The Dissertation Committee for Jared Wade Ellefson Certifies that this is the approved version of the following dissertation:

Engineering the central dogma using emulsion based directed evolution

Committee:

Andrew Ellington, Supervisor

Jeffrey Barrick

Edward Marcotte

George Georgiou

Rick Russell

Engineering the central dogma using emulsion based directed evolution

by

Jared Wade Ellefson, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2016

Dedication

To my mom and dad - for putting me through college and always supporting me

Acknowledgements

The accomplishment of this PhD has been the trial and tribulations of seven years in the Ellington lab. This unique environment has shaped and molded the scientist I have become, whether that be good or bad. This is especially due to the people who have travelled through the lab and in the program as a whole. At some level, special thanks must be given to Andy. Not for the great mentorship one might expect from an advisor, but for creating the environment that brought this island of misfit toys together. Was it incredible foresight to foster this environment of engagement and frustration, or was it the in vivo selection process that semi-randomly brought us all together?

At the start of my time in the Ellington lab, Eric Davidson played an important role in breaking my dependence on mentorship. His stonewall approach to dealing with Andy and sparse encouragement was pivotal to making me independent early on. Numerous other people have been a source of inspiration and amusement throughout my time here, namely: Tony, Randy, Andre, Johnny, Dan, Aziz, Xi, Steven, Drew, and Raghav. Without the support of my close friends Jon and Mike, this would have been much more difficult. Michael Ledbetter for helping with the fly husbandry and always chanting "pickles... pickles."

A very special thanks goes out to Adam Meyer. His friendship and collaboration will always be remembered. Without his support I wouldn't have had the successes I've had. Ross Thyer for putting up with us making fun of him and for his high-copy-highly-

constitutive-promoter philosophy. And to Jimmy Gollihar for challenging and enabling me.

To my family. Mom, for the constant support you gave me. And Kevin - I hope you realize this would never have happened without you. All my brothers, thanks for being there. Colin, why am I always so high up on the boob-list? And stop giving me ponies on Christmas! And Dacia, for the unwavering support even when I'm full of self doubt.

Engineering the central dogma using emulsion based directed evolution

Jared Wade Ellefson, Ph.D.

The University of Texas at Austin, 2016

Supervisor: Andrew D Ellington

The central dogma of molecular biology forms the most basic (and fundamental) paradigm of how life operates. Despite its elegant simplicity, scientists are still uncovering enigmas of the central dogma - which has been shaped throughout billions of years of the Darwinian process. Even though the core concepts of the central dogma have largely been untouched by evolution (the universality of the genetic code, amino acid utilization, DNA/RNA base identity) scientific advances have shown that these fundamental properties can be altered dramatically. This implies the architectures of life are pliable and likely the result of extreme optimization and fine tuning of semi-random events that took place soon after the origin of life.

Reengineering the parameters of life offers a unique way of testing evolutionary processes and perceived optimality of its components. Naturally, coaxing proteins and nucleic acids to function in an unnatural fashion is difficult. Development of techniques to enable these changes has relied heavily on the exploitation of water-in-oil emulsions (or, *in vitro* compartmentalization), which allows directed evolution at the single cell or even single molecule level. In particular, compartmentalized partnered replication (CPR) is a dual mode selection technique, coupling the *in vivo* functionality of a gene with the *in vitro* amplification via emulsion PCR. The CPR technique has enabled the development of synthetic promoter recognition by T7 RNA polymerase, unnatural amino acid

incorporation by aminoacyl tRNA synthetase engineering, genetic code reassignment through tRNA evolution, and transcriptional regulation using repressors with novel allosteric effector molecules and operator binding sites. Using a similar technique, the template recognition of an Archaeal DNA polymerase was altered such that the polymerase utilizes both DNA and RNA templates with similar efficiencies. This resulted in a reverse transcriptase that can functionally proofread on RNA templates.

These technologies will continue to play a pivotal role in the future development of particular aspects of the central dogma. As certain steps in this process are tweaked to have alternative functionalities and combined together, the gap between natural life and synthetically modified life widens and gives the Darwinian process of evolution new areas to explore.

Table of Contents

List of Tables

List of Figures

INTRODUCTION

The molecular basis of life on Earth relies on the information flow between DNA, RNA, and proteins – which was so elegantly postulated by Francis Crick in the central dogma of molecular biology. This framework accounts for the general under-workings behind organismal phenotype and genetic heredity. It describes the residue by residue transfer of sequential information; how one type of molecule becomes another and how sequence information is passed down through the generations of evolutionary time.

Since this framework was put forth, researchers have begun to probe the limits of the central dogma. The molecular machines driving the transfer of sequence information can sometimes utilize non-natural substrates. Additional molecular engineering of these proteins and nucleic acids can further the reach of the substrate analogs utilized. As time has gone on, more sophisticated *in vivo* and *in vitro* screening and selection methodologies have begun to create the tools necessary to create alternative central dogmas using synthetic substrates.

THE CENTRAL DOGMA IN NATURE

Nucleic acid replication

Watson and Crick's model for the structure of DNA was profound because it provided a mechanism for heritability (Watson and Crick, 1953). One half of the double stranded duplex provides the information to synthesize the other half. This mechanism

was elegantly demonstrated by Meselson and Stahl, who proved that nucleic acid replication is semiconservative (Meselson and Stahl, 1958). Since then, the biomolecules behind the replication have been elucidated, and while the replication machinery is complex the core component is the protein polymerase (Leman and Noguchi, 2013).

In vitro reconstitutions of the replication machinery show that polymerization can occur simply given a polymerase, a hydroxyl group to prime the initiation of the complement strand, and substrate deoxynucleotide triphosphates (Lehman et al., 1958). Since this discovery, the most pervasive example of nucleic acid replication systems is the polymerase chain reaction (PCR) (Mullis and Faloona, 1987). In PCR, priming oligonucleotides that bind to each strand of a target region (i.e. “facing each other”) undergo sequential rounds of primer extension that, in effect, will create another template for the opposing primer oligonucleotide - ultimately resulting in exponential amplification of the specified DNA sequence. This technique has revolutionized the whole of biology, giving the ability to not only amplify small amounts of DNA but allow it to be put together in any imaginable way.

Since its inception, the PCR technique has undergone massive improvements. Perhaps the most important technical improvement came when thermostable DNA polymerases were introduced (Saiki et al., 1988). This allowed polymerase enzymes to remain active after heat denaturation of the template DNA. Further advances were made when error-correcting DNA polymerases were used in the PCR reaction - increasing replication fidelity. Archaeal Family-B polymerases (the organism’s replicative polymerase) contain an editing domain (3’-5’ exonuclease) that is stimulated upon base

misincorporation (Fidalgo da Silva and Reha-Krantz, 2007). Adding components of the replication holoenzyme, namely the PCNA clamp, has also been shown to improve PCR by increasing the processivity of the polymerase complex (Kitabayashi et al., 2002).

In addition to DNA to DNA replication, RNA has been shown to complete the same replication cycle. These are generally considered a special case for the central dogma, and are normally only found in small viral elements. The most well studied example of this is the Q-beta replicase (Haruna et al., 1963). This bacteriophage polymerase is responsible for the replication of the viral RNA genome, by iterative copying of the plus and minus strands. This RNA replicase has been used for biotechnology applications, but template sequence requirements limit its utility (Cahill et al., 1991). Interestingly, the Q-beta replicase was used for some of the first *in vitro* selection experiments by Sol Speigelman which demonstrated that RNA molecules could evolve to gain new functions (Mills et al., 1967).

In addition to protein-based nucleic acid replication, there have been significant advances towards the understanding of nucleic acid polymerization in a prebiotic world. Because the components of the prebiotic world are evolutionarily extinct, the best way to understand it is to reconstruct its key elements. For instance, ribozymes can be selected *in vitro* from large randomized pools of RNA and have been shown to perform a wide variety of chemistries (Walter and Engelke, 2002). Most impressively, the Bartel lab created the first ribozyme-based polymerase (Johnston et al., 2001). This polymerase is capable of template directed RNA polymerization, which provides the means of a prebiotic world to replicate its genetic material and begin Darwinian evolution. Since its

inception, the ribozyme polymerase has undergone numerous improvements by utilizing emulsion-based directed evolution techniques (Wochner et al., 2011; Zaher and Unrau, 2007). The first versions of the polymerase could only extend a primer oligonucletide several bases, but after extensive evolution and engineering it has been demonstrated to synthesize over 200 bases (the length of the ribozyme) (Attwater et al., 2013). The demonstration of a ribozyme polymerase, and eventually of a simple self replicating entity, provides the simplest possible mechanism of the central dogma.

Our ability to control nucleic acid polymerization in a controlled fashion has revolutionized biology. The tools invented from these fundamental biological inquiries have enabled *de novo* DNA synthesis, cloning, diagnostic, therapeutic, sequencing, and evolutionary insights. Into the future, polymerases will play a pivotal role especially in applications for next-generation DNA and RNA sequencing technologies.

Transcription

Transcription plays a pivotal role in biology, by acting as the messenger between DNA and the translational apparatus. Cellular transcription is unique from the nucleic acid replicators mentioned above, in that they do not require a complementary primer to initiate polymerization. The *de novo* synthesis of RNA transcripts is carried out upon the transcription machinery's recognition of a promoter sequence (Chamberlin et al., 1970). The process of promoter recognition is vastly different between replication complexes, and between prokaryotes and eukaryotes. Promoter initiation can be influenced by a wide range of factors, including: ancillary proteins that bind upstream or downstream in the DNA, the availability of DNA binding sites by competitive binding of other proteins or

sequestration into chromatin structures, and the cascade of gene expression leading up to transcription which can be influenced by small molecules or signaling pathways (Madan Babu, 2003).

While the transcription machinery is integral for the production of messenger RNAs, the role of RNA transcripts extends far beyond this in the cell. Non-coding RNAs have a variety of functional roles in cells. Perhaps the most famous, and integral to the central dogma, is the ribosome and tRNAs that decode mRNA messages (Ramakrishnan, 2002). Certain RNAs can serve to modulate the functional readout, perhaps most famously in self-splicing introns for tRNA maturation or in the splicosomal ribonucleo-protein complex which removes introns from coding RNAs (Cech, 1990; Will and Luhrmann, 2011). Other non-coding RNAs have gained much attention recently, the discovery of the CRISPR/CAS system has revolutionized biotechnology by allowing specific cleavage of DNA sequences given a complementary guide RNA sequence (Jinek et al., 2012).

The precise control of gene expression is heavily reliant on the regulation of transcription. In bacteria, this is controlled by transcription factors which bind to DNA sequences recruiting or competing for transcription machinery binding. In biosynthetic operons, expression is regulated heavily by positive or negative feedback. DNA binding proteins that are allosterically regulated by the products of the pathway will alter conformational states to reflect the internal parameters of the cell (Rogers et al., 2015).

Reverse Transcription

The unidirectional view of the central dogma was turned on its head in 1970 when Temin and Baltimore discovered the first reverse transcriptases in mammalian RNA viruses (Baltimore, 1970; Temin and Mizutani, 1970). Since the RT's discovery, it has been found across the domains of life (Xiong and Eickbush, 1990). The RT was initially thought to have arisen early in life's history serving to convert RNA to more stable DNA genomes (Darnell and Doolittle, 1986). Since then, it has found a number of different roles such as telomere maintenance, retrotransposon copying and insertion, and viral replication (Boeke and Stoye, 1997). The conversion of an RNA molecule into DNA has enabled significant advances in biology, most importantly enabling the conversion of mature RNAs into cDNA. However, the RTs use in biotechnology has been limited by their poor stability and lack of a proofreading domain.

