCACCAGGTCT
ACTCCCGCCCGGACAGCCCCTCCAGCCAGGTGGGCGGGGTCATGCCCCAGGCGC
AGACCAACGGTCAGTTGGGTGTTCCCCAG

#### >lid d2

TGCGAGTGCAACGACAAGTTGATTGTCTGTTTGCGCCACTATACGGTACTCTGCG GCTGTGCTCCTGAGAAGCACACTCTGATATATCGTTATACACTGGACGAGATGC CGCTGATGCTGCAGAAACTTAAGGTGAAGGCGCACAGCTTCGAGCGATGGCTTT CACGTTGCCGTGACATAGTCGATGCACATACACCCACCTCGGTGACCCTGCAGG AGCTGCAGGAATTGTGCAAGGAGGCCGAGACGAAAAAGTTTCCCTCCTCGCTGC TCATCGATCGCCTTAATGCTGCGGCCGTTGAGGCAGAGAAA

# >lilli\_d1

TTGTATAGCCAAAATTATAACATGGAGGAGTACGAACGGCGAAAAAGACGTGA GCGTGAGAAAATCGAGCGCAACAGGGCATACAGATTGACGATCGGGAGACTA GTCTATTCGGGGAGCCGCGTCGGCTGACTGAGGGAGATGCGGAGATCACCGCCG CCCTGGGTGAGTTCTTCGAGGCGCGGGGAGTACATCAACAATCAGACTGTGGGAA TCAGCCGGAGTGCGCCCGGCGCCGGCAATCCGCGCCTGCAGCCCAATCTGGCGC CGCAAGCCAAATCCCTGGGGCATTCACCCTCCTCCGCCTCCTCAGCAGCTGGGC CCACTGCCGCGTCCGCGACCACTTCGCTGCCGGCCAG

#### >lilli\_d2

AGGGGGACATTGCCAACGCCAACACGCCGTCCTCCATATCACCATCGAACTCG
GTGGGCTCGCAGGCTCCGGTTCGAATACGCCGCCAGGCAGAATAGTGCCTCCA
GATATACACAATATGCTGTGCAAGCAGAATGAGTTTCTGAGCTATCTAAATAGC
GCTCACGAGTTGTGGGATCAAGCCGATCGATTGGTGCGCACAGGCAATCATATA
GATTTCATCCGAGAACTGGATCACGAGAACGGCCCGCTGACGCTGCATAGCACC
ATGCACGAGGTGTTCCGGTACGTGCAGGCGGGTCTCAAGACGCTCAGGGATGCC
GTGTCGCATCCGACG

#### >E75 d1

CACTTGACAGCCGGAGCTGCCCGCTACAGAAAGCTAGATTCGCCCACGGATTCG GGCATTGAGTCGGGCAACGAGAAGAACGAGTGCAAGGCGGTGAGTTCGGGGGG AAGTTCCTCGTGCTCCAGTCCGCGTTCCAGTGTGGATGATGCGCTGGACTGCAGC
GATGCCGCCCAATCACAATCAGGTGGTGCAGCATCCGCAGCTGAGTGTGGTG
TCCGTGTCACCAGTTCGCTCGCCCCAGCCCTCCACCAGCAGCCATCTGAAGCGA
CAGATTGTGGAGGATATGCCCGTGCTGAAGCGCGTGCTGCAGGCTCCCCCTCTGT
ACGATACCAACTCGCTGATGGACGAGGCCTACAAGCCGCAC

# >E78 d1

TGTCACGACGCCTTGGCGGGAACGGCAAACGAGCTGACCGTCTACGATGTCATC
ATGTGCGTGTCGCAGGCGCACCGCCTCAACTGCTCCTACACGGAGGAACTGACC
AGAGAGCTCATGCGTCCCGTGACGGTGCCACAAAATGGGATTGCCAGCACA
GTGGCCGAGAGTCTGGAGTTCCAGAAGATCTGGCTGTGGCAACAGTTCTCGGCC
AGGGTGACGCCTGGCGTTCAGCGGATTGTGGAGTTTGCGAAACGCGTACCTGGC
TTCTGTGATTTCACCCAAGATGACCAG

#### >E78 d2

CAGCTTGAGATACTCTACGATTCTGACTTTGTCAACGCCTTGCTGAACTTTGCCA ACACGCTGAACGCCTACGGGCTGAGTGACACCGAAATCGGACTCTTCTCGGCCA TGGTGCTGCTTGCCTCGGATCGAGCTGGACTCAGCGAGCCCAAGGTGATCGGCA GGGCCAGGGAACTGGTGGCCGAGGCGCTGCGCGTACAGATCCTGCGTTCGCGG GCAGGATCCCCACAGGCGCTGCAGCTGATGCCGGCGCTGGAAGCCAAGATACC CGAGCTGAGATCCTTGGGGGCCCAAGCACTTCTCACACCTAGACTGG

#### >DHR3 d1

ACAATAATCGATCCCGAATTTATTAGTCACGCGGATGGCGATATCAACGATGTG
CTGATCAAGACGCTGGCGGAGGCGCATGCCAACACAAATACCAAACTGGAAGC
TGTGCACGACATGTTCCGAAAGCAGCCGGATGTGTCGCGCATTCTCTACTACAA
GAATCTGGGCCAAGAGGAACTCTGGCTGGACTGCGCCGAGAAGCTTACACAAA
TGATACAGAACATAATCGAATTTGCTAAGCTCATACCGGGATTCATGCGCCTAA
GTCAGGACGATCAGATATTACTGCTGAAGACGGGCTCCTTTGAGCTGGCGATTG
TTCGCATGTCCAGACTGCTTGATCTCTCACAGAACGCGGTTCTCTACGGCGACGT
GATGCTGCCCCAGGAG

#### >DHR3 d2

TCGGAAGAGATGCGTCTGGTGTCGCGCATCTTCCAAACGGCCAAGTCGATAGCC
GAACTCAAACTGACTGAAACCGAACTGGCGCTGTATCAGAGCTTAGTGCTC
TGGCCAGAACGCAATGGAGTGCGTGGTAATACGGAAATACAGAGGCTTTTCAAT
CTGAGCATGAATGCGATCCGGCAGGAGCTGGAAACGAATCATGCGCCGCTCAA
GGGCGATGTCACCGTGCTGGACACACTGCTGAACAATATACCCAATTTCCGCGA
TATTTCCATCTTGCACATGGAATCGCTGAGCAAGTTCAAGCTGCAGCACCCG

### >EcR\_d1

AAGAAGGAGATTCTTGACCTTATGACATGCGAGCCGCCCCAGCATGCCACTATT
CCGCTACTACCTGATGAAATATTGGCCAAGTGTCAAGCGCGCAATATACCTTCCT
TAACGTACAATCAGTTGGCCGTTATATACAAGTTAATTTGGTACCAGGATGGCTA
TGAGCAGCCATCTGAAGAGGATCTCAGGCGTATAATGAGTCAACCCGATGAGA
ACGAGAGCCAAACGGACGTCAGCTTTCGGCAT

### >EcR d2

GTCTTCTACGCAAAGCTGCTCTCGATCCTCACCGAGCTGCGTACGCTGGGCAACC AGAACGCCGAGATGTGTTTCTCACTAAAGCTCAAAAAACCGCAAACTGCCCAAGT TCCTCGAGGAGATCTGGGACGTTCATGCCATCCCGCCATCGGTCCAGTCGCACCT TCAGATTACCCAGGAGGAGAACGAGCGTCTCGAGCGGCTGAGCGTATGCGGG CATCGGTTGGGGGCGCCATTACCGCCGGCATTGATTGCGACTCTCCACTTC G

#### >DHR78 d1

ACTCGCCAGCTAGCGGATATCGATAAGATCGAACCGTTGAAGATCTCGAAGATG GCAAATCTCACCAGGACCCTGCACGACTTTGTCCAGGAGCTCCAGTCACTGGAT GTTACTGATATGGAGTTTGGCTTGCTGCGTCTGATCTTCAATCCAACGCT CTTGCAGCAGCGCAAGGAGCGGTCGTTGCGAGGCTACGTCCGCAGAGTCCAACT CTACGCTCTGTCAAGTTTGAGAAGGCAGGGT

#### >Dis d1

AACTACTCCTCGCCCTCGCCCAGCAACTCCATCCAGTCCATCTCGAGCATTGGAT CGCGCAGCGTGGTGGCGAGGAGGGCCTCAGCCTGGCAGCGAGAGTCCGCGC GTCAATGTGGAAACGGAGACACCTTCGCCATCGAACTCGCCGCCCCTTAGTGCT GGTAGCATTTCGCCAGCGCCCACGTTGACCACCTCGTCGGGATCGCCGCAGCAC CGCCAGATGTCGCGGCACAGCCTCAGT

## >Dis\_d2

CAACAGCTGTTGGACTCGCGGCTGCTCTCCTGGGAGATGCTGCAGGAGACGACG GCGCGACTGCTCTTCATGGCGGTGCGCTGGGTCAAGTGCCTCATGCCCTTCCAGA CGCTCTCCAAGAACGACCAGCATTTGCTGCTCCAGGAATCCTGGAAGGAGCTCT TCCTGCTCAACCTCGCCCAATGGACTATACCGCTGGATCTAACGCCCATACTGG AATCACCGCTCATCCGCGAACGGGTGCTGCAGGACGAGGCCACACAAACGGAG ATGAAGACGATCCAGGAG

#### >ERR d1

ATGTCCGACGCGTCAGCATCTTGCACATCAAACAGGAGGTGGACACTCCATCGGCGTCCTGCTTTAGTCCCAGCTCCAAGTCAACGGCCACGCAGAGTGGCACAAAC

GGCCTGAAATCCTCGCCCTCGGTTTCGCCGGAAAGGCAGCTCTGCAGCTCGACG ACCTCTCTATCCTGCGATTTGCACAATGTATCCTTAAGCAATGATGGCGATAGTC TGAAAGGA

### >DHR38 d2

TCGATGAGCGAGGCAGATAAGGTGCAACAGTTTTACCAGCTGCTGACCAGCTCC
GTGGACGTGATCAAGCAGTTCGCCGAGAAGATTCCCGGCTACTTCGATCTCCTG
CCGGAGGATCAGGAGCTGCTCTTCCAGAGCGCATCGCTGGAACTGTTCGTCCTG
CGGCTGGCCTATCGCGCCAGGATCGATGACACCAAGCTGATCTTCTGCAACGGC
ACGGTGCTCCACCGCACCCAGTGCCTGCGCTCCTTCGGCGAGTGGCTCAACGAC
ATCATGGAGTTCAGCCGCAGCCTGCACAACCTGGAGATCGACATCTCCGCCTTC
GCCTGCCTCTGTGCCCTAACCCTGATCACAGAACGCCATGGCCTGCGGGAGCCG
AAGAAGGTGGAGCAGCTCCAGATGAAGATCATTGGCAGTCTG

#### >ftz-f1 d1

#### >DHR39 d1

#### >DHR39 d2

GCAAGTCAGCAGCAGCCGCACCAGCGACTACATCAACTAAATGGATTTGGAGG TGTACCCATTCCCTGCTCTACTTCTCTCCAGCCAGCCCTAGTTTGGCAGGAACTT CGGTCAAGTCGGAAGAGGTGGCGGAGACGGGCAAGCAAAGCCTCCGAACGGG AAGCGTACCACCACTACTGCAGGAAATCATGGATGTAGAGCATCTGTGGCAGTA CACCGATGCAGAGCTGGCCCGCATCAACCAACCACTGTCC

>DHR4 d1

GAGCGGGACCGGGAACGGGAACGGGAACAGTCCATCAGCTCCTCGCA GCAGCACCTAAGTCGGGTCTCCGCCAGTCCACCCACTCAGCTGTCCCACGGCAG CCTGGGACCCAACATTGTGCAGACGCACCATCTTCACCAGCAACTCACACAGCC GCTGACGCTGCGCAAGAGCAGCCCGCCCACAGAGCACCTGCTCAGTCCAT GCAACATCTCACA

# >DHR4 d2

AGTCGAGCCTCTCCCGATTCGCTGGAAGAGAGCCCTCTACCACAACGACCACA GGTCGTCCAACGCTCACGCCCACGAATGGGGTGCTGTCCTCCGCCTCGGCGGGC ACGGGGATTTCCACAGGAAGCAGCGCCAAGCTGAGCGAGGCTGGTATGAGTGT GATACGGTCCGTGAAGGAGGAGCGCTTGCTCAACGTATCCAGCAAGATGCTGGT GTTCCATCAGCAGCGGGAGCAAGAG

### >BRC d1

GCAGAGGACACACACACCCATCTGGCTCAGATACAGAACCTGGCCAATTCCGG CGGCCGCACGCCGCTCAACACGCACACCCAATCGCTCCCACATCCGCACCATGG CTCGCTACACGACGACGGCGGCTCCTCGACCCTCTTCAGCCGCCAGGGAGCGGG TTCGCCACCGCCGACGGCGGTGCCATCGCTGCCCTCGCACATCAACAACCAGCT GCTGAAGCGCATGGCCATGATGCACCGCAGCAGT

### >BRC d2

GCCAATGCCAACGACGAACACAGCAACGACTCCACCGGCGAGCACGATGCCAA
TCGCTCCAGTAGCGGCGACGGCGCAAGGGATCCCTGAGCTCCGGCAACGACG
AGGAGATCGGCGACGGACTCGCCTCCCATCATGCCGCCCCTCAGTTCATCATGT
CGCCGGCGGAGAACAAGATGTTCCATGCAGCCGCCTTCAACTTTCCCAATATCG
ATCCCTCAGCGTTGCTAGGTCTCAACACACAGTTGCAGCAGTCCGGTGACCTCG
CCGTGTCCCCT

#### >E74 d1

GTCAGCTACGATCTCCTACATGCTGGAGCTGGGCGGATTCCAGCAGCGGAAG GCGAAGAAGCCGCGCAAACCGAAGCTGGAGATGGGCGTAAAGCGGCGCAGCC GGGAGGGATCCACCACCTATCTGTGGGAGTTCCTCCTCAAACTGCTCCAGGATC GCGAATACTGTCCGCGTTTCATCAAGTGGACGAACCGGGAGAAGGGCGTCTTCA AGCTGGTCGACTCGAAG

#### >E74 d2

CAAGGTGGATGGCCAGCGGCTGGTCTACCAGTTCGTGGATGTGCCCAAGGACAT CATCGAGATTGACTGCAACGGT

# >E93 d1

ACGCCCAACGCCTGAAACTGCCCCTTTTCGAGGCGGGTCCACAGGCGTTATCC
TTTCAGCCGAACATGTTCTGGCCCCAGACGAACGCCACGAATGCCTACGGCCTG
GACTTCAATCGCATCACGGAGGCGATGCGGAATCCCCAGGCCTCCAATCACCAC
GGCCTGATGAAGAGTGCCCAGGACATGGTGGAGAACGTGTACGATGGCATCAT
CAGGAAGACGCTG

### >E93 d2

AGGCACAACTGCGCAAACTGAGCCACCTGTCCGAGCACAATGGCAGCGATCT GGGCGAGGATGTGGATCGTGGATCGCCGAAAATGGGGCGACATCCGGCCTGTG GCAATGCCAGTGCCAATCAGGGCGCACCGCCATCCATTCCGCTGGATGCCAATG TCCTGCTGCACACTCTGATGCTGGCTGCTGGGATTGGTGCAATGCCGAAGCTGGA TGAAACGCAAACGGTGGGCGACTTTATCAAGGGTCTGCTGGTGGCCAACAGTGG TGGC

# >mld\_d1

CAAGTGAGCGAGCTGCGAACGAGTCACCATTGTCTGTACTGCGAGGAGCGATTC
ACCAACGAAATTTCCCTGAAGAAGCACCATCAGCTGGCGCACGGAGCGCTGAC
AACAATGCCATATGTGTGCACAATCTGCAAGCGAGGGTATCGCATGCGTACGGC
GCTGCATCGGCACATGGAGTCGCACGATGTGGAGGGACGGCCGTACGAGTGCA
ACATTTGTCGCGTCCGGTTCCCGCGACCATCGCAACTGACGCTGCACAAGATCA
CGGTGCACCTGCTCTCCAAGCCGCACACTGCGACGAGTGTGGCAAACAGTTTG
GCACGGAGAGCGCCCTCAAGACGCACATTAAGTTCCACGGAGCTCACATGAAA
ACCCATTTGCCGCTGGGCGTATTCCGCAACGAGGAT

#### >salm/salr d1

>salm/salr d2

TGCAATGCGATGAACCAGATCGCCCAGTCCGTAATGCCGGCGGCTCCATTCAAC CCACTGGCACTCAGCGGTGTTCGCGGCAGCACCACCTGCGGCATCTGCTACAAG ACATTCCCCTGCCACTCGGCGCTGGAGATCCACTACAGGAGCCACCAAAGA GCGGCCATTCAAGTGCAGCATCTGTGATCGCGGCTTTACGACCAAGGGCAACCT GAAGCAACACATGCTAACTCATAAAATCCGCGATATGGAGCAAGAAACCTTCA GAAATCGTGCCGTAAAGTATATGAGTGAGTGGAACGAAGATCGC

## >ac\_d1

GGACCCTCTGTTATCCGGAGAAATGCCCGGGAACGCAACCGCGTAAAGCAGGT CAACAATGGCTTCAGCCAACTACGACAACATATCCCTGCGGCCGTAATAGCCGA TTTAAGCAATGGTCGCCGGGGAATTGGTCCCGGCGCCAATAAAAAACTGAGCA AAGTTAGCACACTGAAAATGGCAGTAGAGTACATACGGCGCTTGCAGAAA

#### >ac d2

GAAAACGACCAGCAGAAACAGAAACAGTTGCATTTGCAGCAGCAGCAACATTTGCA CTTTCAGCAGCAGCAACAGCATCAACACTTATACGCCTGGCACCAAGAGTTGCA GTTGCAATCTCCAACTGGCAGCACAAGTTCCTGCAACAGCATTAGCTCTTATTGC AAGCCAGCAACATCGACGATTCCGGGAGCAACACCTCCTAACAATTTTCATACC AAGTTGGAAGCCAGTTTTGAAGACTACCGTAACAATTCCTGCAGTTCTGGTACTG AAGATGAGGACATCCTCGACTATATATCACTCTGGCAGGACGAC

#### >sc d1

GCTCCATATAATGTAGACCAATCCCAGTCGGTCCAAAGGCGCAATGCTAGAGAA CGAAATCGTGTAAAGCAGGTGAACAACAGCTTCGCCAGGTTGCGGCAACATAT ACCACAATCCATAATCACGGATTTGACAAAGGGTGGTGGTCGAGGACCTCACAA AAAGATCTCCAAAGTAGACACACTGCGCATTGCCGTCGAGTACATCCGGAGGCT TCAGGATCTGGTGGATGACCTAAATGGGGGCAGCAATATTGGTGCCAACAATGC AGTCACCCAG

### >l(1)sc d1

GAGCAATTGCCATCGGTAGCCAGACGAAATGCCCGTGAACGCAATCGCGTGAA GCAGGTGAACAATGGATTCGTCAATCTCCGCCAGCATTTGCCTCAAACTGTGGT AAACTCGCTGTCCAATGGAGGACGTGGTAGCAGCAAGAAGTTATCCAAGGTGG ACACACTGCGAATCGCCGTTGAATATATTCGAGGACTACAGGACATGCTTGATG ATGGCACTGCT

# >l(1)sc d2

ACTCGTCACATCTACAATTCCGCCGATGAAAGTAGCAACGATGGCAGCAGCTAT AACGATTACAACGATAGTTTGGACAGTTCGCAACAGTTCTTGACGGGAGCCACC CAGTCTGCCCAATCCCACTCGTACCACTCCGCCTCGCCCACGCCGTCGTACTCCG GATCCGAGATTTCCGGAGGTGGCTATATCAAACAGGAACTACAAGAGCAGGAC CTCAAATTCGACTCCTTTGATAGCTTCAGTGACGAGCAGCCAGATGACGAGGAG CTACTCGATTATATTTCATCTTGGCAAGAG

## >ase d1

### >Dsx d1

GTTTCGGAGGAGAACTGGAATAGCGACACGATGTCCGACTCGGACATGATCGA CTCAAAGAACGACGTCTGTGGCGGAGCCTCCAGTTCCAGCGGCAGCTCGATTTC GCCGAGGACACCACCGAACTGCGCCCGCTGCCGCAATCATGGCCTAAAGATCA CCCTAAAGGGACACAAGCGGTACTGCAAGTTCCGCTACTGCACGTGCGAGAAGT GCCGACTGACGGCGGACCGCCAGCGGGTGATGGCTCTGCAAACGGCCTTGAGG CGAGCCCAGGCGCAGGATGAGCAGCGGGCGCTGCACATGCACGAG

#### >Dsx d2

### >Ovo/Svb d1

GCCTACGGCATAATACTCAAGGATGAACCGGACATTGAGTACGACGAGGCCAA GATCGATATTGGCACCTTTGCGCAGAACATTATCCAGGCAACGATGGGCAGCTC CGGTCAGTTCAATGCCAGCGCCTATGAGGATGCTATAATGTCGGACCTGGCCAG TTCGGGTCAGTGCCCAATGGAGCCGTCGATCCGCTTCAGTTCACAGCCACTCTG ATGCTGAGTTCGCAGACCGATCATTTACTGGAGCAGCTGTCCGATGCCGTGGACT TGAGTTCATTCCTGCAAAGGAGCTGCGTG