Translation

The decoding of mRNAs into proteins is of paramount importance to cells, as estimates put it at a quarter of the cells total mass. The sequential transfer of information is determined by two factors: the step-by-step decoding of an mRNA on the ribosome by codon-anticodon interactions, and less appreciated, the specificity of the aminoacyl tRNA synthetase for its cognate tRNA. The decoding architecture is represented by the genetic code, which reads mRNAs in triplets and in some rare cases quadruplets. Since there are four possible RNA bases, a triplet code represents the near universal 64-term codon table. This universality is not trivial, and points towards an evolutionarily ancient common ancestor that preserved this code even billions of years ago (Woese, 1998). Due to this

surprising fact, DNA sequences from divergent organisms (such as *H. sapiens* and *E. coli*) can be translated in both hosts to create functional proteins. Although the genetic code is shared even among distant organism, there are many examples of alternative codes, even within human mitochondria.

As mRNAs are decoded on the ribosome, they move through three base intervals creating peptide bonds between aminoacylated tRNAs in the A and P sites of the ribosome (Steitz, 2008). tRNAs with the corresponding anti-codon sequence are shuttled to the ribosome by EF-Tu, and the sequential addition of amino acids is completed by a number of high energy reactions as the ribosome translocates across the mRNA. Upon reaching a stop signal (such as an amber codon), a release factor will bind to the stop codon and signal the ribosome complex to dissociate – terminating peptide synthesis.

The decoding of mRNAs is perhaps the most error prone process of the central dogma, making errors roughly 1 in every 1,000 amino acids (Loftfield and Vanderjagt, 1972). One possibility for this high error rate could be the incorrect aminoacylation of tRNAs by the synthetases, but studies have demonstrated its contribution to be minor, as the enzymatic aminoacylation of tRNAs only makes errors roughly 1 in 10^6 catalytic events (Söll, 1990). However, this is largely dependent on the specificities of tRNAs to their cognate synthetases, as tRNA identity elements can cross-react. The rate at which certain codons contain errors is highly variable suggesting misincorporation is partially based on inappropriate binding of noncognate tRNAs which may be driven by tRNA competition of the A-site (Kramer and Farabaugh, 2007). Some recent evidence also

suggests that EF-Tu performs an amino acid identity check by binding both the amino acid and tRNA elements (Schrader et al., 2011).

ALTERATIONS TO THE CENTRAL DOGMA

Through engineering efforts the central dogma has been shown to be pliable, accepting unnatural substrates (Fig. I.1). Here, the altered specificity of some of the systems is discussed which shows that even highly conserved proteins and nucleic acids have the potential to utilize novel substrates. This begs the question, how far can we deviate from what nature left us?

Nucleic acid polymerases have been demonstrated to utilize both modified substrates and modified templates. Originally, investigations into polymerase mutants capable of utilizing modified substrates were performed to improve sequencing techniques by enhancing a polymerases' ability to incorporate chain terminators or dye labeled nucleotides (Vander Horn et al., 1997). Further investigations probed the substrate specificity of Archaeal Family-B polymerases by sequence alignments of conserved regions involved in substrate discrimination in the active site (Joyce and Steitz, 1995) . Small scale screens of mutations in these regions relaxed the substrate specificity improving incorporation of dideoxyribonucleotides and even ribonucleotides (Gardner and Jack, 1999). This engineered polymerase (termed Therminator) contained mutations Y409G and A488L which are implicated as a steric gate - blocking ribonucleotide incorporation and relaxing nucleotide selectivity, respectively. Additionally, the Therminator polymerase could incorporate threose nucleic acids (TNAs), which only have a 4 carbon sugar backbone (Ichida, 2005). In fact, it was later shown that in

combination with several additional mutations (E664K and V93Q) the Archaeal polB enzyme could effectively incorporate RNA to the point where mRNAs longer than one kilobase could be polymerized (Cozens et al., 2012).

While the Therminator polymerase was discovered based on small scale screens of conserved motifs of Archaeal polB, more drastic changes needed to be made to improve the incorporation of modified bases and allow the incorporation of even more divergent base analogs. The interactions made between polymerases and the substrates, templates, and nascent strand are extensive, making rational mutagenesis nearly impossible, and library design challenging. A number of research groups have developed selection methodologies that enable the high throughput scanning of mutations for a desired functionality. Various *in vivo* and *in vitro* approaches have been developed to tackle this problem (Holmberg et al., 2005). The two most successful *in vitro* strategies have relied upon phage display technologies and emulsion based compartmentalized self replication (CSR; discussed below). In phage display selections, polymerases are physically encoded into the gene sequence of M13 gene p3 (Jestin et al., 1999). Primer-template complexes are tethered to the phage by covalently cross-linking, and nucleotide addition can be selected by incorporation of a biotinylated nucleotide followed by affinity pull-down of the phage (Xia et al., 2002). Using this basic methodology DNA polymerases have been engineered to incorporate ribonucleotides, including modified ribonucleotides such as 2'-O-methyl RNA (Fa et al., 2004).

Building on top of this work, researchers have become interested in more and more chemically divergent substrates. For instance, Kool demonstrated that hydrophobic

base analogs, geometrically similar to purines and pyrimidines, could be used as substrates by natural polymerases (Schweitzer and Kool, 1995). These findings point towards geometry of base pairs being an important factor in polymerization, and not specific base hydrogen bonding . The Hirao and Romesberg laboratories have also created nucleic acid analogs based on hydrophobic base composition, but unlike the Kool bases, have diverged structurally (Berger et al., 2000; Hirao et al., 2006). The Benner laboratory has also constructed artificial bases, but uniquely these can have altered hydrogen bonding between bases (Yang et al., 2006). The rearranged hydrogen bonding pairs between nucleobases should give more specificity than the simple geometric packing present in the other unnatural base pairs (UBPs). Most experiments with the UBPs have tested polymerization with single UBP addition. Most natural polymerases utilized for their incorporation struggle with the successive polymerization of these bases, presumably because the altered conformation of the resulting duplex causes polymerase stalling (Lutz et al., 1999). Even under optimized conditions, the fidelity of these reactions has been quite limited, making errors roughly 1 in 100 bases incorporated (Malyshev et al., 2012; Yang et al., 2011).

Recently it has been shown that UBPs can be both transported into *E. coli* cells, as well as, incorporated during episome replication by the endogenous polymerase machinery (Malyshev et al., 2014). This technology is in its infancy and even with heavily optimized conditions the UBP sequence is lost within several generations. But, this provides an optimistic outlook on the use of UBPs into the future. Most intriguingly if UBP systems could be stably used *in vivo*, being utilized for both DNA replication and

transcription, then synthetic mRNA and tRNAs could be made with a third base pair. This would expand the number of usable codons to 216, as compared with the 64 in the standard genetic code (Thyer and Ellefson, 2014). This would greatly enable efforts to expand the genetic code with nonstandard amino acids by creating new codons for additional amino acids instead of competing with the amber codon for incorporation. The widespread use of UBPs across the cell may be challenging. The Romesberg bases are hydrophobic in nature, and their replication has caveats *in vitro* (Lavergne et al., 2013). It will remain to be seen if large stretches of hydrophobic bases will be able to be polymerized, as there is no additional specificity enforced by hydrogen bond pairs. For widespread use across a genetically augmented cell, the rearranged hydrogen bonding of Benner bases may provide the specificity needed. However, most modifications to the nucleotides are not as forgiving and have required engineering and *in vitro* selection methods to improve their function with unnatural nucleotides (Chen and Romesberg, 2014).

Transcriptional machinery can also be engineered to utilize substrates outside of the central dogma. This work has focused mainly on the incorporation of modified nucleotide triphosphates instead of modified template recognition. This field is vastly underexplored due to the lack of a directed evolution methodology that can directly select for unnatural nucleotide incorporation. Selection methodologies have instead relied upon selections for wild-type function followed by screening of the mutant libraries for desired function (Holmberg et al., 2005). Most directed evolution efforts have relied heavily upon use of the RNA polymerase from T7 bacteriophage (T7 RNAP). This polymerase is

used widely because it is a monomeric enzyme (not requiring accessory factors) and it recognizes a short well defined promoter sequence (Milligan et al., 1987). The crystal structure of T7 RNA polymerase has defined the active site for nucleic acid polymerization (Cheetham et al., 1999), and mutations were subsequently identified that relax substrate specificity (Brieba and Sousa, 2000). These mutations, namely Y639F and H784A, render the polymerase capable of a minor amount of polymerization with fluoro-, amine-, deoxy-, and 2'O-methyl modified NTPs. A directed evolution approach identified further positions and mutations to allow increased incorporation efficiencies (Chelliserrykattil and Ellington, 2004). This approach relied on an autogene self-replication system to identify active polymerases, which were then screened for activity with modified substrates. The functional selection *in vivo* selection followed with *in vitro* screening is a valid approach for identifying mutants, but requires much work on the backend to screen the variants. In addition the libraries almost certainly must be targeted and well defined, preventing exploration of mutations which may be critical to unnatural nucleotide addition but are not predicted from the structural or phylogenetic data.

Modification of the genetic code has been a longstanding goal in synthetic biology. Particularly, the concept of adding new amino acids to the code is an intriguing one. Nonstandard amino acids (NAAs) are phenomenal tools for understanding and engineering proteins particularly by aiding the determination of x-ray crystallography (Sakamoto et al., 2009), the creation of orthogonal cross linking reagents (Deiters and Schultz, 2005; Zhang et al., 2002), the direct incorporation of post translational modifications (Neumann et al., 2009), and the incorporation of photo-reactive side groups

to active gene function upon light induction (Hino et al., 2005). Additionally, expanded genetic codes are capable of increasing the fitness of an organism by proving novel chemistry (Hammerling et al., 2014).

Altering the genetic code by incorporation of nonstandard amino acids was first demonstrated by using global replacement strategies. In a brilliant experiment by Wong, tryptophan auxotrophs of *B. subtilis* were grown on 4-fluorotryptophan, a tryptophan analog, in the absence of L-tryptophan (Wong, 1983). Because the tryptophan aminoacyl tRNA synthetase could utilize the 4-fluorotryptophan, global replacement of tryptophan for the NAA occurred. Initially this lowered the fitness of the organism, which is unsurprising considering the amino acid substitution would alter the entire proteome. After several serial passages a strain was identified that not only utilized the tryptophan analog, but preferentially grew on it. This was a significant finding - it demonstrated that the genetic code is not a stagnate entity but could be modified to contain amino acids not found naturally.

While the Wong strain was evolved to utilize a tryptophan analog, this organism still had a twenty amino acid code – with a rare amino acid swapped for a structurally very similar one. To truly expand the genetic code, a codon would have to be reassigned to allow the incorporation of a 21st amino acid. The amber codon (UAG) has been utilized for almost all genetic code expansion experiments because it is the rarest codon is most organisms. Since it is the rarest codon, it is the least likely to have considerably negative effects across the proteome of the host. Examples of stop codon suppressors can be found throughout nature, and are postulated to be important for genome evolution by

facilitating the creation of fusion proteins after gene duplication (Wong et al., 2008). In addition to having a free codon to utilize for incorporation of a NAA, an orthogonal tRNA:synthetase pair is needed. Orthogonal pairs are simply defined as a tRNA that will not react with any of the host's synthetases, and concurrently, a synthetase that will not aminoacylate endogenous host tRNAs. Orthogonal pairs are typically transplanted from distantly related organisms in hopes that the components will not cross-react when placed in the intended host organism. For instance, the first demonstration of this concept used an *E. coli* tyrosyl tRNA:sythetase pair to suppress amber codons in *S. cerevisiae* (Edwards and Schimmel, 1990). Once an orthogonal pair has been established, mutations to the amino acid binding pocket can be made without any fitness effects to the host organism. Pioneering efforts from Peter Schultz's lab have utilized orthogonal tRNA synthetases to site specifically incorporate dozens of NAAs (Liu and Schultz, 2010). Basic variations of this theme have been successful, allowing the incorporation of specific NAAs in bacteria (Wang et al., 2001), yeast (Chin et al., 2003), mammalian cells (Liu et al., 2007), and even nematodes (Greiss and Chin, 2011).

By far the trysoyl tRNA Synthetase from *M. jannaschii* has been the most utilized orthogonal pair for genetic code expansion. The synthetase lacks an anticodon recognition domain, which allowed the swapping of the tRNA anticodon to enable amber codon suppression – with minimal effects to the catalytic activity of amino acid charging (Steer and Schimmel, 1999). Some cross reactivity with endogenous machinery was observed as determined by amber suppression activity solely in the presence of the tRNA (Wang et al., 2000). This was overcome by randomization of residues in the tRNA using

an *in vivo* negative and positive selection scheme. In this scheme, cross-reactive tRNA variants (in the absence of the cognate synthetase) will cause amber suppression of a toxic gene - negatively selecting these variants. A subsequent positive selection is performed with the cognate synthetase which will allow cells to survive by the amber suppression of an antibiotic resistance marker. Using this system, tRNAs with increased orthogonality were selected (Wang et al., 2001). This same positive and negative selection approach can be used to engineer the synthetase to specifically incorporate NAAs. Randomization of the amino acid binding pocket could alter the substrate specificity to charge the cognate tRNA with a novel amino acid. By negatively selecting the library variants in the absence of the desired amino acid, synthetases that incorporate a standard amino acid will produce a toxic gene. Performing positive selection in the presence of the desired amino acid will allow the amber suppression of a resistance gene. By iterative cycling between positive and negative selection, synthetases which are specific for the desired amino acid are enriched in the population. Using this approach, NAAs with many functional groups have been incorporated into proteins using the *M. jannaschii* pair including: ketones (Wang et al., 2003), hydroxyl-amines (Xie and Schultz, 2006), thioester (Xie and Schultz, 2006), alkyne (Deiters and Schultz, 2005), and alkenes (Zhang et al., 2002).

Genetic code expansion in *E. coli* has largely relied on amber suppression of engineered tRNA:synthetase pairs. Inherently, this is a non-efficient process because amber suppression by tRNAs directly competes with the native release factor machinery. In *E. coli*, release factor 1 (RF1) is responsible for termination of translation in response

to the amber codon. Theoretically, deleting the RF1 gene all together would eliminate the competition with NAA amber suppression efficiency - however, deletion of RF1 is lethal in *E. coli* (Rydén and Isaksson, 1984). Recently researchers have developed viable RF1 knockout strains that improve amber suppression efficiency (Johnson et al., 2011; Mukai et al., 2010). These strains rely on the presence of amber suppressors to achieve viability, and required genomic editing to proteins with amber stop codons that suffer fitness defects with the C-terminal addition of amino acids. To bypass these issues all together, a monumental effort by Isaacs and Church has led to an *E. coli* strain that contains no amber stop codons (Isaacs et al., 2011; Lajoie et al., 2013). This 'amberless' organism greatly improves the NAA incorporation machinery and results in increased expression of proteins with novel amino acid sidechains.

In addition to modification of the tRNA:synthetase pair and recoding the organism, ribosomes themselves have been engineered. For example, ribosomes have been mutated in the peptidyl transferase center that allowed the incorporation of D-amino acids and therefore the synthesis of backbone modified proteins (Dedkova et al., 2003). Chin and colleagues have developed orthogonal ribosomes in *E. coli* by mutating the anti-shine dalgarno sequence on the 16S subunit, which led to enhancement of amber suppression efficiency (Wang et al., 2007). The orthogonal ribosome opens the opportunity for mutating the ribosome without catastrophic cellular consequences, as 'normal' messages would be read by the wild-type ribosome. However, modifications to the 23S would not be as accepted because orthogonality cannot be achieved in the same fashion. To overcome this, an incredible engineering feat was achieved by the fusion of

the 16S and 23S subunits of the bacterial ribosome (Orelle et al., 2015). This enables the possibility of mutation of both subunits without severe fitness effects *in vivo*.

While still in its infancy, the evolutionary and biotechnology consequences of expanded genetic codes are beginning to take shape. The evolutionary potential of expanded genetic codes was tested using bacteriophage T7 as a model organism (Hammerling et al., 2014). This was chosen because the host, *E. coli*, could be freshly infected every several generations of phage – preventing the mutation of the NAA machinery. Serial passages increased fitness of the population and in some phage lineages, amber codons became fixed in the population. For instance, a holin lysis protein contained an amber codon (and therefore the NAA, 3-iodotyrosine) which was demonstrated to increase the fitness of the phage in an amino acid specific manner. This simple experiment points towards NAAs being valuable biotechnology tools into the future. One such use would be to use NAAs as a biocontainment strategy, relying on NAAs for survival (Mandell et al., 2015; Tack et al., 2016; Thyer et al., 2015). This was possible by making the function of essential genes dependent on the NAA, either in the form of metabolic genes or resistance markers. These techniques have largely relied on exploitation of the folding dynamics of proteins, but it is likely that NAAs will be shown to be advantageous in the active site of enzymes as well.

EMULSIONS AS A TOOL FOR MOLECULAR EVOLUTION

A link between genotype and phenotype is a requirement for the evolution of living systems. This is seen throughout nature, as cells go through great lengths to keep out foreign genetic information. Spontaneous genotypic mutations that correspond to

positive fitness phenotypes need to reproduce and outcompete mutations that are less beneficial. For instance, a continuous aqueous phase of mixed phenotypes and genotypes will often result in evolutionary instability as tested by *in vitro* evolutionary systems (Breaker and Joyce, 1994; Ichihashi et al., 2013; Mills et al., 1967). Instability is a result of parasitic genotypes that arise by positive phenotypes incorrectly recognizing non-cognate genetic templates diminishing enrichment of positive genotypes. For all intents and purposes this linkage is presumably a fundamental feature of life itself (Deamer and Dworkin, 2005; Meyer et al., 2012; Szathmáry and Demeter, 1987).

In most living systems the genotype-phenotype linkage is achieved by containment of genetic information into a cell. A fatty acid membrane envelops the genome of a given organism preventing cross metabolism with other cells that may be around. At the origin of life, how was such a linkage maintained, especially without the complex machinery involved in membrane biosynthesis? Several alternative strategies have been postulated, such as nucleic acid targeting sequences ('guides') to coordinate proximity of the correct genotype-phenotype pairs (Lincoln and Joyce, 2009; Vaidya et al., 2012), or the eutectic phase of ice crystals preventing cross reactivity of genotypes (Attwater et al., 2010).

Compartmentalization has been an invaluable tool for the *in vitro* evolution of biomolecules. In 1998, Tawfik and Griffiths published a seminal paper describing the creation of water-in-oil emulsion compartments to perform *in vitro* molecular evolution (Tawfik and Griffiths, 1998). By combining particular recipes of oil, surfactant (amphiphilic molecules), and an aqueous solution using a mechanical force, miniature

compartments ranging from 2-4 microns could be created. In a single milliliter there are up to 10^{10} emulsion droplets that are approximately four femtolitres in volume (roughly the volume contained in a bacterial cell). Each of these emulsion bubbles act as a distinct molecular bioreactor, enabling a massively parallel experiment. This is achieved because compartments do not cross react - large molecules such as nucleic acids and proteins cannot escape a given compartment. This sets up the necessary requirement for molecular evolution - the genotype and phenotype linkage.

Initially, molecular evolution experiments were focused around the modification of the template molecule itself. DNA templates are emulsified with an *in vitro* transcription / translation mix, such that a single template will enter a single emulsion bubble. The DNA template is designed with a promoter sequence (typically the T7 RNA polymerase promoter), which will transcribe the encoded gene. Subsequently, translation of the encoded gene will occur from the mRNA transcript using the translation apparatus in the mix. Usually, this machinery comes from cell lysates that are depleted of endogenous mRNA, but can also be purified reconstituted systems such as the PURE system (Shimizu et al., 2001).

Using the basic principles of emulsions, Tawfik and coworkers over the next several years put out successive proof of principle variations of this technique. In the initial paper, the *in vitro* selection was based around the methylation of the DNA encoding HaeIII methyltransferase which protected templates from digestion during HaeIII endonuclease treatment (Tawfik and Griffiths, 1998). In a subsequent experiment, the amino acids involved in DNA template recognition were investigated by

randomization of residues followed by reselection of the native recognition sequence (Lee et al., 2002). This demonstrated the utility of emulsion based directed evolution for probing enzyme function. Following this, the DNA recognition of the HaeIII methyltransferase was modified – validating that the technique could alter enzymatic function (Cohen et al., 2004). Amino acid residues in the HaeIII methyltransferase were randomized and tasked with recognition and methylation of an altered DNA site (GGCC to AGCC) to protect from subsequent digestion from the NheI endonuclease. The resulting mutants of HaeIII methylated this altered site with several hundred-fold improvement of catalytic efficiency, but the mutant was promiscuous for several DNA sites. Given results from other experiments, the specificity could likely be improved by negative selection steps.

Additional variations of the emulsion scheme were developed. One of the main disadvantages of the emulsification is that once emulsified manipulation of the genes of proteins inside is difficult. For instance, the function of a particular enzyme (buffer conditions, substrates, temperature) may not be compatible with the *in vitro* transcription / translation system used to produce the proteins in the first place. Additionally, the previous examples have been limited to enzymatic modification of the template DNA. Many gene functions of interest however do not modify DNA. In order to bypass these issues, several additional techniques were layered. The use of fluorescence based cell sorting (FACS) was coupled with the emulsion based system by creating a double emulsion (Bernath et al., 2004). Because FACS instruments are not designed to microfluidically handle oil phase, oil based droplets were put through a ‘secondary

emulsion' which made the main constituent an aqueous phase (a so called water in oil in water emulsion). Accumulation of fluorescent based products inside individual emulsion bubbles enables the FACS instrument to sort individual compartments. A second scheme was also invented to bypass these limitations, wherein streptavidin coated beads were included in the emulsion reaction. These beads serve to capture products of the emulsion reaction during the *in vitro* transcription and translation phase. For instance, DNA can be captured by the addition of a biotin labeled tag and proteins can be captured with antibodies affixed to the bead's surface (Sepp et al., 2002). Upon breaking the emulsion, genes and proteins are linked together by physical attachment to the bead and can be probed for function. For example, enzymatic labeling of a fluorescent dye to the bead enables FACS based enrichment of active genes. These techniques were used for the evolution of a phosphotriesterase for faster catalytic turnover from a large library of variants ($>10^7$) (Griffiths and Tawfik, 2003).

In addition to the selection of protein enzymes, the emulsion based techniques have been used for selection of catalytic RNAs. The Bartel ligase was evolved using a bead based affixing technique followed by subsequent FACS (Levy et al., 2005), selecting RNA ligases that were more efficient at multiple catalytic turnover. Similarly, a Diels-Alderase ribozyme was selected from a random pool of ribonucleotides by emulsification and catalytic attachment of a biotin to the template by formation of the cyclohexene ring (Agresti et al., 2005). And perhaps most impressively, ribozyme polymerases have been evolved using emulsion based methods (Wochner et al., 2011; Zaher and Unrau, 2007). These methods have greatly improved the polymerase ribozyme

originally selected by David Bartel's group (Johnston et al., 2001), creating polymerases capable of extending a primer hundreds of nucleotides (Attwater et al., 2013; Wochner et al., 2011).

Emulsions also have the capacity for evolving much more complex phenotypes. Several evolutionary systems have been evolved in emulsion compartments that almost begin to resemble phage like lifecycles. For instance, synthetic operons have been evolved in emulsions that require the cooperativity of multiple gene functions. In one such system, the combined function of streptavidin and the biotin ligase, BirA, must function together to survive (Levy and Ellington, 2008; Lu and Ellington, 2014). Another example is the simple self-replicating T7 system, termed the 'autogene,' which requires a feedback based system for self amplification. Emulsification of the autogene system allowed the evolution of T7 RNAP over multiple cycles of self-replication (Davidson et al., 2012). In the most life-like example, the Q-beta replicase was evolved in an *E. coli* transcription / translation system (Ichihashi et al., 2013). This intriguing study found that emulsification was essential for evolving the Q-beta replicase by preventing accumulation of molecular parasites. Over the course of evolution, the system adapted to mitigate the molecular parasites that would arise throughout genome replication. These examples show that emulsion-based compartmentalized systems can evolve and have protocellular-like behavior.