>Ovo/Svb\_d2

GGTTTGCTGGCGCCATCGCCCACCGTTTCCGTTTTGAATGAGAGCAAAGTCTTGC AGCGGCGCTTGGCCTTGCCGCCGGATCTGCAGCTTGAGTTTGTGAACGGCGGCC ATGGCATTAAGAACCCGCTGGCCGTGGAGAATGCCCACGGTGGCCATCACCGA ATTCGCAACATCGATTGCATTGATGATCTCAGCAAGCATGGC

# >dFOXO d2

GGTGGCTTCCAATTATCGCCCGATTTCCGGCAACGCGCCTCATCCAATGCCAGTT CCTGCGGACGCCTGAGCCCCATTAGGGCGCAGGATCTTGAGCCCGACTGGGGAT TCCCCGTTGACTACCAGAACACAACGATGACGCAGGCCCACGCCCAGGCGCTC GAGGAGCTGACGGGCACAATGGCGGATGAGCTGACGCTGTGCAACCAGCAGCA GCAAGGGTTCAGTGCCGCCTCGGGACTTCCCTCTCAG

#### >ey d1

# >ey\_d2

GCGATGTACTCCAACATGCATCATACGGCGTTATCCATGAGCGATTCATATGGG GCGGTTACGCCGATTCCGAGCTTTAACCACTCAGCTGTCGGTCCGCTCGCCC CATCGCCAATACCGCAACAGGGCGATCTTACCCCTTCCTCGTTATATCCGTGCCA CATGACCCTACGACCCCCTCCGATGGCTCCCGCTCACCATCACATCGTGCCGGGT GACGGTGGCAGACCTGCGGGCGTTGGCCTAGGCAGTGGC

# >toy\_d1

### >toy\_d2

TCCTCATTAGGATCAATGACCCCGTCATGCTTACAACAACGCGATGCCTATCCTT ACATGTTCACGATCCGTTATCACTAGGATCTCCCTATGTGTCAGCCCACCATCG AAACACAGCTTGCAACCCCTCAGCTGCGCACCAACAGCCCCCTCAGCATGGCGT TTATACCAATAGTTCTCCAATGCCATCATCAAACACAGGTGTCATTTCTGCGGGC GTTTCGGTGCCTGTCCAGATTTCAACGCAAAATGTATCTGACCTAACGGGAAGC AATTACTGGCCACGTCTT

#### >Stat92E d2

GGCATGTGGAAAGCAGGCTGCATTATGGGCTTCATCAACAAGACCAAGGCTCA GACTGATCTGCTGCGTTCAGTCTATGGTATCGGCACTTTCCTGCTCCGTTTCTCCG ACAGCGAGCTCGGTGGAGTCACTATCGCCTACGTAAACGAAAATGGACTGGTCA CCATGCTAGCGCCATGGACTGCACGGGATTTCCAGGTGCTGAACCTGGCCGATC GCATTCGAGATCTGGACGTGCTTTGCTGGCTGCACCCTAGCGACCGCAATGCGA GTCCCGTGAAGAGGGACGTCGCCTTCGGTGAGTTCTACTCAAAGCGTCAA

## $>Rx_d1$

CAGGAGAAATCGGAGAGTCTTCGCCTGGGCCTGACCCACTTCACCCAGCTGCCA CATCGCCTGGGATGCGGTGCATCCGGACTGCCCGTGGATCCCTGGCTATCACCA CCACTGCTCAGCGCATTGCCAGGCTTTCTCTCGCATCCGCAGACTGTGTACCCAA GCTATCTGACGCCGCCGCTCAGCCTGGCGCCCGGAAATCTGACCATGAGCAGTC TGGCGGCCATGGGCCACCACCATGCCCACAATGGGCCGCCCC

## >hbn d1

TTTATGAATCAGGACAAGGCCGGTTACCTTCTGCCCGAACAAGGTCTGCCGGAG
TTTCCGCTGGGCATTCCCCTGCCGCCACATGGACTTCCGGGTCATCCGGGCTCGA
TGCAGTCGGAGTTCTGGCCCCCGCACTTTGCCCTCCACCAGCACTTTAATCCCGC
TGCAGCCGCTGCCGCCGGCTTGCTCCCCAGCACCTGATGGCGCCCCACTACAA
GCTGCCCAACTTCCACACCCTGCTCTCCCAGTACATGGGCCCTGAGCAATCTGAAT
GGGATCTTTGGCGCG

## >otp\_d1

AAGACAACCAATGTCTTCCGCACCCCGGGCGCCCTGCTGCCCTCCCATGGACTT
CCGCCGTTTGGGGCCAATATAACCAATATCGCCATGGGCGATGGTCTCTGTGGC
ACGGGAATGTTTGGCGGAGATCGCTGGAGCGTGGGTGTCAATCCAATGACGGCA
GGCGACTCCATGATGTACCAGCACAGTGTGGGCGGAGTCAGTTGTGGGCCCAGT
GGTTCGCCGAGCGCCACCACCCCGCCGAACATGAACAGCTGCTCCTCGGTGACC
CCGCCGCCACTTTCCGCGCAGCCGAACTCCAGCCAAAACGAGCTGAACGGCGA
GCCCATGCCGCTGCAC

### >dwg\_d1

>dwg d2

# 3. Drosophila transcription factor gene fragments (expression optimized):

#### >AB1

CGTAGTCCGATTAGTCTGGAACATGATCGCCAAAAAAGTAGTCCGGTTCGTGTT
AAAAGTTTTAGTATCGCGGACATCCTGGGCCGTGGTCACGAGGAGGACCGTGTT
GAGAAGAAGCCGGAACGCATCGCGATCCCGAACGCGCTGCCAGGTGTGCCAGC
GCCGCTGGCCCTGATCAACGAACGTCTGCAAATTCCACTGGTTCCGAACTGTCC
ACCGCCAGCGGCCCTGCATACGTTTCTGAGTCCAGCGCTGTGT

#### >AB11

GCGTACCATCTGGGTAGTCACCTGGGCCAGACGAGTACCCCAACCACGACCCA
AGCCACCCTGAGCGGTCAAATGGACAAATTTGCGCTGGAACGTAGTAGCTATCT
GAGCAATAGCAGTCAAGCCAGTGCGTATAGCGCGTACCTGGACCCGAGTGTTCT
GACCAAAGCCTATTTCGACAGCAAAATGTATCAGGACCGTGCCGCCAATTATGC
GTTTGATATTAGCAAAAATCTATGGTGCCCAACAACATGCCGCGGCCCATCATCA
ATGC

#### >AC12

#### >AF4

#### >AG3

GGCCAATGGACCGAGGACGACTATCAGCGCGCGCAAAGCAGCAGTGTGAGCCC
ACGTGGCGGCTACTGCAGTAGCGCGAGTACCCCGCATAGCAGCGGCGGCAGTT
ACGGTGCGACGGCGGCCAGTGCCGCCGTGGCGGCGACGGCGAATGGCTACGCC
CCAGCCCCGAATATGGGTACCCTGAGTAGTACTCCAGGCAGCGTGTTTAACAGT
ACGAGTCGCGTTAGCAGCCTGAGTTTTAATCCATTTGCCCTGCCGACCTGTAATA
CGCAAGGTTACAGCACCCAACTGGTTACGAGTACCAAGTG
T

#### >AR1

CTGTATCTGTGTCTGAGTCGTGACGATGCGATTAGCACGTGCATCTGCACCGAAT GTTGCAGCCAGCTGGAGAGTTTCCACAACTTTTGGAAGCTGGTTGAACTGAAGC AGACCACCCTGTGCAGTCAATTTCTGGCCATTGACTGCGATGTTAACTGGAGCG AAGATGGCAGTGAAACGCAGCTGGACGCGCAACCGCAGCTGCTGCTGGAGCCA GCGGAGGAGCCAAAGGTTGTTACGCCAACCACGGCCAATAAATTCCCATGTATG TTTTGCGAAAAAAAGCTTCTGC

#### >B11

### >D5

GAGAATGGCATGATGTATACCAATCTGGACTGCATGTATCCGACCGCGCAAGCG CAAGCCCCAGTTCATGGCTATGCGGGTCAAATTGAGGAGAAGTATGCCGCGGTG CTGCACGCGAGTTACGCGCCAGGTATGGTTCTGGAAGATCAAGACCCAATGATG CAGCAAGCGACGCAAAGTCAGATGTGGCATCATCAGCAACACCTGGCGGGCAG TTACGCGCTGGATGCCATGGATAGC

# >F9

CTGGAAATTAACGATTTTCCACAACAAGCCCGTTGGAAAGTTACGAGTAAAGAA GCCCTGGCGCAGATCAGCGAATATAGTGAGGCGGGTCTGACGGTGCGCGGTAC GTATGTTCCACAAGGTAAAAATCCACCAGATGGTGAACGTAAACTGTATCTGGC CATTGAAAGTTGTAGTGAACTGGCCGTTCAAAAAGCGAAACGCGAAATTACGC GCCTGATTAAAGAAGAACTGCTGAAACTGAGTAGTGCGCATCATGTTTTTAATA AAGGTCGTTATAAAGTGTGT

#### >K1

CACGACTTCCATCCGGCGTATCGTATCCCGGGCTACATGGAGCAGCTGTACAGC CTGCAGCGCACGAATAGTGCGAGTAGCTTTCATGATCCATATGTTAACTGTGCCA GCGCGTTCCACCTGGCCGGTCTGGGTCTGGGCAGCGCCGACTTCCTGGGTAGCC GTGGCCTGAGTAGTCTGGGTGAACTGCATAATGCGGCGGTGGCCGCCGCGTGT

#### >K3

GCCAGCAACAGTAGTACCAGCAGTGAAGCGAGCAATAGTAGTCAACAGAATAA TGGTTGGAGTTTTGAGGAACAATTTAAACAAGTGCGCCAACTGTATGAGATTAA CGACGACCCAAAGCGTAAAGAATTTCTGGATGATCTGTTCAGTTTTATGCAGAA ACGCGGTACGCCGATTAACCGTCTGCCGATTATGGCCAAGAGTGTTCTGGATCTG TATGAACTGTATAATCTGGTGATCGCCCGTGGTGGTTGTGGATGTGC

#### >K4

AGTGCCCATACCAGTCCACAACAACGTGAAGCCCTGAATCTGAGTGATAGTCCG CCAAATCTGACGAATATTAAACGTGAACGTGAGCGCGAACCAACGCCGGAGCC AGTGGACCAGGACGACAAGTTCGTGGACCAGCCGCCACCGGCCAAACGTGTTG GTAGCGGTCTGCCGCCGGGCTTTCCAGCGAATTTTTACCTGAACCCACACAA TATGGCGGCCGTGGCCGCGTGT