Dividing a single reaction into billions of small reaction vessels is an extremely useful technique beyond molecular evolution technologies. The ability to encapsulate a single DNA template and perform amplification reactions has proven valuable for next-

generation sequencing technologies (Margulies et al., 2005; Shendure, 2005). Single DNA molecules can be replicated giving rise to monoclonal amplification, which can create a bead homogenously labeled with DNA from a complex pool of DNAs. The same principle can be used for amplifying RNAs by the addition of a reverse transcriptase to clonally amplify single RNA molecules (Nakano et al., 2005). Similar techniques can be used to identify rare single nucleotide polymorphisms that may be disease markers (Dressman et al., 2003). Emulsion PCR based amplification of rare alleles offers a number of advantages over single reaction amplification, as rare sequences will not be lost due to being out competed in PCR replication. Additionally, pairs of genes can be linked inside emulsion compartments which has proven extremely valuable for the pairing of antibody variable regions in the heavy and light chains (DeKosky et al., 2013, 2015).

POLYMERASE EVOLUTION USING COMPARTMENTALIZED SELF REPLICATION

A variation of the methods developed by Griffiths and Tawfik was conceived by Philipp Holliger for the evolution of thermostable DNA polymerases (Ghadessy et al., 2001). This selection scheme relies on the self-replication of thermostable DNA polymerases inside of an emulsion reaction. Compartmentalization of *E. coli* cells bearing the polymerase variants (and corresponding templates) with buffers, dNTPs, and primers provides the means for self replication to occur upon thermal cycling.

This compartmentalized self replication (CSR) scheme has several features that are distinct from the aforementioned methods. First, instead of using *in vitro* transcription and translation mixes, the protein production is achieved inside of *E. coli* cells. Libraries

of polymerase variants are constructed and transformed into cells, which also serve to maintain a genotype-phenotype linkage. Expression of polymerases inside of cells has several advantages over the use of lysates. Multiple copies of the template are available during the subsequent reaction (for multicopy plasmids) and protein expression inside of *E. coli* is generally more robust than those achieved in lysates. Second, the emulsion itself must support thermal cycling and sustained high temperatures without collapse of individual compartments. Third, the CSR directly amplifies the templates inside of the emulsion instead of relying on the post-emulsion selection of templates through modification, sorting, or affinity binding. The direct amplification enables modulation of selection parameters by altering the reaction conditions present to the polymerase. These customizable parameters enable the selection of polymerases with an assortment of properties (Fig. I.2).

Perhaps the simplest CSR evolution scheme is the addition of molecules that would normally inhibit PCR polymerization. In the original manuscript, it was shown that variants of Taq polymerase could be evolved that have increased resistance to heparin (which binds to the polymerase's active site inhibiting primer extension) (Ghadessy et al., 2001). In a subsequent study, gene shuffling of Taq with orthologous sequences followed by CSR containing various environmental inhibitors led to the discovery of polymerases that are inhibitor resistant (Baar et al., 2011). These polymerases are of diagnostic and forensic value, as oftentimes environmental or biological samples contain inhibitors to PCR.

The modification of DNA primer sequences in CSR can also lead to the evolution of polymerases with interesting properties. For instance, designing primer sequences that purposefully contain mismatches at their 3' end can evolve polymerase variants with unique properties. Polymerases evolved under these conditions are less fussy about the 3' complementarity and can bypass common template blocking lesions such as abasic sites (Ghadessy et al., 2004). These polymerases were also shown to behave more favorably when incorporating unnatural base analogs such as 7-deaza-dGTP, as incorporation of these bases might prevent further extension and inhibit PCR. The primer modification technique was further employed for more drastic mismatches in the primer-template complex in a subsequent selection (d'Abbadie et al., 2007). The resulting polymerases were used for ancient DNA samples that contain damaged genetic information which would otherwise prevent effective PCR amplification for sequencing reactions.

During self-replication in the emulsion compartment, swapping the native dNTPs for nucleotide analogs puts evolutionary pressure during CSR for the incorporation of these unnatural bases. In order to achieve effective self-replication polymerases must expand their substrate specificity to accommodate the substrates in the given reaction. For instance, substitution of dATP for ATP in the CSR reaction forces polymerase variants to incorporate ATP in order to efficiently self-replicate. Using this approach, several rounds of evolution produced polymerases that were capable of being a DNA, RNA, and reverse transcriptase simultaneously (Ong et al., 2006). This approach has been used for other base analogs, such as, fluorescently labeled dNTPs and hydrophobic base analogs (Loakes et al., 2009; Ramsay et al., 2010). However, the incorporation of

such bases is a large evolutionary hurdle and required the self-replication of just the DNA comprising the active site. In addition, reverse transcriptases which could convert this information back into DNA would be needed.

To bypass these issues a new selection methodology was created that would enable the evolution of polymerases with wildly altered sugar backbones. Instead of relying on the entire self-replication of polymerase sequences, a strategy was developed that would instead enable the capturing of the plasmids that encoded the polymerase variants (Pinheiro et al., 2012). Polymerase libraries were constructed and selected to polymerize a number of modified sugar dNTPs including: CeNAs, LNAs, ANAs, FANA, and TNAs. These polymerases could support *in vitro* evolution of functional nucleic acids including aptamers and ribozymes that are entirely composed of synthetic DNA analogs (Taylor et al., 2015).

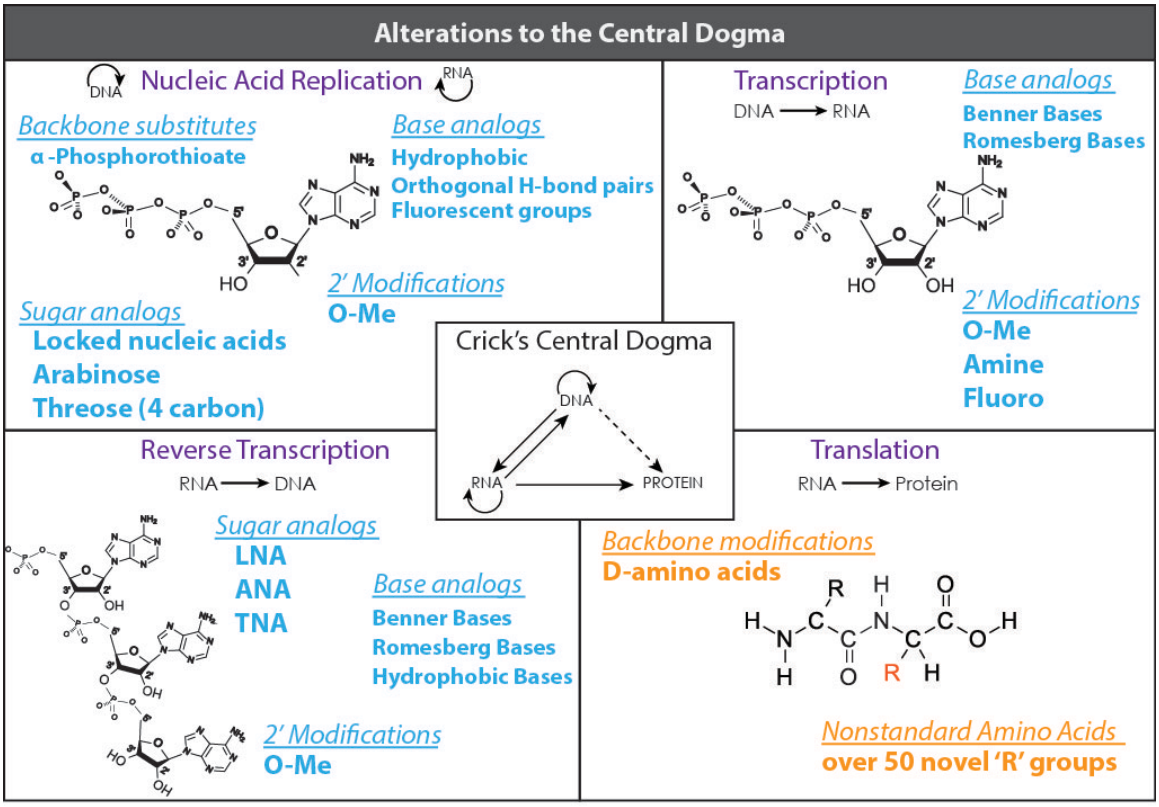


Figure I.1 : Alterations to the central dogma

Although Crick's Central Dogma is highly conserved (performing the chemistries of life's functions), it is surprisingly malleable. Natural substrate flexibility or engineering efforts have left the central dogma with many altered functionalities.

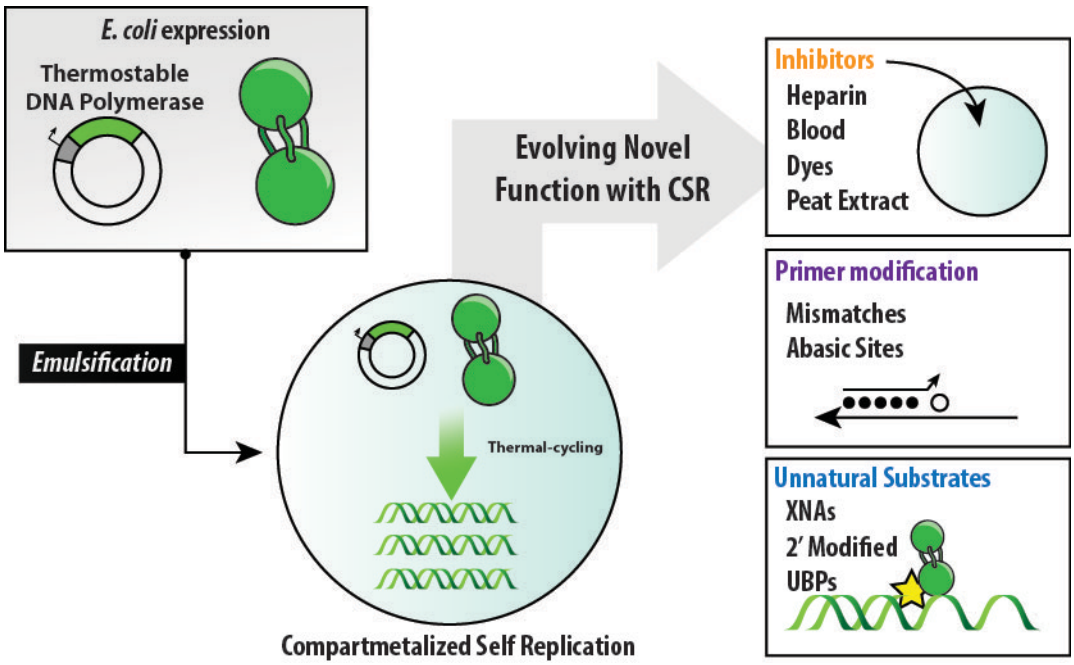


Figure I.2 : Evolving novel function using Compartmentalized Self Replication

Compartmentalized Self Replication (CSR) is widely adaptable for evolving polymerase function. Slight modifications to the selection parameters have enabled the evolution of polymerases with many novel functionalities. For example, CSR has evolved polymerases with inhibitor resistance, utilization of mismatched 3' ends, and the use of unnatural substrates.

CHAPTERS

Chapter 1: Directed Evolution of Genetic Parts and Circuits by Compartmentalized Partnered Replication

Most existing in vivo and in vitro directed evolution methods suffer from inadvertent selective pressures (i.e. altering organism fitness), resulting in the evolution of products with unintended or suboptimal function. To overcome these barriers, here we present compartmentalized partnered replication (CPR). In this approach, synthetic circuits are linked to the production of Taq DNA polymerase so that evolved circuits that most efficiently drive Taq DNA polymerase production are enriched by exponential amplification during a subsequent emulsion PCR step. We apply CPR to evolve a T7 RNA polymerase variant that recognizes an orthogonal promoter and to reengineer the tryptophanyl tRNA-synthetase:suppressor tRNA pair from Saccharomyces cerevisiae (Hughes and Ellington, 2010) to efficiently and site-specifically incorporate an unnatural amino acid into proteins. In both cases, the CPR-evolved parts were more orthogonal and/or more active than variants evolved using other methods. CPR should be useful for evolving any genetic part or circuit that can be linked to Taq DNA polymerase expression.

INTRODUCTION

With the emergence of sophisticated DNA synthesis and cloning techniques, the creation of in vivo-based synthetic circuitry has become commonplace (Gibson, 2011; Hughes et al., 2011). However, in contrast to the relative ease with which one can design

electronic circuits exhibiting predictable and precise behavior, it is not yet possible to reliably design biological pathways of equivalent complexity (Temme et al., 2012a). Most methods for refining synthetic circuits so that they behave in a more predictable and reliable manner rely on the bioengineering principles of screening and selection; however, this process has proven challenging in part because of the inherent linkage of the host organism's fitness and the circuit's function (Tan et al., 2009). Evolution and engineering techniques including phage-assisted continuous evolution (PACE) (Esvelt et al., 2011) and multiplex automated genomic engineering (MAGE) (Wang et al., 2009) were developed to better facilitate engineering of synthetic circuits, but each technique has limitations. PACE relies on a viral replication cycle, thus limiting control over the evolutionary process; while MAGE has relied on manual screening of the libraries generated and is therefore time consuming and laborious. Here we present compartmentalized partnered replication (CPR), an engineering platform capable of efficiently improving synthetic gene circuitry by utilizing advantages of both in vivo and in vitro approaches.

In a previously described related method called compartmentalized self-replication, DNA polymerases produced in cells facilitate the compartmentalized in vitro amplification of their own genes (Ghadessy et al., 2001),(Ghadessy and Holliger, 2007). While this technique has proven to be effective for the evolution of novel DNA polymerases, the method is of limited utility for the evolution of other genes or genetic circuits. CPR addresses this limitation by allowing coupling of Taq DNA polymerase production to a variety of other gene functions; these gene functions form the synthetic

circuit of interest. As such, synthetic circuits that most efficiently drive Taq DNA polymerase production in cells in the in vivo step will be preferentially amplified during the subsequent compartmentalized in vitro PCR step (Fig. 1.1). More specifically, a library of genetic circuits or parts designed to drive expression of Taq DNA polymerase is transformed into E. coli. Host cells are then separated into emulsion compartments that contain primers specific for the genetic circuit or part. Upon thermal cycling, cell lysis occurs and the compartment-to-compartment variations in abundance of Taq DNA polymerase result in preferential amplification of genes in compartments containing the most Taq DNA polymerase; these compartments correspond to the cells transformed with the most efficient genetic circuits or parts. CPR is a modular platform that should in theory be capable of evolving almost any genetic part or circuit that can influence Taq DNA polymerase production or activity. Here, to demonstrate the capacity of CPR to effectively evolve genetic circuitry and to facilitate comparison to other molecular engineering methods, CPR was wired to evolve circuitry for two important and ongoing synthetic biology efforts: the generation of orthogonal transcription machinery and expanding the genetic code.

RESULTS

Evolution of T7 RNA Polymerase Promoter Recognition

We first constructed a genetic circuit in which transcription of Taq DNA polymerase relies on T7 RNA polymerase (T7 RNAP) binding to and activating its promoter (Figure 1b). As a proof of principle, we initially coupled the production of Taq DNA polymerase to a wild-type (WT) T7 RNA polymerase promoter (PT7), and

generated a library of T7 RNA polymerase variants in which the six amino acids of the T7 RNAP specificity loop (Cheetham et al., 1999) (R746, L747, N748, R756, L757, Q758; RLN…RLQ) were fully randomized. The library was transformed into E. coli, and the cells were emulsified in the presence of primers specific for the T7 RNAP gene. Presumably, cells expressing those T7 RNAP variants that most efficiently bound to and activated PT7 contained higher amounts of Taq DNA polymerase. Heat lysis released Taq DNA polymerase and T7 RNAP gene variants into individual emulsion bubbles, and the emulsion bubbles were subjected to thermal cycling. After four iterative rounds of this process, the library enriched for variants driving robust expression from the PT7 (As shown by GFP expression from PT7) and converged on a consensus sequence resembling that of wild-type T7 RNAP, although some variation remained at the non-critical positions L747 and L757.

Based on this result, we attempted to evolve more orthogonal circuitry. Specifically, the randomized specificity loop library was tasked with driving Taq DNA polymerase expression from a promoter, PCGG, which differed from PT7 at -11, -10, and -9 positions. After seven rounds of CPR, a single sequence “RVH…EMQ” dominated the population, and transcriptional activity of an evolved variant bearing this motif (CGG-R7-8) was analyzed. While the CGG-R7-8 T7 RNAP activated expression of GFP driven by the PCGG promoter, this activation was equivalent to only 2% of the activity of WT T7 RNAP on the WT PT7 promoter. Therefore, we reasoned that additional T7 RNAP mutations, not present in the initial library, were needed to compensate for the changes to the promoter sequence.

To further improve the activity of the CGG-R7-8 polymerase on the PCGG promoter, we performed an addition five rounds of selection by CPR, with a larger (ca. 500bp) region of the CGG-R7-8 polymerase immediately flanking the specificity loop subjected to error prone PCR before the first and fourth round. We transformed the Round 12 population into E. coli expressing a PCGG-driven GFP reporter, and characterized the 8 T7 RNAP variants that drove the highest GFP expression. The most active T7 RNA polymerase variant from this population, which we termed CGG-R12-KI, displayed 20% the activity of the WT T7 RNAP:promoter pair .

Additional mutations beyond those seen in CGG-R12-KI were also frequently observed in variants in the round 12 population. We inserted combinations of three of these additional highly represented mutations to CGG-R12-KI; this further improved its in vivo activity to roughly 40-60% of the WT T7 RNAP:promoter pair. The five most active variants were mixed in equal ratios and selected by CPR for an additional 4 rounds (with error prone PCR before the first, third, and fourth round). This resulting round 16 population closely resembled the mutant CGG-R12-KIRV which robustly drove the expression of GFP from the PCGG promoter but demonstrated only minimal activity on the PT7 promoter (~1% cross-reactivity) (Fig. 1.2). In turn, the WT T7 RNAP did not markedly drive expression of the PCGG-GFP reporter in vivo. When expressed, purified, and assayed in vitro, CGG-R12-KIRV and WT T7 RNAP also demonstrated less than 0.1% promoter cross-creativity.

Although T7 RNAP variants bearing altered promoter specificity have been evolved and engineered over two decades by various means (Chelliserrykattil et al., 2001;

Esvelt et al., 2011; Raskin et al., 1993; Temme et al., 2012b), the CPR-evolved CGG-R12-KIRV is the most orthogonal and among the most active T7 RNAP:promoter pair when compared with previously described engineered and evolved T7 RNAP variants (Fig. 1.2).

Evolution of Aminoacyl tRNA Synthetase Amino Acid Specificity

Next, CPR was applied to evolve an orthogonal aminoacyl tRNA synthetase:tRNA pair facilitating site-directed incorporation of an unnatural amino acid, 5-hydroxy-L-tryptophan (5HTP). Previously, the tryptophanyl tRNA synthetase from Saccharomyces cerevisiae (ScWRS) and its corresponding suppressor tRNA have been adapted to the orthogonal suppression of amber codons in E. coli(Hughes and Ellington, 2010). To redirect CPR for evolving synthetic translation machinery, we generated a Taq DNA polymerase variant containing amber stop codons, and attempted to evolve variants of both a tRNA synthetase and its cognate suppressor tRNA with altered substrate specificity and improved amber suppression efficiency.

Based on the previously described crystal structure of the yeast tryptophanyl tRNA synthetase, we randomized three residues adjacent to the presumed position of the 5-hydroxy moiety in the tryptophan analogue (T107, P254, and C255) (Zhou et al., 2010). The library was transformed into an E. coli strain which expressed an amber codon-containing Taq DNA polymerase and grown in media supplemented with 5HTP. After each of three rounds of CPR, the library became more enriched for tRNA synthetase variants capable of amber suppression. Mock selections indicated that we could expect up to 100-fold enrichment per round of CPR; thus three rounds of selection

and amplification should have substantively narrowed the initial pool of tRNA synthetases. Individual tRNA synthetase variants were picked and screened for activity using a modified E. coli strain with an amber codon in the β-galactosidase gene. Several tRNA synthetase variants effectively suppressed the amber codon, resulting in a visibly blue colony, only when 5HTP was present (Fig. 1.3). The most active and specific utilizer of 5HTP , 5OH-R3-13, had the amino acid substitutions T107C, P254T, and C255A. Mutational analysis of single and double mutants of 5OH-R3-13 demonstrated that although it does not directly contact the 5-hydroxy moiety, the T107C mutation was the crucial step towards specificity (Fig. 1.4). In the wild-type tRNA synthetase binding pocket, T107 appears to be hydrogen-bonding with the carbonyl oxygen between residues P254 and C255. Mutation of T107 may eliminate this bond and instead allow the 5-hydroxyl moiety to hydrogen-bond with the carbonyl oxygen. Further specificity is gained after mutating P254C, potentially stabilizing the pocket with an additional hydrogen bond to T127 or perhaps allowing greater flexibility to the beta-sheet adjacent to the 5-hydroxyl moiety, while C255A may prevent a possible steric clash.

To determine if the variant 5OH-R3-13 site-specifically incorporated 5HTP, the enzyme was co-expressed with a modified version of dihydrofolate reductase (DHFR) that contained an amber codon at position 10 (V10amber) in the presence of rich 2xYT media supplemented with 5HTP. Purified DHFR proteins were characterized by top-down ultraviolet photodissociation mass spectrometry (Shaw, J.B., Li, W., Holden, D.D., Zhang, Y., Griep-Raming, J., Fellers, R.T., Early, B.P., Thomas, P.M., Kelleher, N.L., Brodbelt, 2013). The presence of a single mass shift of ~16 Da, indicative of the single

incorporation of 5HTP in place of tryptophan, was detected in the 5OH-R3-13 samples but not in the WT tRNA synthetase samples (Fig. 1.3). Fragmentation analysis verified that 5HTP was in fact incorporated at position 10, corresponding to the location of the amber codon. Incorporation fidelity was estimated to be approximately 85%, despite growth in rich media in the presence of abundant tryptophan. When compared with the wild-type ScWRS enzyme the evolved 5OH-R3-13 synthetase showed 1,500-fold improvement incorporating 5HTP (ratio of peak density in.