#### >K5

#### >1.5

GGCGGTAACAGTCATCCGGCGATCGGCGTTATGAGTCTGTTTGATCCACGTTACA TGGACAGCCCGGGTAACGTTCACAGTATGACCCGTCAGCGCTACATCGAAGCCA CCGGTGGTGGCGCGGGTGCCGGCACGGGCGCGGGTACCGTACCATCAACGTT AACCCAACCACGGCGATGAATCTGCAGAGCATCAACTTCAACAGTTGT

#### >M6

GGCGCCCGGTAATGCGAACGCCGCAATGCGAACGGTCAAAA CAACCAGCCGGTGGCATGCCGGTTCGCCCAAGTGCGTGTACCCCGGATAGTCG CGTTGGTGGTTATCTGGATACCAGCGGCGGTAGTCCAGTTAGCCATCGTGGTGGT AGTGCGGGTGAACGTTAGTGTTAGCGGTGGTAATGGTAACGCGGGCGCGTT CAGAGTGGCGTGTGT

# >bcd\_d2

#### >cad d1

CGCACCAGCCCAAGCAAGCCGCCGTACTTCGATTGGATGAAAAAAACCGGCCTAT
CCAGCGCAACCACAACCGGGCAAAACCCGTACGAAAGACAAGTACCGTGTGGT
TTACACCGACTTTCAGCGCCTGGAACTGGAAAAAGAATACTGTACGAGCCGTTA
CATTACCATCCGTCGTAAGAGTGAGCTGGCCCAAACCCTGAGCCTGAGCGAACG
TCAGGTTAAGATTTGGTTTCAGAACCGCCGCGCCAAAGAACGTAAGCAGAATAA
GAAGGGTAGCGATCCAAACGTGATGGGTGTT
>hb d2

AAACATAAGAATCAGAAACCGTTTCAATGTGACAAGTGCAGTTATACCTGTGTT
AATAAAAGTATGCTGAATAGTCACCGTAAAAGCCACAGTAGTGTGTACCAATAC
CGTTGTGCCGATTGTGACTATGCCACCAAGTATTGCCACAGCTTTAAGCTGCACC
TGCGTAAATACGGTCATAAACCAGGCATGGTTCTGGATGAAGATGGCACCCCAA
ATCCAAGTCTGGTTATTGACGTGTATGGTACCCGCCGCGGTCCAAAAAGTAAAA
ATGGTGGTCCGATTGCCAGTGGTAGTGGTAGTGGCAGCCGCAAAAGCAATG
TTGCGGCCGTTGCGCCG

#### >lab d1

#### >lab d2

GCGGCCCATCAGCACCATCAGAATAGCGTTAGCCCGAATGGCGGTATGAAT CGTCAACAACGTGGCGGTGTTATTAGTCCGGGTAGTACCAGTAGCAGTACC AGTGCCAGTAATGGCGCGCACCCGGCGAGCACCCAAAGCAAAAGCCCGAACCA TAGTAGCAGCATCCCGACGTACAAGTGGATGCAGCTGAAACGCAATGTTCCAAA GCCGCAAGCCCGAAACTGCCGGCGAGTGGCATCGCCAGTATGCATGACTATCA GATGAACGGCCAGCTGGATATGTGTCGT

## $>pb_d2$

AACAGCAATAGTAAAAAAGAGTTGCCAAGGCTGTGAACTGCCAAGCGACGATAT
TCCAGATAGCACCAGCAATAGTCGTGGCCACAATAATAATACGCCGAGTGCGA
CGAATAATAATCCGAGCGCGGGCAACCTGACGCCGAATAGCAGTCTGGAAACC
GGTATTAGTAGCAACCTGATGGGTAGTACGACGGTGAGTGCGAGCAACGTGATC
AGTGCGGACAGCAGTGTTGCGAGCAGTGTGAGTCTGGACGAAGACATTGAGGA
GAGTAGCCCAATTAAAGTTAAAAAAAAAGGATGACGGCCAGGTGATTAAGAAAG
AAGCGGTTAGCACCAGTAGTAAGGCGAGTCCATTCGGTTATGAAAATAGTACGC
CAAGTCTGGTTAGTTTTCGTCGTGATAGCGACGCCAGC

### >Dfd d1

#### >Scr d2

AACCATAGTGGCAGCGGCGTTAGTGGCGGTCCGGGTAACGTGAATGTGCCAATG CACAGTCCGGGCGGTGGCGACAGCGACAGTGAAAGTGATAGTGGCAATGAAGC CGGTAGCAGTCAAAACAGTGGCAACGGTAAAAAAAATCCACCGCAAATCTACC CATGGATGAAGCGCGTTCATCTGGGCACGAGCACCGTGAATGCCAATGGTGAA ACGAAACGTCAACGCACCAGCTATACG

### >Antp d1

ACGAGTTATTTCACCAATAGTTACATGGGCGCCGATATGCACCACGGTCATTAC
CCGGGCAACGGTGTGACGGATCTGGACGCGCAGCAAATGCACCACTATAGCCA
GAATGCCAATCATCAAGGCAATATGCCATATCCGCGCTTCCCACCGTATGATCG
CATGCCATATTATAATGGTCAAGGTATGGACCAGCAACAGCAACATCAGGTGTA
CAGTCGCCCAGACAGTCCGAGCAGTCAGGTTGGTGGTGTGATGCCACAAGCGCA
AACCAACGGCCAGCTGGGCGTGCCACAA

#### >lid d2

TGCGAATGTAATGACAAACTGATTGTGTGTCTGCGCCATTATACGGTTCTGTGTGGCTGTGCCCCAGAAAAGCATACCCTGATTTATCGCTACACCCTGGATGAGATGC

# >lilli\_d1

## >lilli\_d2

CACCGCGGTGATATTGCGAACGGTAATACGCCGAGTAGCATTAGTCCAAGTAAC
AGTGTGGGTAGTCAGGGTAGCGGCAGTAATACCCCGCCGGGCCGTATCGTGCCA
CCAGATATTCACAACATGCTGTGTAAACAGAATGAATTTCTGAGCTACCTGAAC
AGCGCGCATGAACTGTGGGATCAAGCCGATCGTCTGGTTCGTACCGGTAATCAT
ATTGATTTCATCCGCGAACTGGACCATGAGAACGGTCCACTGACGCTGCACAGT
ACGATGCACGAGGTTTTTCGTTACGTTCAAGCCGGCCTGAAAACCCTGCGTGAT
GCCGTTAGCCATCCGACC

### >E75 d1

#### >E78 d1

TGCCATGACGGTCTGGCGGGTACGGCCAATGAACTGACGGTGTACGACGTTATT
ATGTGCGTTAGCCAAGCGCATCGTCTGAACTGTAGCTACACCGAGGAACTGACG
CGCGAACTGATGCGCCGTCCGGTGACCGTGCCACAAAATGGCATTGCGAGCACG
GTGGCCGAAAGCCTGGAATTTCAAAAAGATCTGGCTGTGGCAGCAGTTCAGTGCC
CGAGTTACCCCGGGTGTGCAACGTATCGTGGAATTCGCCAAGCGTGTTCCAGGTT
TCTGTGATTTCACGCAGGACGACCAG

### >E78\_d2

CAACTGGAGATTCTGTACGATAGCGATTTTGTTAATGCGCTGCTGAACTTTGCGA ACACCCTGAATGCCTATGGTCTGAGTGATACCGAGATCGGTCTGTTTAGTGCGAT GGTGCTGCTGGCGAGCGATCGCGGGGCCTGAGTGAACCAAAAGTTATTGGTCG TGCGCGTGAACTGGTTGCCGAAGCGCTGCGTGTTCAAATTCTGCGTAGTCGCGCG GGTAGTCCACAAGCGCTGCAACTGATGCCAGCCCTGGAGGCGAAAATTCCAGA GCTGCGTAGTCTGGGCCCCAAACATTTTAGCCATCTGGATTGG

#### >DHR3 d1

ACGATCATTGATCCGGAATTTATCAGCCATGCGGACGGCGACATCAACGACGTT CTGATCAAAACGCTGGCCGAGGCCCATGCGAACACCAAGCTGGAGGC CGTGCATGACATGTTCCGCAAGCAGCCGGACGTGAGCCGTATTCTGTATTACAA GAATCTGGGCCAGGAGGAGCTGTGGCTGGATTGCGCGGAAAAGCTGACGCAGA TGATCCAAAACATCATTGAGTTCGCCAAGCTGATTCCAGGCTTCATGCGTCTGAG CCAGGATGACCAAATCCTGCTGCTGAAGACCGGCAGTTTCGAGCTGGCCATCGT GCGCATGAGTCGCCTGCTGGACCTGAGTCAAAACGCCGTGCTGTACGGCGACGT TATGCTGCCGCAGGAG

#### >DHR3 d2

### >EcR d1

AAAAAAGAGATTCTGGATCTGATGACGTGCGAACCACCACAACATGCGACGAT
TCCGCTGCTGCCGGATGAAATTCTGGCGAAATGTCAAGCGCGCAATATCCCAAG
CCTGACCTACAACCAACTGGCCGTGATTTACAAGCTGATCTGGTACCAGGACG
TTACGAACAGCCAAGTGAAGAAGATCTGCGTCGTATCATGAGTCAACCAGACG
AGAACGAAAGCCAGACGGACGTTAGTTTCCGCCAT

### >EcR d2

GTGTTTTATGCGAAACTGCTGAGCATTCTGACCGAACTGCGTACCCTGGGCAACC AAAATGCCGAAATGTGTTTTAGCCTGAAACTGAAGAATCGTAAACTGCCAAAAT TCCTGGAAGAAATTTGGGACGTGCATGCCATTCCACCGAGTGTGCAGAGTCATC TGCAAATCACCCAAGAAGAAAATGAACGTCTGGAACGCGCCGAGCGTATGCGT GCCAGTGTGGGTGCCATTACGGCCGGTATTGATTGTGACAGCGCCAGTACG AGT

#### >DHR78 d1

ACCCGCCAGCTGGCCGATATTGACAAGATTGAACCGCTGAAGATCAGCAAGAT GGCCAACCTGACCCGCACGCTGCATGACTTCGTGCAGGAACTGCAGAGTCTGGA CGTTACGGACATGGAATTTGGTCTGCTGCGCCTGATCCTGCTGTTCAACCCGACC CTGCTGCAGCAGCGCAAGGAGCGCAGTCTGCGTGGCTATGTTCGCCGTGTTCAA CTGTATGCGCTGAGTAGCCTGCGTCGTCAAGGT

### >Dis\_d1

AACTACAGCAGCCCGAGCCCGAGCAATAGTATCCAAAGTATTAGTAGCATTGGC AGTCGCAGCGGCGGCGAGGAAGGCCTGAGCCTGGGTAGTGAAAGTCCGCG CGTGAATGTGGAGACCCGAGACCCCAAGTCCGAGCAATAGTCCACCGCTGAGTG CGGGCAGTATCAGCCCAGCGCCAACGCTGACGACCAGCAGTGGTAGTCCACAG CATCGTCAAATGAGCCGCCACAGTCTGAGC