Evolution and Optimization of Orthogonal tRNAs

Optimization of orthogonal tRNAs (as opposed to tRNA synthetases) can also have a substantial impact on the efficiency of unnatural amino acid incorporation (Young et al., 2010). As the S. cerevisiae tRNA is already highly efficient, it is unlikely that its interaction with other parts of the E. coli translation machinery, such as elongation factors or the ribosome, will be optimized. Three tRNA libraries were constructed randomizing either the anticodon stem, acceptor stem, or loop sequences (Fig. 1.5). Each library contained roughly 10^6 different tRNA variants, and we subjected each library to ten rounds of CPR. To increase selection pressure, we progressively increased the number of amber codons in the open reading frame of the Taq DNA polymerase (up to six), as well as, reducing the expression of the orthogonal tRNA synthetase. Moreover, to show that CPR could potentially be used to co-evolve entire genetic circuits (as opposed to a single part in a circuit), the wild-type tRNA synthetase was also allowed to mutate during the selection of the suppressor tRNA libraries. After ten rounds, several circuits containing co-evolved tRNA synthetase and tRNA pairs were assayed via flow

cytometry by their ability to suppress GFP bearing 3 amber codons. Although some neutral and silent mutations were detected in the tRNA synthetases, as might be expected most of the mutations accrued within the tRNA libraries. Seven of ten tRNA variants displayed more efficient amber suppression than the parental suppressor tRNA (which itself had previously been improved over the wild-type yeast suppressor tRNA by rational mutagenesis). The two best tRNA variants (40A and 49A) were roughly 3- and 4-fold more active than the parental tRNA (AS3.4) and in consequence likely 12-fold better than the wild-type tRNA (Fig. 1.3). Both suppressor tRNA variants were assayed for cross-utilization by E. coli tRNA synthetases by determining whether they could suppress a single amber codon in a β-galactosidase or GFP gene, in the absence of functional ScWRS. The 49A tRNA variant displayed some background charging by endogenous synthetases, but the 40A variant (loop mutations U16G, G43U, U58G) appeared to be completely orthogonal.

We hypothesized that the efficiency of the CPR-evolved 5HTP incorporating 5OH-R3-13 tRNA synthetase would be improved by pairing it with the CPR-evolved 40A suppressor tRNA. When combined, these optimized parts drove 8.5-fold higher expression of GFP containing 3 amber codons in the presence of 5HTP relative to the parental AS3.4 tRNA (Fig. 1.3). Similarly, when DHFR (containing one amber) was expressed in cells expressing both the 5OH-R3-13 synthetase and 40A suppressor tRNA, mass spectrometry experiments confirmed no loss of fidelity of 5HTP incorporation despite the improved suppression efficiency. The efficiency with which 5HTP is incorporated into these proteins containing multiple amber codons is particularly

impressive given that other evolved synthetase:tRNA pairs have only been shown to be capable of suppressing one amber codon with an unnatural amino acid (Johnson et al., 2011).

DISCUSSION

Advantages of CPR Compared to Other Directed Evolution Techniques

One advantage of compartmentalized partnered replication is its ability to select for gene part or circuit function in vivo without being confounded by the need to simultaneously select for organismal fitness. This is because the short duration of the in vivo functional selection limits fitness effects that may disfavor highly active circuits. Another advantage is that the subsequent in vitro replication phase enables the exponential amplification of the most active circuits. These attributes likely lead to better functioning circuits being selected over a range of functionality (Fig. 1.6), meaning that even genetic parts or circuits resulting in initially weak phenotypes (in other words, inducing only a small difference in Taq DNA polymerase expression) can be established in the population and that the most active variants can ultimately come to dominate. Finally, the discontinuous nature of the selection process offers a much-needed element of control that limits parasite accumulation (Breaker and Joyce, 1994; Bull and Pease, 1995), avoids extraneous mutations impacting the selection (Dickinson et al., 2013; Goldsmith and Tawfik, 2009), and enables a part to be evolved largely in the context in which it will be used.

Future Directions

The evolution of transcription and translation machinery suggests that CPR can evolve enzymes and RNAs that are highly active and specific. By simply rewiring the in vivo circuit-based architecture to favor the production of Taq DNA polymerase in other ways, CPR can potentially be adapted to the evolution of regulatory parts such as transcription factors, repressors, and riboswitches, as well as larger genetic ensembles such as operons and biosynthetic pathways.

MATERIALS AND METHODS

T7 RNAP library design and selection

Site saturation mutagenesis was used to randomize the residues R746, L747, N748, R756, L757, and Q758 of the T7 RNAP promoter specificity loop. The degenerate oligonucleotide was synthesized in-house on an Expedite 8900 synthesizer using reagents and phosphoramidites purchased from Glen Research at a 40 nmol synthesis scale. Degeneracy was introduced into the oligonucleotides via the use of trimer phoshoramidites containing a mixture of 20 trimer (codon) phosphoramidites encoding all twenty amino acids. Primer-sets each introduced different silent mutations ("water-marks") up and down stream of the randomized regions. These allowed for specific amplification as well as identification upon sequencing.

The non-randomized portions of T7 RNAP were amplified from pQE-RSS. “T7 RSS” is a previously synthesized version of the T7 RNAP that was optimized for reduced secondary structure of the mRNA(Davidson et al., 2012). All selections and initial characterizations were performed with this codon set to avoid contamination during the

selection process. All protein purification and comparison to other mutants were performed with wild-type codon set in order to provide a more fair representation of said mutants. In all cases, every polymerase used the same codon set for a given assay.

The N-terminal portion of the T7 RNAP coding sequence was amplified from pQE-RSS; the C terminal portion of the T7 RNAP coding sequence was amplified from pQE-RSS. The three portions were then assembled by overlap PCR (N-term:library:C-term at a 1:2:1 molar ratio). The assembly PCR and an empty pQE vector were each digested with BamHI and HindIII (New England Biolabs). Insert and vector were mixed at a 2.3:1 molar ratio (~2.5 μg total), incubated with 2000 U T4 DNA ligase (New England Biolabs), and incubated at 14oC for at least 15 hours. All PCR steps (other than those in emulsion PCR) were performed using Accuprime Pfx DNA Polymerase (Life Technologies) per manufacturer's instructions. All PCR and digestion products were gel purified using QIAquick Gel Extraction Kit (Qiagen) and ligations were purified with SV Wizard PCR clean-up (Promega) prior to electroporation.

The wild-type Taq DNA polymerase gene was cloned into a modified pACYC-duet (Novagen) backbone with a single T7 promoter. Bl21 gold cells (Agilent) were transformed with pACYC-Taq (or its derivative with altered promoter) and grown in bulk overnight. 250 μl of this culture was subcultured in 20 ml 2xYT medium and grown at 37oC for 2 hours (OD600 ~0.5). The culture was then spun and washed with ice cold 10% glycerol 4 times, with the fourth resuspension in 100 μl 10% glycerol. This cell slurry (~200 μl total) was combined with 2-10 μl purified ligation and using electroporated 0.2 cm cuvettes at 2.5 kV in an E. coli pulser (Biorad). This routinely

resulted in 2 x 107 CFUs (multiple replicates were pooled for early rounds in order to attain full coverage).

100 µl overnight transformation cultures were subcultured in 2 ml 2xYT medium, grown for 1 hour (OD600 ~0.6) and induced with 0.3 mM IPTG at 37oC for 4 hours. 200 µl of the induction culture was centrifuged (10 min: 5,000g) to pellet the cells. The supernatant was removed and cells were gently resuspended in 20 µl 10x PCR buffer (500 mM KCl, 100 mM Tris-HCl pH 8.3, 15 mM MgCl2) 10 µl dNTP mix (4 mM each), 4 µl CPR.F primer (20 uM), 4 µl CPR.R primer (20 µM), and 162 µl water.

Emulsification was performed by slowly adding resuspended cells to 600 µl of spinning oil mix (438 µl Tegosoft DEC (Evonik), 42 µl AbilWE09 (Evonik), and 120 µl Mineral oil (Sigma)). The oil mixture was constantly spun in a tube (Sarstedt 13 ml 95 mm x 16.8 mm) on ice using a stirbar (Spinplus 9.5 mm x 9.5 mm Teflon, Bel-Art) on a magnetic plate (Corning) at the maximum setting (1150 rpm). The cell mixture was slowly added over a 1 minute interval and spun for an additional 4 minutes. The emulsified cells were thermal cycled (95^{o}C:3min, 20 cycles [95^{o}C:30s, 55^{o}C:30s, 72^{o}C:2min/kb], 72^{o}C:5 min) such that cells containing the most active enzymes will also contain the most Taq DNA polymerase and will preferentially PCR amplify. The emulsion was broken in two steps. Firstly, it was spun down by centrifugation (5 min: 10,000g) and the oil (upper) phase was removed. Secondly, 300 µl of H2O and 500 µl chloroform was added and the mixture was vortexed vigorously. The mixture was transferred to a heavy-gel phase-lock tube (5 Prime) and upon centrifugation (2min: 16,000g) the aqueous (upper) phase was collected along with any nucleic acids present.

To purify PCR amplified DNA from plasmid DNA we used a 5' biotinylated primer such that products amplified from Taq DNA polymerase can be purified away from plasmid DNA using streptavidin coated beads (MyOne Streptavidin C1 Dynabeads, Invitrogen). Purified DNA was used as a template for re-amplification using primers specific to the watermark introduced. This PCR product was used in an assembly PCR, followed by digestion and ligation as above.In the later rounds of the PCGG selection (after isolation of CGG-R7-8), a larger region of the polymerase coding sequence was recovered (and thus allowed to evolve).

CGG-R7-8 was subject to error-prone PCR and used as the input for CGG-R8. CGG-R12-KIV, CGG-R12-KIR, CGG-R12-KIRV, CGG-R12-KIGR, and CGG-R12-KIGRV were combined in equal amounts, subject to error-prone PCR and used as the input for CGG-R13. The recovered product of CGG-R10, CGG-14, and CGG-R15 were subject to error-prone PCR as described. Briefly the reaction mixture was composed of 50 mM KCl, 10 mM Tris-HCl pH 8.3, 2.5 mM MgCl2, 5 ug/ml BSA, 0.35 mM dATP, 0.4 mM dCTP, 0.2 mM dGTP, 1.35 mM dTTP, 0.5 mM MnCl2, 0.5 μM each primer, 2 ng/μl template, and 0.8 U/μl Taq DNAP (New England Biolabs) and was thermal cycled (95^{o}C:4 min, 25 cycles [95^{o}C:30s, 55^{o}C:30s, 72^{o}C:2m],72^{o}C:5 min). This achieved the expected 1 mutation per 500 bp. Individual variants from PT7-R4, CGG-R7, CGG-R12, CGG-R16 were sequenced and analyzed using Geneious software (Biomatters ltd).

In vivo RNAP activity assays

A FACS optimized variant of GFP, GFP mut2(Cormack et al., 1996), was cloned in place of the Taq DNAP open reading frame in pACYC-Taq. For measures of in vivo

activity of T7 RNAP, variants (plasmid) or pools (ligation) were electroporated into BL21 gold cells containing pACYC-GFPmut2 (or its derivative with altered promoter). Transformations were grown at 37^{o}C overnight. 100 µl of the culture was grown in 2 ml 2xYT medium at 37^{o}C for 1 hour (OD600 ~0.6) and induced at 0.05 mM IPTG for 4 hours. This concentration of IPTG was chosen in order to limit metabolic overload on the host and prevent saturation of signal. After induction, cells were measured for OD600 on a Synergy-HT plate reader (Bio-Tek) and GFP fluorescence (Excitation/ Emission 481/507) on a Safire monochromator (Tecan).

Images of T7 RNA polymerase-driven GFP expression shown in Figures 2a and S1 were generated by spinning down 10 ml of induced culture, decanting the supernatant, and resuspending cells in 500 µl PBS. The resuspended cells were excited with a UV transilluminator and imaged by Canon DSLR 500D digital camera.

In vivo RNAP cross-reactivity assays

A promiscuous mutant of the phenylalanine aminoacyl-tRNA synthetase, PheS A294G(Kast and Hennecke, 1991; Thyer et al., 2013), was cloned in place of the Taq DNAP open reading frame in pACYC-Taq. T7 RNAP mutant plasmids were electroporated into BL21-gold cells containing pACYC-PheS with PheS A294G driven by the wild-type T7 promoter. Transformations were grown at 37oC overnight. 100 µl of the culture was grown in 2 ml 2xYT medium at 37oC for 1 hour (OD600 ~0.6) and induced at 0.05 mM IPTG for 4 hours. Cells were diluted with media (containing the same antibiotics and IPTG as the growth media) to OD600s of 0.1, 0.01, and 0.001. 5 µl of each dilution was plated on 0 mM, 5 mM, 10 mM, 15 mM, or 20 mM 4-chloro-DL-

phenylalanine (Cl-Phe; Sigma). The plating media also contained 0.4% glycerol, 0.5% yeast extract, 1% NaCl, 1.5% agar, 50 μg/ ml kanamycin, 24 μg/ ml chloramphenicol, and 0.05 mM IPTG. Plates were grown at 37oC for 20 hours and imaged with ambient white light on a FluorChem Q (Protein Simple). Mutant cross-reactivity may be judged be the dose dependent cytotoxicity of Cl-Phe, which is only lethal (at the concentrations used) when PheS A294G is expressed.

In vitro transcription assays

For in vitro transcription assays, T7 RNAP variants were purified by standard Ni-NTA 6xHis (N-terminal) methods. The plasmid pQE-T7RSS (or a derivative thereof for T7 RNAP mutant) was transformed in BL21-gold (Agilent). Cells were grown in 2xYT media at 37oC until reaching OD600 ~0.7-0.8 at which point 1 mM IPTG was added. Cells were grown four hours at 37^{o}C. Following induction, cells were harvested by centrifugation and resuspended in binding buffer (50 mM Tris-Hcl, pH8.0, 0.5 M NaCl, 5 mM imidazole). Resuspended cells were lysed via sonication on ice using 40% probe amplitude for 2 minutes (1s ON, 1s OFF). Cell debris was pelleted by centrifugation (30min: 20,000g). His-tagged T7 RNAP was purified by immobilized metal affinity chromatography (IMAC). The lysate was run over 1 ml (bead volume) Ni-NTA gravity column pre-equilibrated with binding buffer. The column was washed with 10x column volumes of wash buffer (50 mM Tris-Hcl, pH 8.0, 0.5 M NaCl, 20 mM imidazole). T7 RNAP was eluted off the column by the addition of 4x column volumes of elution buffer (50 mM Tris-Hcl, pH 8.0, 0.5 M NaCl, 250 mM imidazole). Dialysis was performed in final storage buffer (50 mM Tris-Hcl, pH 8.0, 100 mM NaCl, 1 mM DDT, 1 mM EDTA).

Dialates were adjusted to 1 mg/ml and added to an equal volume of glycerol (final concentration 0.5 mg/ml).

Transcription reactions contained 40 mM Tris-HCl pH 7.0, 30 mM MgCl2, 6 mM spermadine, 6 mM each NTP, 10 mM DTT, 0.5 μM T7 RNAP, 0.5 μM DNA template, and 0.17 mg/ml DFHBI(Paige et al., 2011) in DMSO. Reactions were incubated for up to 2 hours at 37oC with spinach fluorescence (Excitation/ Emission 469/501) reading taken every minute in a Safire monochromator (Tecan). Spinach templates were made by thermal cycling 2 μM AJM.19 with 2 μM AJM.20 (for PT7-spinach) or 2 μM AJM.21 (for PCGG-spinach) with Accuprime Pfx in its standard buffer (94°C:2 min, 12 cycles [94°C:15 s, 50°C:30 s, 68°C:30 s], 68°C:1 min). Templates were purified by QIAquick Gel Extraction Kit (Qiagen).

Aminoacyl tRNA synthetase CPR for site specific 5-hydroxy-L-tryptophan incorporation

The binding pocket of the Saccharomyces cerevisiae tryptophanyl tRNA-Synthetase (ScWRS) was mutagenized by site-saturation mutagenesis at residues T107, P254, and C255 using NNS randomized oligonucleotides. These residues were chosen due to their proximity to the 5 position of tryptophan(Zhou et al., 2010) in the binding pocket of the synthetase. Libraries were prepared by overlap PCR using PFU Ultra II fusion HS (Agilent) and cloned into pRST.2TAA, a modified version of pRST.11BAS3.4(Hughes and Ellington, 2010) containing two ochre (TAA) stop codons, using HindIII and XhoI restriction endonucleases. The resulting library was transformed into E. coli BL21(DE3) (Invitrogen) harboring the plasmid pACYC.Taq.1Amb (W167Amber) which contains an

amber codon in the open reading frame of Taq DNA polymerase. The efficiency of transformation was >106 for each round of selection indicating several fold coverage of the library. Transformed cells were grown in 2xYT media (Sigma) overnight at 37^{o}C in carbenicillin (Cellgro) and chloramphenicol (Sigma). The following morning 10 µl of cells were seeded into 1 ml of fresh 2xYT media with appropriate antibiotics and 1 mM 5-L-Hydroxytryptophan (5HTP) (Sigma) and grown at 37^{o}C for 2 hours. The expression of the library of mutant synthetases and the Taq DNA polymerase was initiated by the addition of 1 mM IPTG. Cells were induced at 30°C for 7 hours. Cells were harvested (200 µl) by centrifugation (8 min: 3,000g) and removal of the supernatant. Cells were resuspended in 168 µl CPR buffer (45 mM KCL, 9 mM Tris-HCl (pH8.3), 1.4 mM $MgCl_2$, 0.5 nM each CPR primer (JE.33 and JE.35), and 250 µM each dNTP).

Emulsification was performed by slowly adding resuspended cells to 600 µl of spinning oil mix (438 µl Tegosoft DEC (Evonik), 42 µl AbilWE09 (Evonik), and 120 µl Mineral oil (Sigma)). The oil mixture was constantly spun in a tube (Sarstedt 13 ml 95 mm x 16.8 mm) on ice using a stirbar (Spinplus 9.5 mm x 9.5 mm Teflon, Bel-Art) on a magnetic plate (Corning) at the maximum setting (1150 rpm). The cell mixture was slowly added over a 1 minute interval and spun for an additional 4 minutes. The emulsified cells were thermal cycled (95^{o}C:3min, 20 cycles [95^{o}C:30s, 55^{o}C:30s, 72^{o}C:2min/kb], 72°C:5 min) to selectively amplify functional variants. The emulsion was broken in two steps. Firstly, it was spun down by centrifugation (5 min: 10,000g) and the oil (upper) phase was removed. Secondly, 300 µl of H_2O and 500 µl chloroform was added and the mixture was vortexed vigorously. The mixture was transferred to a heavy-

gel phase-lock tube (5 Prime) and upon centrifugation (2 min: 16,000g) the aqueous (upper) phase was collected. To purify CPR amplified DNA we used a 5' biotinylated primers, thus, products amplified by Taq DNA polymerase can be purified away from plasmid DNA using streptavidin coated beads (MyOne Streptavidin C1 Dynabeads, Invitrogen). Purified DNA was used as a template for re-amplification using nested primers. Reamplification products were cloned into the pRST.2TAA; completing one round of CPR selection. Three rounds of selection were carried out using the same reaction conditions. The pool activity was assayed using a β-galactosidase colony spot assay (Figure 3a). Blue colonies were counted for each of the rounds and divided by the total, indicating that each round of selection enriched for active synthetases. Synthetase variants were assayed for 5HTP specificity by patch plating colonies onto plates with or without the unnatural amino acid 5HTP (1 mM). Colonies that turned blue only in the presence of 5HTP were putative candidates for selective incorporation. As expected, some active variants were pulled from the initial Round 0 pool due to the small library size. Variant 5OH-R3-13 was chosen for further analysis as it displayed the most pronounced β-galactosidase activity.

tRNA CPR for improved amber suppression

CPR was used to select for enhanced amber suppression by randomizing key regions of the tRNA. Libraries were generated using the orthogonal tRNAAS3.4 as a starting point and residues in the anticodon stem (AS), acceptor stem (AX), and the loop sequences were randomized. CPR selections were carried out essentially as described for the 5HTP selection, except the stringency of selection was modulated to increase

selective advantage for the best suppressor tRNAs. To increase selective pressure on the tRNA, we modulated the: time of induction, temperature of induction, IPTG concentration, and most importantly the number of amber codons in the Taq DNA polymerase open reading frame. The recovery strategy also changed from streptavidin based capture to using CPR primers with unique sequences on the 5' end. During reamplification primers will anneal to the unique sequence preventing amplification of the contaminating plasmid DNA. Additional diversity was introduced by error-prone PCR after Rounds 4, 5, and 9. tRNA libraries were pooled after Round 3. Amber codons were introduced into the Taq DNA polymerase gene at positions W167, W169, W179, W211, W243, and W318. In pACYC-Taq, constructs with more than one amber codon were introduced additively from N to C terminus. After ten rounds of selection, tRNAs were screened by flow cytometry for the ability to suppress three amber codons in GFPmut2 (Figure 3c, S16). The most active tRNA variants, 40A and 49A, were further examined to determine if they work promiscuously with other synthetases; this was done using the β-galactosidase assay. tRNA variant 40A was completely inactive in the absence of a functional synthetase, while variant 49A displayed partial activity, suggesting it might be interacting with other synthetase machinery expressed by the host (data not shown).

β-galactosidase screening

tRNA synthetase variants cloned into pRST.11B backbones were transformed into CA274 E. coli cells, which contain an amber codon at position 125 of the LacZ gene. Individual colonies were patch plated onto 2xYT media plates containing 100 μg/ml

carbenicillin, 40 μg/ml X-Gal (Sigma), and 0.1 mM IPTG, with or without 1 mM 5HTP. Cells were grown at 37^{o}C for approximately 6 hours which resulted in visibly blue colonies for active synthetases.

GFP-FACS screening

Variants from either synthetase or tRNA libraries were cloned into pRST.11B backbones and analyzed via GFP. A FACS optimized variant of GFP, GFP mut2(Cormack et al., 1996), was cloned in place of the Taq DNAP open reading frame in pACYC-Taq and amber codons were introduced at positions Y39, Y151, Y182(Wang et al., 2001), resulting in plasmid pACYC.GFPmut2.3Amb. E. coli BL21(DE3) were co-transformed with individual synthetase or tRNA library variants and pACYC.GFPmut2.3Amb, and grown overnight at 37oC in 2xYT media. The following day, 10 μl of cells were diluted into 1 ml fresh 2xYT media (containing appropriate antibiotics and 5HTP when necessary) and grown for 2 hours at 37oC. Expression of both the synthetase and GFP were induced with the addition of 1 mM IPTG and grown for 4 hours at 37^{o}C. Cells were harvested by centrifugation (4^{o}C) and resuspended in phosphate buffered saline. Fluorescence analysis was performed on either a FACSCalibur (BD Biosciences) flow cytometer or on a Safire monochromator (Tecan) using GFP fluorescence (Excitation/ Emission 481/507).

DHFR purification and mass spectrometry

E. coli dihydrofolate reductase (DHFR) was used to assess the incorporation of the 5HTP amino acid. Plasmid pACYC.DHFR_V10Amb contains an amber codon at V10 of DHFR. Plasmids pRST.11B, pRST.5HTP, and pRST.5HTP.40A were co-

expressed with pACYC.DHFR_v10Amb in BL21(DE3). The strains were grown in 2xYT media at 37^{o}C until reaching OD600 ~0.7-0.8 at which point 1 mM 5HTP and 1 mM IPTG were added. Cells were grown overnight at 30^{o}C. Following induction cells were harvested by centrifugation and resuspended in binding buffer (50 mM Tris-HCl, pH8.0, 0.5 M NaCl, 5 mM imidazole). Resuspended cells were lysed via sonication on ice using 40% probe amplitude for 2 minutes (1s ON, 1s OFF). Cell debris was pelleted by centrifugation (30min: 20,000g). The His-tagged DHFR was purified by immobilized metal affinity chromatography (IMAC). The lysate was run over 1 ml (bead volume) Ni-NTA gravity column pre-equilibrated with binding buffer. The column was washed with 10x column volumes of binding buffer, 3x column volumes of wash 1 buffer (50 mM Tris-Hcl, pH 8.0, 0.5 M NaCl, 20 mM imidazole), and an additional 3x column volumes of wash 2 buffer (50 mM Tris-HCl, pH 8.0, 0.5 M NaCl, 30 mM imidazole). DHFR was eluted off the column by the addition of 4x column volumes of elution buffer (50 mM Tris-HCl, pH 8.0, 0.5 M NaCl, 250 mM imidazole). DHFR samples were centrifuged (10min: 16,000g) to remove insoluble protein and then loaded onto a FPLC and fractionated by size exclusion chromatography into 10 mM Tris-pH 8.0. The purest DHFR fractions were used for mass spectrometry analysis.