### >Dis\_d2

CAACAACTGCTGGACAGTCGCCTGCTGAGCTGGGAAATGCTGCAAGAAACGAC
CGCCCGTCTGCTGTTCATGGCGGTGCGTTGGGTTAAGTGTCTGATGCCATTCCAA
ACGCTGAGTAAGAATGACCAACATCTGCTGCTGCAAGAGAGTTGGAAAGAACT
GTTCCTGCTGAATCTGGCCCAATGGACGATCCCACTGGATCTGACGCCAATCCTG
GAAAGTCCACTGATTCGCGAGCGTGTTCTGCAGGATGAGGCCACGCAAACGGA
AATGAAAACGATTCAAGAG

### >ERR d1

ATGAGTGACGCGTTAGTATCCTGCACATTAAGCAGGAAGTGGATACCCCAAGT GCCAGTTGTTTTAGTCCGAGTAGCAAAAGTACGGCGACCCAGAGCGGCACGAAT GGTCTGAAGAGCAGTCCGAGTGTTAGCCCAGAGCGCCAGCTGTGCAGTAGTACG ACGAGCCTGAGTTGTGACCTGCACAACGTGAGTCTGAGTAACGATGGCGATAGT CTGAAAGGT

#### >DHR38 d2

AGCATGAGCGAGGCGGACAAGGTGCAGCAGTTCTACCAACTGCTGACCAGCAG
TGTTGACGTGATCAAGCAATTCGCCGAGAAGATTCCGGGTTACTTCGATCTGCTG
CCAGAAGATCAGGAGTTGCTGTTTCAGAGCGCCAGTCTGGAACTGTTTGTGCTGC
GTCTGGCCTACCGTGCCCGTATCGACGACACGAAGCTGATCTTCTGCAACGGCA
CCGTGCTGCATCGTACGCAATGCCTGCGTAGTTTCGGCGAATGGCTGAACGATAT
CATGGAATTTAGTCGCAGCCTGCATAATCTGGAGATTGATATTAGCGCGTTCGCC
TGTCTGTGCGCCCTGACCCTGATTACCGAACGCCATGGCCTGCGCGAACCGAAG
AAAGTGGAACAACTGCAAATGAAAATTATTGGTAGTCTG

# >ftz-f1\_d1

CGTAATAAATTCGGTCCAATGTACAAGCGTGATCGCGCCCGTAAGCTGCAGGTT
ATGCGTCAACGTCAACTGGCGCTGCAAGCGCTGCATAATAGCATGGGCCCGGAT
ATTAAACCAACCCCAATTAGCCCGGGTTACCAGCAGGCCTATCCGAACATGAAT
ATTAAACAAGAGATTCAAATCCCGCAAGTGAGCAGTCTGACCCAGAGCCCGGA
CAGTAGCCCGAGCCCGATTGCCATCGCCCTGGGTCAAGTGAACGCCAGTACGGG
CGGCGTGATTGCCACGCCGATGAATGCC

#### >DHR39 d1

ATGCCGAATATGAGCAGTATCAAGGCCGAGCAACAGAGTGGCCCGCTGGGCGG
TAGCAGCGCTACCAGGTTCCAGTGAACATGTGCACCACGACGGTGGCGAACA
CCACGACCACGCTGGGTAGTAGCGCCGGTGGTGCCACGGGTAGTCGCCACAAC
GTTAGCGTTACCAACATTAAGTGCGAACTGGATGAGCTGCCGAGTCCAAATGGC
AATATGGTGCCGGTGATCGCGAACTATGTGCACGGTAGCCTGCGTATTCCGCTG
AGTGGCCACAGTAATCACCGCGAGAGTGACAGTGAAGAAGAACTGGCCAGTAT
CGAGAATCTGAAAGTTCGCCGTCGCACGGCCGCCGATAAAAAATGGTCCGCGCCC
GATGAGTTGGGAGGGTGAACTGAGTGATACCGAAGTGAACGGCGGC

#### >DHR39 d2

GCCAGTCAACAACAACCACATCAGCGTCTGCACCAACTGAACGGTTTCGGTGGC GTTCCAATCCCATGCAGTACGAGCCTGCCAGCGAGTCCAAGTCTGGCGGGTACG AGTGTTAAAAGTGAGGAAATGGCCGAAACCGGTAAACAAAGCCTGCGTACCGG CAGCGTGCCACCGCTGCTGCAGGAAATCATGGATGTTGAACATCTGTGGCAGTA TACGGATGCCGAGCTGGCCCGTATTAACCAGCCACTGAGC

#### >DHR4 d1

GAACGTGATCGTGATCGCGAACGTGAGCGCGAGCAGAGTATTAGTAGTAGCCA ACAACACCTGAGTCGTGTGAGCGCCAGCCCGACGCAACTGAGCCATGGTA GTCTGGGCCCAAACATCGTTCAAACGCACCATCTGCATCAACAACTGACCCAAC CACTGACGCTGCGTAAGAGCAGTCCGCCAACCGAGCATCTGCTGAGCCAAAGC ATGCAACATCTGACG

#### >DHR4 d2

AGTCGTGCCAGTCCAGATAGTCTGGAAGAGAAACCAAGTACGACCACGACGAC CGGTCGTCCAACGCTGACGCCAACCAATGGTGTTCTGAGTAGCGCGAGTGCGGG CACGGGTATCAGTACCGGTAGTAGTGCCAAGCTGAGCGAAGCCGGCATGAGCG TGATTCGCAGCGTGAAAGAAGAACGTCTGCTGAATGTTAGTAGTAAAATGCTGG TGTTCCATCAACAGCGCGAACAGGAG

>BRC d1

GCGGAGGATACCCACAGTCATCTGGCGCAGATTCAAAATCTGGCCAATAGCGGT GGTCGCACCCCACTGAATACGCACACGCAGAGTCTGCCACACTCACACCACGGT AGTCTGCATGACGATGGCGGTAGCAGCACCCTGTTCAGCCGTCAGGGTGCCGGC AGCCCGCCCCAACCGCCGTGCCGAGCCTGCCAAGCCATATCAATAACCAACTG CTGAAGCGCATGGCGATGATGCATCGCAGTAGC

# >BRC d2

GCGAACGCGAACGATGAACACAGTAATGATAGTACCGGCGAACACGATGCGAA TCGTAGCAGCAGTGGCGACGGTGGCAAAGGCAGCCTGAGTAGTGGCAATGACG AGGAGATCGGCGATGGTCTGGCGAGTCACCACGCCGCCGCCAATTTATTATGA GTCCGGCGGAAAACAAAATGTTCCATGCGGCGGCGTTTAACTTCCCGAACATTG ATCCGAGCGCGCTGCTGGGCCTGAATACGCAACTGCAGCAAAGTGGTGATCTGG CGGTTAGTCCG

## >E74\_d1

GTTAGCTACGATCTGAGTTATATGCTGGAACTGGGCGGCTTTCAACAACGTAAG GCGAAGAAACCACGCAAACCAAAGCTGGAGATGGGTGTGAAGCGCCGCAGCC GTGAAGGTAGCACCACCTACCTGTGGGAATTCCTGCTGAAGCTGCTGCAGGATC GCGAGTACTGCCCGCGCTTTATTAAATGGACGAACCGCGAAAAAAGGTGTGTTTA AACTGGTGGATAGTAAA

#### >E74 d2

ACCACGTATCTGTGGGAATTTCTGCTGAAACTGCTGCAAGATCGTGAGTATTGTC CACGCTTATTAAGTGGACCAACCGTGAAAAAGGTGTTTTTTAAACTGGTTGATAG TAAAGCCGTGAGTCGCCTGTGGGGCATGCACAAGAACAAGCCGGACATGAATT ACGAAACGATGGGCCGTGCGCTGCGTTACTACTACCAACGCGGTATTCTGGCCA AAGTTGACGGTCAGCGTCTGGTGTATCAGTTTGTGGACGTGCCGAAGGACATCA TTGAAATCGATTGCAATGGC

## >E93\_d1

ACGCCAAATGGTCTGAAACTGCCACTGTTTGAAGCGGGCCCACAAGCGCTGAGT TTTCAGCCGAATATGTTCTGGCCGCAAACGAACGCCACGAATGCCTACGGTCTG GATTTCAATCGCATTACGGAAGCGATGCGTAATCCACAGGCGAGTAACCATCAC GGTCTGATGAAGAGTGCCCAAGATATGGTTGAAAAATGTTTATGATGGCATCATC CGCAAGACGCTG

#### >E93 d2

CGCGCGCAACTGCGCAAACTGAGCCACCTGAGCGAGCACAACGGCAGTGACCT GGGCGAGGACGTGGATCGCGGCAGTCCAAAAATGGGCCGCCATCCGGCGTGTG GCAACGCCAGCGCCAATCAAGGTGCGCCACCGAGCATCCCGCTGGATGCGAAT GTTCTGCTGCACACGCTGATGCTGGCCGCCGCCATCGCCGAAACTG GATGAGACGCAAACGGTGGGTGACTTTATTAAAGGCCTGCTGGTTGCGAATAGC GGTGGC

#### >mld d1

CAAGTGAGCGAACTGCGTACGAGTCATCATTGCCTGTATTGTGAGGAACGCTTC
ACCAACGAGATCAGTCTGAAAAAGCATCACCAACTGGCCCATGGCGCGCTGAC
CACGATGCCGTACGTGTACCATCTGCAAACGTGGCTACCGCATGCGTACCGC
GCTGCACCGTCACATGGAAAGCCACGATGTGGAGGGCCGCCCGTACGAATGCA
ACATCTGCCGTGTTCGCTTCCCGCGTCCGAGCCAGCTGACCCTGCACAAAATTAC
GGTGCACCTGCTGAGTAAGCCACACACCTGCGACGAATGCGGTAAGCAGTTCGG
TACCGAGAGTGCGCTGAAGACCCACATCAAATTTCATGGTGCGCATATGAAAAC
GCATCTGCCGCTGGGCGTGTTCCGCAACGAGGAT

#### >salm/salr d1

TTCTTCAACCCGATTAAACACGAAATGGCGGCCCTGCTGCCACGTCCGCACAGC
AACGACAACAGCTGGGAAAACTTTATCGAAGTTAGCAACACCTGCGAGACCAT
GAAACTGAAAGAACTGATGAAGAACAAGAAGATTAGTGACCCGAATCAGTGTG
TGGTTTGCGACCGCGTTCTGAGCTGTAAGAGTGCCCTGCAAATGCATTATCGTAC
GCACACCGGTGAACGTCCATTCAAGTGCCGTATCTGTGGTCGCGCGTTCACGAC
CAAAGGCAACCTGAAGACCCACATGGCGGTGCATAAAATCCGCCCACCAATGC
GTAACTTTCACCAATGCCGGTTTGTCACAAGAAATACAGCAATGCCCTGGTGC
TGCAACAGCACATTCGTCTGCACACGGGTGAGCCAACGGACCTGACCCCAGAG
CAAATCCAA

#### >salm/salr d2

TGTAATGCCATGAATCAGATTGCCCAAAGTGTTATGCCAGCGGCCCCGTTCAATC CACTGGCCCTGAGCGGTGTTCGTGGTAGTACGACGTGTGGCATTTGTTACAAGAC CTTTCCATGCCACAGTGCGCTGGAGATTCATTACCGTAGTCACACCAAAGAACG TCCGTTTAAATGCAGTATTTGTGATCGTGGCTTTACCACCAAGGGCAATCTGAAA CAACATATGCTGACCCATAAAATTCGTGATATGGAACAAGAAACGTTCCGTAAT CGCGCGGTGAAGTACATGAGTGAATGGAACGAAGACGT

### >ac d1

GGCCCAAGTGTTATTCGTCGTAACGCCCGTGAACGTAATCGCGTTAAACAGGTG AACAACGGTTTCAGTCAGCTGCGCCAGCACATTCCAGCCGCGGTGATTGCCGAT CTGAGCAACGGCCGCCGTGGCATTGGCCCAGGTGCCAACAAGAAGCTGAGTAA AGTTAGTACCCTGAAGATGGCGGTTGAATATATTCGCCGTCTGCAAAAA

>ac d2

#### >sc d1

GCCCCGTATAACGTTGATCAAAGTCAGAGCGTTCAGCGCCGTAACGCCCGTGAA CGTAACCGCGTTAAACAGGTTAATAACAGCTTTGCCCGTCTGCGCCAGCACATT CCACAGAGTATCATTACGGATCTGACCAAAGGCGGTGGTCGTGGTCCGCATAAA AAGATTAGCAAAGTGGACACGCTGCGTATCGCCGTTGAGTATATTCGTCGTCTGC AAGACCTGGTGGACGATCTGAACGGTGGTAGTAATATTGGCGCGAATAATGCGG TTACGCAG

#### >l(1)sc d1

GAACAACTGCCGAGTGTTGCCCGTCGCAATGCCCGTGAACGCAATCGTGTGAAA CAAGTTAATAATGGTTTTGTGAATCTGCGTCAGCATCTGCCGCAGACGGTTGTGA ACAGTCTGAGCAATGGCGGTCGTGGCAGTAGCAAAAAACTGAGCAAGGTGGAT ACGCTGCGTATTGCCGTTGAGTATATTCGTGGCCTGCAAGATATGCTGGATGATG GTACGGCC

#### >l(1)sc d2

ACGCGCCACATTTATAACAGTGCCGATGAAAGCAGTAACGACGGTAGTAGCTAC
AACGACTACAACGATAGCCTGGATAGCAGCCAACAGTTTCTGACGGGTGCGAC
CCAGAGTGCGCAAAGTCACAGTTACCACAGTGCCAGCCCAACCCCAAGCTACA
GTGGCAGTGAGATCAGCGGTGGTTGTTATATTAAACAAGAGCTGCAGGAGCAG
GATCTGAAGTTTGACAGTTTTGATAGTTTCAGCGATGAACAGCCAGACGATGAG
GAGCTGCTGGACTACATCAGTAGCTGGCAGGAA

### >ase d1

>Dsx d1

GTTAGTGAGGAGAACTGGAACAGTGATACCATGAGCGATAGTGACATGATCGA CAGCAAAAATGATGTTTGTGGTGGCGCCCAGTAGTAGTAGTAGTAGTATTAG CCCACGTACCCCGCCGAATTGTGCGCGCTGCCGTAACCACGGTCTGAAAATCAC GCTGAAAGGCCACAAACGTTATTGTAAATTTCGTTATTGTACGTGTGAAAAGTGT CGTCTGACGGCCGATCGCCAACGTGTTATGGCCCTGCAGACGGCCCTGCGTCGT GCGCAGGCCCAAGATGAACAACGTGCGCTGCACATGCACGAG

#### >Dsx d2

### >Ovo/Svb d1

GCCTACGGCATCATCCTGAAGGATGAACCAGATATTGAATATGATGAAGCCAAA ATCGACATCGGCACCTTCGCCCAAAACATCATTCAAGCCACCATGGGTAGTAGC GGCCAGTTCAATGCCAGCGCCTATGAGGACGCCATTATGAGTGATCTGGCCAGT AGCGGTCAATGCCCGAATGGTGCGGTGGACCCACTGCAGTTCACCGCGACGCTG ATGCTGAGTAGTCAGACCGACCATCTGCTGGAGCAGCTGAGTGACGCGGTTGAC CTGAGCAGTTTTCTGCAACGCAGCTGTGTT

### >Ovo/Svb d2

GGTCTGCTGCCCAAGTCCAACCGTTAGTGTGCTGAATGAGAGTAAAGTGCTG CAACGCCGTCTGGGCCTGCCACCAGACCTGCAGCTGGAATTTGTGAATGGCGC CATGGCATTAAGAACCCACTGGCCGTGGAAAATGCGCACGGCGCCACCACCG TATCCGTAACATTGATTGCATCGATGATCTGAGTAAACATGGT

### >dFOXO d2

GGCGGCTTTCAACTGAGTCCAGACTTCCGTCAACGTGCCAGTAGTAACGCCAGT AGTTGCGGCCGCCTGAGCCCAATTCGCGCCCAAGACCTGGAGCCAGATTGGGGT TTTCCAGTTGATTATCAAAATACGACGATGACCCAAGCGCATGCCCAAGCGCTG GAAGAACTGACGGGTACCATGGCCGATGAACTGACCCTGTGTAATCAGCAGCA ACAAGGCTTTAGTGCCGCGAGCGGTCTGCCAAGTCAA

### >ey\_d1

AGTACGAGTGCCGCAATAGTATCAGTGCGAAAGTGAGTGTTAGTATCGGTGGC AATGTTAGCAACGTGGCGAGCGGTAGCCGCGGTACGCTGAGTACCGAT CTGATGCAAACCGCCACCCACTGAATAGTAGTGAAAGCGGCGGTGCCAGCAA TAGCGGTGAGGGCAGTGAACAAGAAGCCATTTATGAGAAACTGCGTCTGCA ACACGCAACATGCCGCCGGTCCAGGCCCGCTGGAGCCAGCGCGTGCCGCCG CTG

### $>ey_d2$

GCCATGTACAGTAACATGCATCACACCGCGCTGAGTATGAGCGATAGCTATGGC GCCGTTACCCCAATTCCAAGTTTTAACCATAGTGCGGTTGGTCCGCTGGCGCCAC CAAGTCCGATCCCGCAGCAGGGTGATCTGACCCCAAGTAGTCTGTACCCATGTC ACATGACGCTGCGTCCACCACCGATGGCCCCAGCCCATCATCATATTGTTCCGG GCGACGGTGGTCCTCCAGCCGGCGTTGGCCTGGGCAGTGGT

## >toy\_d1

AGCGCGATTAATGTGGCCGAGCGTACGAGTAGTGCGCTGGTGAGCAACAGCCTG CCGGAAGCCAGCAATGGCCCAACGGTGCTGGTGAAGCCAATACGACCCA CACGAGTAGTGAGAGTCCGCCACTGCAACCGGCCGCCCACGTCTGCCACTGAA CAGCGGTTTTAATACCATGTACAGTAGCATTCCGCAGCCGATTGCGACCATGGC CGAAAACTATAAT

# >toy\_d2

AGTAGTCTGGGTAGTATGACCCCAAGTTGCCTGCAACAGCGTGATGCGTACCCA
TACATGTTCCACGATCCACTGAGCCTGGGCAGCCCGTATGTGAGTGCCCACCAT
CGTAATACGGCCTGTAACCCAAGTGCGGCCCCACCAACAGCCACCACAGCACGG
CGTGTACACGAACAGTAGCCCGATGCCAAGCAGTAACACGGGTGTTATTAGTGC
GGGTGTGAGTGTTCCGGTTCAGATCAGTACCCAAAATGTTAGTGATCTGACCGGT
AGTAATTACTGGCCACGTCTG

#### >Stat92E d2

GGTATGTGGAAAGCCGGTTGCATTATGGGCTTCATCAACAAAACGAAAGCGCAG ACGGATCTGCTGCGTAGTGTTTATGGTACTTTCTGCTGCGTTTCAGCGA CAGCGAACTGGGTGGTGACGATTGCCTACGTGAATGAGAACGGTCTGGTGAC CATGCTGGCCCCGTGGACGGCGCGTGATTTCCAAGTGCTGAATCTGGCGGACCG TATTCGTGACCTGGATGTGCTGTTTGGCTGCATCCGAGCGATCGTAATGCCAGC CCAGTTAAGCGTGACGTGGCCTTCGGTGAATTTTATAGTAAGCGTCAA

#### >Rx d1

CAAGAGAAGAGTGAGAGTCTGCGTCTGGGTCTGACGCATTTTACCCAGCTGCCA CATCGTCTGGGCTGCGGCCCAGTGGCCTGCCGGTTGATCCATGGCTGAGCCCG CCACTGCTGAGTGCCCTGCCGGGCTTCCTGAGTCATCCGCAAACGGTGTACCCG AGCTACCTGACCCGCCGCTGAGCCTGGCCCCAGGTAATCTGACCATGAGTAGC CTGGCCGCCATGGGCCATCATCATGCGCACAATGGTCCACCACCA

# >hbn\_d1

TTTATGAATCAGGATAAGGCCGGCTATCTGCTGCCAGAGCAAGGCCTGCCGGAA
TTTCCACTGGGTATTCCACTGCCACCACACGGTCTGCCAGGTCATCCGGGTAGTA
TGCAAAGCGAATTCTGGCCGCCACATTTTGCCCTGCATCAACACTTTAATCCGGC
CGCCGCCGCCGCGGGGTCTGCTGCCACAACATCTGATGGCCCCGCACTACAA
GCTGCCGAATTTCCATACCCTGCTGAGTCAGTATATGGGCCTGAGCAATCTGAAC
GGCATCTTTGGTGCG

# >otp\_d1

AAAACGACGAATGTTTTCGTACCCCGGGCGCCCTGCTGCCAAGTCATGGCCTG CCACCGTTCGGTGCCAATATCACCAATATTGCCATGGGCGACGGTCTGTGCGGC ACCGGCATGTTTGGTGGTGATCGTTGGAGCGTTGGTGAATCCAATGACCGCCG GCGACAGCATGATGTACCAGCACAGCGTGGGCGGCGTGAGTTGTGGCCCAAGT GGTAGTCCAAGCGCCACGACCCCACCGAACATGAATAGCTGCAGCAGTGTGAC CCCACCGCCACTGAGCGCGCAGCCGAATAGCAGTCAAAACGAGCTGAATGGTG AACCAATGCCACTGCAC

# >dwg\_d1

GCCTTTGATCATCTGCGCCGTCATAAACTGACCCACACGGGTGAGCGTCCGTAC GCCTGTGATCTGTGATAAAGCGTATTACGATAGTAGCAGTCTGCGCCAACAT AAAATCAGCCATACCGGCAAGAAAGCGTTTACGTGTGAAATCTGTGGTGTGGGT CTGAGTCAAAAAAGTGGCTATAAGAAACAC

### >dwg d2

# 4. Chip oligonucleotide sequences

>*LacZ* $\alpha$ :

LacZ-1,

MGRAAYTCNGARGARGCNMGRACNCATCATCATCACCACCACTGAGCGGCATG ACTCGACCATCCGATTTTTT

LacZ-2,

CKNGCYTCYTCNGARTTYCKCCANGANGCRAANGGNGGRTGNGCNGCYGCATG ACTCGACCATCCGATTTTTT

LacZ-3,

GARAAYCCNGGNGTNACNCARYTRAAYMGRYTRGCNGCNCAYCCNCCNGCATG ACTCGACCATCCGATTTTTT

LacZ-4,

GNGTNACNCCNGGRTTYTCCCARTCYCKYCKYTGYARNACNACNGCYAGCATGA CTCGACCATCCGATTTTTT

LacZ-5,

ACAGGAAACAGCTATGACNATGATHACNYTRGCNGTNGTNYTRCARMGGCATG ACTCGACCATCCGATTTTTT

LacZ-6.

TCATNGTCATAGCTGTTTCCTGTGTGAAATTAATGGGTAACATGATCCGCATGAC TCGACCATCCGATTTTTT

LacZ-7,

CGNAAYAGYGARGARGCNCGNACNCATCATCATCACCACCACTGAGCGGCATG ACTCGACCATCCGATTTTTT

LacZ-8.

CGNGCYTCYTCRCTRTTNCGCCARCTNGCRAANGGNGGRTGNGCNGCATGA CTCGACCATCCGATTTTT

LacZ-9,

GARAAYCCNGGNGTNACNCARCTNAAYCGNCTNGCNGCNCAYCCNCCNGCAT GACTCGACCATCCGATTTTTT

LacZ-10,

GNGTNACNCCNGGRTTYTCCCARTCNCGNCGYTGNAGNACNACNGCNAGCATG ACTCGACCATCCGATTTTTT

LacZ-11,

ACAGGAAACAGCTATGACNATGATHACNCTNGCNGTNGTNCTNCARCGGCATG ACTCGACCATCCGATTTTTT

LacZ-12,

TCATNGTCATAGCTGTTTCCTGTGTGAAATTAATTAGGTTAGTACCGGGCATGAC TCGACCATCCGATTTTTT

>RFP

RFP-f1-1

TTCAAATGGGAACGTGTTATGAACTTCGAAGACGGTGGTGTTGTTACCGTTACCC AGGACGCATGACTCGACCATCC

GATTTTTT

RFP-f1-2

TTCATAACACGTTCCCATTTGAAACCTTCCGGGAAGGACAGTTTCAGGTAGTCCGGGATGGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f1-3

CCAGTACGGTTCCAAAGCTTACGTTAAACACCCGGCTGACATCCCGGACTACCT GAAACTGCATGACTCGACCATCC

GATTTTTT

RFP-f1-4

CGTAAGCTTTGGAACCGTACTGGAACTGCGGGGACAGGATGTCCCAAGCGAAC GGCAGCGGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f1-5

ACGAAGGTACCCAGACCGCTAAACTGAAAGTTACCAAAGGTGGTCCGCTGCCG TTCGCTTGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f1-6

GCGGTCTGGGTACCTTCGTACGGACGACCTTCACCTTCACCTTCGATTTCGAACT CGTGAGCATGACTCGACCATCC

GATTTTTT

RFP-f1-7

CATGCGTTTCAAAGTTCGTATGGAAGGTTCCGTTAACGGTCACGAGTTCGAAATC GAAGGGCATGACTCGACCATCC

GATTTTTT

RFP-f1-8

CATACGAACTTTGAAACGCATGAACTCTTTGATAACGTCTTCGGAGGAAGCCAT CATAGCGCATGACTCGACCATCC

GATTTTTT

RFP-f1-9

AGCCTGGATGACGTTTTCATCAAAATTTCACACAGGAAACAGCTATGATGGCTTCCTCCGGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-1

CGGGAAGTTGGTACCACGCAGTTTAACTTTGTAGATGAACTCACCGTCTTGCAGG GAGGAGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-2

CGTGGTACCAACTTCCCGTCCGACGGTCCGGTTATGCAGAAAAAAACCATGGGT TGGGAAGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-3

TCAGAGCACCGTCTTCCGGGTACATACGTTCGGTGGAAGCTTCCCAACCCATGGT TTTTTGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-4

CGGAAGACGGTGCTCTGAAAGGTGAAATCAAAATGCGTCTGAAACTGAAAGAC GGTGGTCGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-5

TTTAGCCATGTAGGTGGTTTTAACTTCAGCGTCGTAGTGACCACCGTCTTTCAGTT TCAGGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-6

GAAGTTAAAACCACCTACATGGCTAAAAAAACCGGTTCAGCTGCCGGGTGCTTAC AAAACCGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-7

GTGTAGTCTTCGTTGTGGGAGGTGATGTCCAGTTTGATGTCGGTTTTGTAAGCACC CGGCGCATGACTCGACCATCC

**GATTTTTT** 

RFP-f2-8

CCTCCCACAACGAAGACTACACCATCGTTGAACAGTACGAACGTGCTGAAGGTCGTCACTGCATGACTCGACCATCC

GATTTTTT

RFP-f2-9

AAATTGAGACTGGAAACGCACGGTTTCTTAAGCACCGGTGGAGTGACGACCTTC AGCACGGCATGACTCGACCATCC

**GATTTTTT** 

# References

- 1. Smith, H.O. and K.W. Wilcox, *A restriction enzyme from Hemophilus influenzae*. *I. Purification and general properties*. J Mol Biol, 1970. **51**(2): p. 379-91.
- 2. Danna, K. and D. Nathans, Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. Proc Natl Acad Sci U S A, 1971. **68**(12): p. 2913-7.
- 3. Cohen, S.N., et al., Construction of biologically functional bacterial plasmids in vitro. Proc Natl Acad Sci U S A, 1973. **70**(11): p. 3240-4.
- 4. Liu, Q., et al., *The univector plasmid-fusion system, a method for rapid construction of recombinant DNA without restriction enzymes.* Curr Biol, 1998. **8**(24): p. 1300-9.
- 5. Hartley, J.L., G.F. Temple, and M.A. Brasch, *DNA cloning using in vitro site-specific recombination*. Genome Res, 2000. **10**(11): p. 1788-95.
- 6. Walhout, A.J., et al., *GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes.* Methods Enzymol, 2000. **328**: p. 575-92.
- 7. Shuldiner, A.R., L.A. Scott, and J. Roth, *PCR-induced* (ligase-free) subcloning: a rapid reliable method to subclone polymerase chain reaction (*PCR*) products. Nucleic Acids Res, 1990. **18**(7): p. 1920.
- 8. Aslanidis, C. and P.J. de Jong, *Ligation-independent cloning of PCR products (LIC-PCR)*. Nucleic Acids Res, 1990. **18**(20): p. 6069-74.
- 9. Rashtchian, A., *Novel methods for cloning and engineering genes using the polymerase chain reaction.* Curr Opin Biotechnol, 1995. **6**(1): p. 30-6.
- 10. Bitinaite, J., et al., *USER friendly DNA engineering and cloning method by uracil excision*. Nucleic Acids Res, 2007. **35**(6): p. 1992-2002.
- 11. Li, M.Z. and S.J. Elledge, *MAGIC*, an in vivo genetic method for the rapid construction of recombinant DNA molecules. Nat Genet, 2005. **37**(3): p. 311-9.
- 12. Li, M.Z. and S.J. Elledge, *Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC.* Nat Methods, 2007. **4**(3): p. 251-6.
- 13. Marsischky, G. and J. LaBaer, *Many paths to many clones: a comparative look at high-throughput cloning methods.* Genome Res, 2004. **14**(10B): p. 2020-8.
- 14. Klock, H.E., et al., Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. Proteins, 2008. **71**(2): p. 982-94.
- 15. Nisson, P.E., A. Rashtchian, and P.C. Watkins, *Rapid and efficient cloning of Alu-PCR products using uracil DNA glycosylase*. PCR Methods Appl, 1991. **1**(2): p. 120-3.
- 16. Sharp, P.M. and W.H. Li, Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res, 1986. **14**(19): p. 7737-49.

- 17. Xia, X., How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics, 1998. **149**(1): p. 37-44.
- 18. Marais, G. and L. Duret, *Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans.* J Mol Evol, 2001. **52**(3): p. 275-80.
- 19. Hershberg, R. and D.A. Petrov, *Selection on codon bias*. Annu Rev Genet, 2008. **42**: p. 287-99.
- 20. Sharp, P.M. and W.H. Li, *The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications*. Nucleic Acids Res, 1987. **15**(3): p. 1281-95.
- 21. Wang, S., et al., Hemagglutinin (HA) proteins from H1 and H3 serotypes of influenza A viruses require different antigen designs for the induction of optimal protective antibody responses as studied by codon-optimized HA DNA vaccines. J Virol, 2006. **80**(23): p. 11628-37.
- 22. Foster, H., et al., Codon and mRNA sequence optimization of microdystrophin transgenes improves expression and physiological outcome in dystrophic mdx mice following AAV2/8 gene transfer. Mol Ther, 2008. **16**(11): p. 1825-32.
- 23. Burgess-Brown, N.A., et al., Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. Protein Expr Purif, 2008. **59**(1): p. 94-102.
- 24. Zhi, N., et al., Codon optimization of human parvovirus B19 capsid genes greatly increases their expression in non-permissive cells. J Virol, 2010.
- 25. Hale, R.S. and G. Thompson, Codon optimization of the gene encoding a domain from human type 1 neurofibromin protein results in a threefold improvement in expression level in Escherichia coli. Protein Expr Purif, 1998. **12**(2): p. 185-8.
- 26. Zhou, Z., et al., Enhanced expression of a recombinant malaria candidate vaccine in Escherichia coli by codon optimization. Protein Expr Purif, 2004. **34**(1): p. 87-94.
- 27. Ertl, P.F. and L.L. Thomsen, *Technical issues in construction of nucleic acid vaccines*. Methods, 2003. **31**(3): p. 199-206.
- 28. Ill, C.R. and H.C. Chiou, *Gene therapy progress and prospects: recent progress in transgene and RNAi expression cassettes.* Gene Ther, 2005. **12**(10): p. 795-802.
- 29. Stacey, K.J., et al., *The molecular basis for the lack of immunostimulatory activity of vertebrate DNA*. J Immunol, 2003. **170**(7): p. 3614-20.
- 30. Krieg, A.M., et al., *CpG motifs in bacterial DNA trigger direct B-cell activation*. Nature, 1995. **374**(6522): p. 546-9.
- 31. Trinh, R., et al., *Optimization of codon pair use within the (GGGGS)3 linker sequence results in enhanced protein expression.* Mol Immunol, 2004. **40**(10): p. 717-22.
- 32. Boycheva, S., G. Chkodrov, and I. Ivanov, *Codon pairs in the genome of Escherichia coli*. Bioinformatics, 2003. **19**(8): p. 987-98.
- 33. Chen, D., et al., *Expression of enterovirus 70 capsid protein VP1 in Escherichia coli*. Protein Expr Purif, 2004. **37**(2): p. 426-33.

- 34. Muthumani, K., et al., *Issues for improving multiplasmid DNA vaccines for HIV-1*. Vaccine, 2002. **20**(15): p. 1999-2003.
- 35. Donnelly, J.J., B. Wahren, and M.A. Liu, *DNA vaccines: progress and challenges*. J Immunol, 2005. **175**(2): p. 633-9.
- 36. Koide, Y., et al., *DNA vaccines*. Jpn J Pharmacol, 2000. **83**(3): p. 167-74.
- 37. Doria-Rose, N.A. and N.L. Haigwood, *DNA vaccine strategies: candidates for immune modulation and immunization regimens*. Methods, 2003. **31**(3): p. 207-16.
- 38. Andersson, H.A., R.A. Singh, and M.A. Barry, *Activation of refractory T cell responses against hepatitis C virus core protein by ablation of interfering hydrophobic domains*. Mol Ther, 2006. **13**(2): p. 338-46.
- 39. Wuitschick, J.D. and K.M. Karrer, *Analysis of genomic G + C content, codon usage, initiator codon context and translation termination sites in Tetrahymena thermophila.* J Eukaryot Microbiol, 1999. **46**(3): p. 239-47.
- 40. Edwards, N.C., et al., Characterization of Coding Synonymous and Non-Synonymous Variants in ADAMTS13 Using Ex Vivo and In Silico Approaches. PLoS One, 2012. 7(6): p. e38864.
- 41. Sanchez, J., *3-base periodicity in coding DNA is affected by intercodon dinucleotides*. Bioinformation, 2011. **6**(9): p. 327-9.
- 42. Plotkin, J.B. and G. Kudla, *Synonymous but not the same: the causes and consequences of codon bias.* Nat Rev Genet, 2011. **12**(1): p. 32-42.
- 43. McArthur, G.H.t. and S.S. Fong, *Toward engineering synthetic microbial metabolism*. J Biomed Biotechnol, 2010. **2010**: p. 459760.
- 44. Quan, J. and J. Tian, *Circular polymerase extension cloning of complex gene libraries and pathways.* PLoS One, 2009. **4**(7): p. e6441.
- 45. Horton, R.M., et al., Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. Gene, 1989. 77(1): p. 61-8.
- 46. Prodromou, C. and L.H. Pearl, *Recursive PCR: a novel technique for total gene synthesis*. Protein Eng, 1992. **5**(8): p. 827-9.
- 47. Stemmer, W.P., et al., *Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides*. Gene, 1995. **164**(1): p. 49-53.
- 48. Shevchuk, N.A., et al., Construction of long DNA molecules using long PCR-based fusion of several fragments simultaneously. Nucleic Acids Res, 2004. **32**(2): p. e19.
- 49. Smith, H.O., et al., Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15440-5.
- 50. Tian, J., et al., Accurate multiplex gene synthesis from programmable DNA microchips. Nature, 2004. **432**(7020): p. 1050-4.
- 51. Garces, C. and J. Laborda, *Single-step, ligase-free cloning of polymerase chain reaction products into any restriction site of any DNA plasmid.* Anal Biochem, 1995. **230**(1): p. 178-80.

- 52. Sambrook, J. and D.W. Russell, *Molecular cloning : a laboratory manual*. 3rd ed2001, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press. 3 v.
- 53. Quan, J. and J. Tian, *Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries*. Nat Protoc, 2011. **6**(2): p. 242-51.
- 54. Gibson, D.G., et al., One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. Proc Natl Acad Sci U S A, 2008. **105**(51): p. 20404-9.
- 55. Shao, Z. and H. Zhao, *DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways.* Nucleic Acids Res, 2009. **37**(2): p. e16.
- 56. Saaem, I., et al., *In situ synthesis of DNA microarray on functionalized cyclic olefin copolymer substrate.* ACS Appl Mater Interfaces, 2010. **2**(2): p. 491-7.
- 57. Quan, J., et al., *Parallel on-chip gene synthesis and application to optimization of protein expression*. Nat Biotechnol, 2011. **29**(5): p. 449-52.
- 58. Yeung, A.T., et al., *Enzymatic mutation detection technologies*. Biotechniques, 2005. **38**(5): p. 749-58.
- 59. Qiu, P., et al., *Mutation detection using Surveyor nuclease*. Biotechniques, 2004. **36**(4): p. 702-7.
- 60. Carr, P.A., et al., *Protein-mediated error correction for de novo DNA synthesis*. Nucleic Acids Res, 2004. **32**(20): p. e162.
- 61. Fuhrmann, M., et al., *Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage*. Nucleic Acids Res, 2005. **33**(6): p. e58.
- 62. Kosuri, S., et al., *Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips*. Nat Biotechnol, 2010. **28**(12): p. 1295-9.
- 63. Tian, J., K. Ma, and I. Saaem, *Advancing high-throughput gene synthesis technology*. Mol Biosyst, 2009. **5**(7): p. 714-22.
- 64. Welch, M., et al., *You're one in a googol: optimizing genes for protein expression*. J R Soc Interface, 2009. **6 Suppl 4**: p. S467-76.
- 65. Carr, P.A. and G.M. Church, *Genome engineering*. Nat Biotechnol, 2009. **27**(12): p. 1151-62.
- 66. Patwardhan, R.P., et al., *High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis.* Nat Biotechnol, 2009. **27**(12): p. 1173-5.
- 67. Pfleger, B.F., et al., *Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes*. Nat Biotechnol, 2006. **24**(8): p. 1027-32.
- 68. Salis, H.M., E.A. Mirsky, and C.A. Voigt, *Automated design of synthetic ribosome binding sites to control protein expression*. Nat Biotechnol, 2009. **27**(10): p. 946-50.
- 69. Isaacs, F.J., et al., *Engineered riboregulators enable post-transcriptional control of gene expression*. Nat Biotechnol, 2004. **22**(7): p. 841-7.
- 70. Kudla, G., et al., *Coding-sequence determinants of gene expression in Escherichia coli*. Science, 2009. **324**(5924): p. 255-8.

- 71. Tuller, T., et al., *Translation efficiency is determined by both codon bias and folding energy*. Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3645-50.
- 72. Coleman, J.R., et al., *Virus attenuation by genome-scale changes in codon pair bias.* Science, 2008. **320**(5884): p. 1784-7.
- 73. Wu, Z., et al., *Optimization of self-complementary AAV vectors for liver-directed expression results in sustained correction of hemophilia B at low vector dose.* Mol Ther, 2008. **16**(2): p. 280-9.
- 74. Wu, G., L. Nie, and S.J. Freeland, *The effects of differential gene expression on coding sequence features: analysis by one-way ANOVA*. Biochem Biophys Res Commun, 2007. **358**(4): p. 1108-13.
- 75. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science, 2004. **306**(5696): p. 636-40.
- 76. Zhou, X., et al., *Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences*. Nucleic Acids Res, 2004. **32**(18): p. 5409-17.
- 77. Richmond, K.E., et al., Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. Nucleic Acids Res, 2004. **32**(17): p. 5011-8.
- 78. Borovkov, A.Y., et al., *High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides*. Nucleic Acids Res, 2010. **38**(19): p. e180.
- 79. Lamprecht, M.R., D.M. Sabatini, and A.E. Carpenter, *CellProfiler: free, versatile software for automated biological image analysis.* Biotechniques, 2007. **42**(1): p. 71-5.
- 80. Ma, K.-S., et al., Versatile surface functionalization of cyclic olefin copolymer (COC) with sputtered SiO2 thin film for potential BioMEMS applications. Journal of Materials Chemistry, 2009. **19**(42): p. 7914-7920.
- 81. Huang, M.C., et al., *Integrated two-step gene synthesis in a microfluidic device*. Lab Chip, 2009. **9**(2): p. 276-85.
- 82. Consortium, T.E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-640.
- 83. Nakamura, Y., T. Gojobori, and T. Ikemura, *Codon usage tabulated from international DNA sequence databases: status for the year 2000.* Nucleic Acids Res, 2000. **28**(1): p. 292.
- 84. Puigbo, P., I.G. Bravo, and S. Garcia-Vallve, *CAIcal: a combined set of tools to assess codon usage adaptation*. Biol Direct, 2008. **3**: p. 38.
- 85. Xu, L., et al., Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol Biol Evol, 2006. **23**(6): p. 1107-8.
- 86. Wu, G., et al., Simplified gene synthesis: a one-step approach to PCR-based gene construction. J Biotechnol, 2006. **124**(3): p. 496-503.

# **Biography**

Jiayuan Quan was born on November 29th, 1984 in Dalian, China. She attended Peking University in Beijing, China from 2003 to 2007 where she obtained her Bachelor of Science degree in Biotechnology in June 2007. After graduating from Peking University, she was accepted into the PhD program in Biomedical Engineering at Duke University. She performed her thesis research in the laboratory of Jingdong Tian in the field of synthetic biology until July 2012 when she completed her dissertation. During the PhD period, her publications, awards and fellowships are listed below:

#### **Publications:**

- Quan, J., et al. Parallel on-chip gene synthesis and application to optimization of protein expression. Nature Biotechnology 29, 449-452 (2011)
- Quan, J. & Tian, J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. Nature Protocols 6, 242-251 (2011)
- Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. PLoS ONE 4, e6441 (2009). Recommended on Faculty of 1000 Post-Publication Peer Review website, 07 Sep 2009
- Saaem, I, Ma, S, Quan, J, & Tian, J. Error correction of microchip synthesized genes using Surveyor nuclease. Nucleic Acid Research, 2012. 40(3): p. e23
- Quan, J. & Tian, J. Circular polymerase extension cloning (book chapter),
   Springer protocols, Methods in Molecular Biology, Humana Press (2012)

### Awards and fellowships:

- Duke University Graduate Fellowship of Biomedical Engineering
- Duke University Kewaunee Poster Session Junior Graduate Award (2011)
- BioBricks Foundation SB5.0 Young Researcher Travel Award (2011)

- Biomedical Engineering Society Travel Award (2009 and 2010)
- International Genetically Engineered Machine (iGEM) competition Best Experimental Measurement (2010)
- International Genetically Engineered Machine (iGEM) competition Silver Award (2009)
- Sax/Baldridge First Place Award for best small business consulting project (2012)