Top down ultraviolet photodissociation MS

DHFR (V10Amber) was expressed and purified as described above. Following purification, the proteins were buffer exchanged into LC-MS grade water using 3 kDa molecular weight cutoff filters. The proteins were diluted to 10 μM in a solution of 50/49/1 MeOH/water/formic acid. Proteins were infused at a flow rate of 5 μL/min and

ionized by electrospray ionization on a Thermo Scientific Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) modified for ultraviolet photodissociation (UVPD) in the HCD cell as described previously. A 193 nm excimer laser was used for UVPD. Intact molecular weight measurements at maximum resolution were undertaken to confirm the presence or absence of the expected single 5-hydroxy-L-tryptophan incorporation. The site of modification was confirmed using UVPD via a single 5 ns 193 nm laser pulse. The UVPD product ion spectra were also acquired at maximum resolution. Precursor and product ion spectra were interpreted manually and using the Xtract (ThermoFisher) deconvolution algorithm in conjunction with a beta version of ProSightPC 3.0. The Sequence Gazer tool was used to assign possible locations of the ~16 Da observed mass shift. Incorporation efficiency was calculated by subtracting the area of artifactual oxidation (ScWRS - 5HTP control) from the combined areas of the 5HTrp and singly oxidized 5HTP peaks, and dividing by the summed areas of all peaks (Trp, 5HTP, and 5HTP+oxidation). Peak area integration was performed using the 5 most abundant peaks for each isotope cluster. Fold enhancement was calculated by comparing the incorporation rates of 5HTP for ScWRS and R3-13. Incorporation rates for each were arrived at by dividing the area of the 5HTP containing peak by the combined areas of the 5HTP and naturally occurring reduced Trp containing peaks, and then applying a correction to account for artifactual oxidation during sample handling. The correction factor was attained by performing the same calculations as above for the singly oxidized peak using ScWRS in the absence of 5HTP in the media.

Comparison of tRNAs generated by existing methods and CPR

Existing approaches for generating orthogonal tRNAs have relied upon an in vivo life-death selection system. A recent publication (Chatterjee et al., 2013) used this traditional approach to generate optimized versions of the Saccharomyces cerevisiae orthogonal suppressor tRNA. To demonstrate the effectiveness of CPR, tRNAs were compared to each other for cross-reactivity to the native E. coli translation machinery and for amber suppression activity. The optimized tRNAs were cloned into the starting vector (sequences of tRNAs: AS3.4; 40A; H13; H14) in place of the AS3.4 tRNA as well as a vector containing a non-functional version of the synthetase (to test for tRNA orthogonality). The tRNAs (compared to AS3.4 and 40A) are in the AS3.51 background with acceptor stem mutations H13 and H14. To test for tRNA cross-reactivity to E. coli tRNA synthetases, tRNAs (without active S. cerevisiae aminoacyl tRNA synthetase) were expressed with GFPmut2 (Y39TAG; 1Amb), as described above and measured on a monochromator. Resulting fluorescence reflects tRNA charging with endogenous E. coli machinery and therefore lack of orthogonality. To further test orthogonality, tRNAs were transformed into E. coli strain CA274 (LacZ 1Amb). The relative efficiency of amber suppression was tested as described above.

Dynamic range of CPR

CPR mixes were processed as described (above) but purified Taq DNA polymerase (NEB) was added exogenously. A mixture consisting of an abundance of Taq DNA polymerase (0.8 µg; 5.1 x 10^{12} molecules) and 20 ng (3.8 x 10^{9} molecules) of a DNA template "A" was emulsified and added in equal volume to a second emulsion

consisting of a variable concentration of Taq DNA polymerase and 20 ng template "B" (which is identical to template A but contains an internal Hind III restriction endonuclease site). Emulsion PCRs were performed with various ratios of polymerase concentrations (A (0.8 µg):B (N µg)). Emulsions were broken, recovery PCR performed, and followed by a Hind III digest. The DNA was run on a gel resulting in two distinct bands, one corresponding to template A and the other template B. Intensities of template A and B were measured (ImageJ), normalized to the equal ratio condition (1:1), and plotted.

ACKNOWLEDGEMENTS

This work was supported by the National Security Science and Engineering Faculty (FA9550–10-1-0169) the Welch Foundation (F–1654 to ADE and F-1155 to JSB), the National Science Foundation (CHE1012622 to JSB), and the Defense Advanced Research Projects Agency (HR–0011-12-C-0066). JSB thanks Thermo Fisher Scientific with helping on the modifications to the Orbitrap Elite mass spectrometer to allow UVPD.

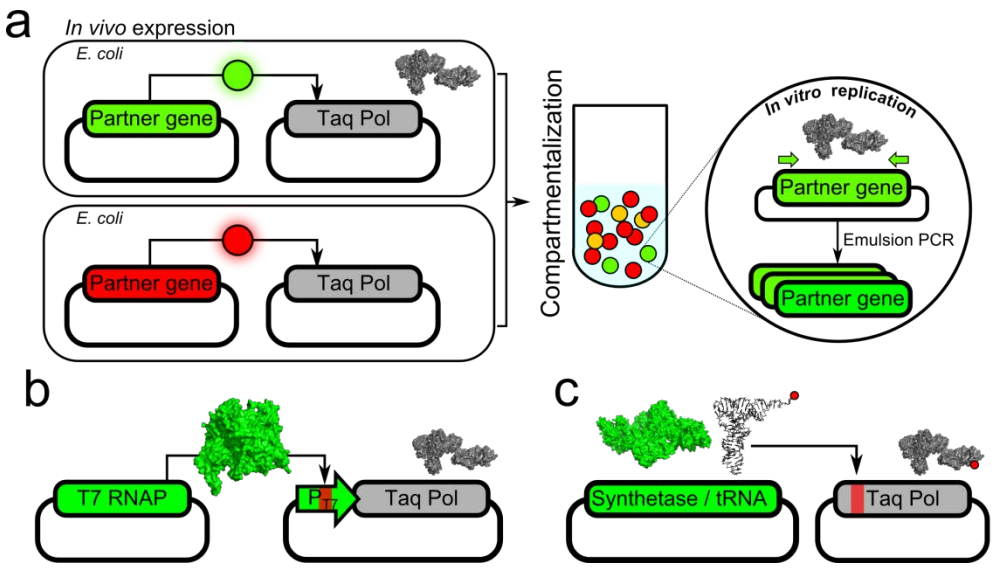


Figure 1.1 : Schematic of general CPR concept.

a, In the *in vivo* expression step, *E. coli* are transformed with genetic circuits or parts (partner genes) designed to drive production of *Taq* DNA polymerase (*Taq* pol). The partner gene-encoded biomolecules that display an active phenotype (green) produce *Taq* pol while inactive biomolecules (red) do not. Whole *E. coli* cells are compartmentalized via a water-in-oil emulsion along with primers, dNTPs, and *Taq* DNA polymerase buffer. Emulsions are thermal cycled, leading to *E. coli* cell lysis and preferential *in vitro* PCR amplification of partner genes that drove production of the most *Taq* DNA polymerase during the *in vivo* expression step. **b**, *In vivo* CPR design for the evolution of orthogonal T7 RNA polymerase: promoter pairs. A T7 RNA polymerase library drives the expression of *Taq* pol from a mutant promoter sequence. **c**, *In vivo* CPR diagram for the evolution of tRNA synthetase:suppressor tRNAs. A tRNA and/or tRNA synthetase library suppresses amber codons in the *Taq* pol gene to generate active polymerase.

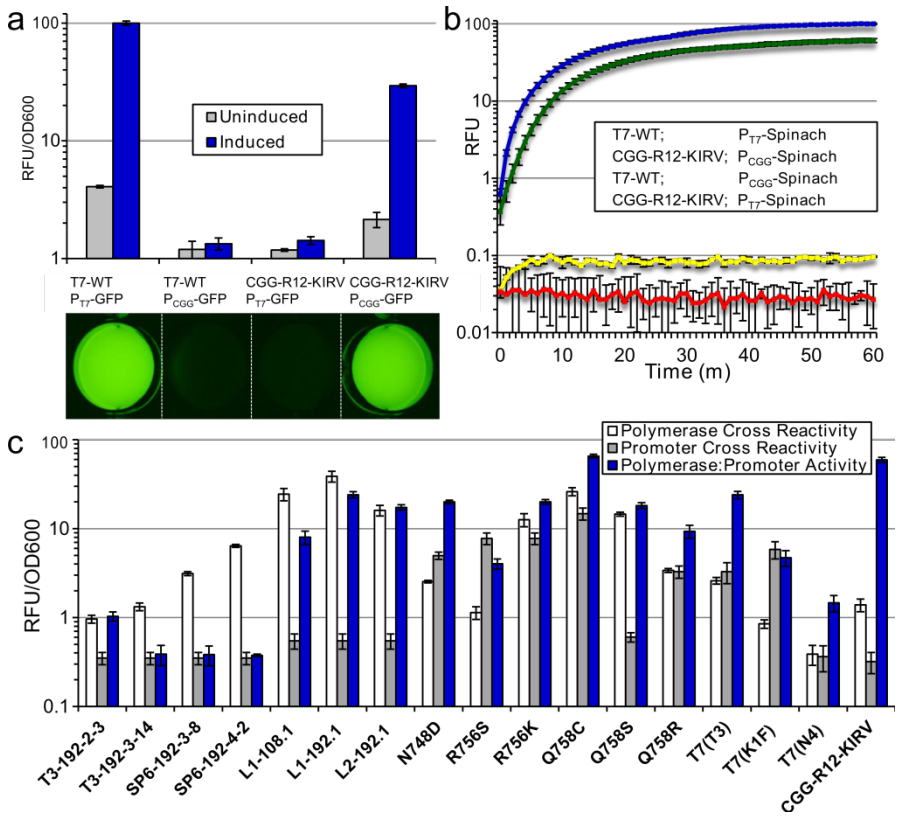


Figure 1.2 : CPR selection of an orthogonal T7 RNA polymerase.

a, Activity of WT T7 RNAP and the CPR-evolved variant CGG-R12-KIRV in *E. coli* cells expressing P_{T7} and P_{CGG}-driven GFP reporters (top). Induced cultures were imaged with UV transillumination and a digital camera (bottom). **b,** Activity of WT T7 RNAP and the CPR-evolved variant CGG-R12-KIRV in an *in vitro* transcription assay using P_{T7} and P_{CGG}-driven expression of the spinach aptamer as a readout. Spinach fluorescence was read every minute and plotted as a function of time. **c,** Activity of several evolved or engineered T7 RNAP variants in *E. coli* cells expressing P_{T7} and P_{cog}-driven GFP reporters. P_{cog} refers to the cognate promoter for the mutant being assayed. Polymerase Cross Reactivity (white) refers to the activity of the mutant T7 RNAP on the WT promoter. Promoter Cross Reactivity (grey) refers to the activity of the WT T7 RNAP on the mutant promoter. Polymerase:Promoter activity refers to the activity of the mutant polymerase on the mutant promoter. Fluorescence was quantified on a Tecan Safire monochromator. The WT pair's value was defined as 100 in each experiment. *In vivo* fluorescence was normalized to OD600; fluorescence/OD600 ratio reported is the average of three independently grown cultures. *In vitro* fluorescence reported is the average of three independently assembled transcription reactions. Error bars represent one standard deviation.

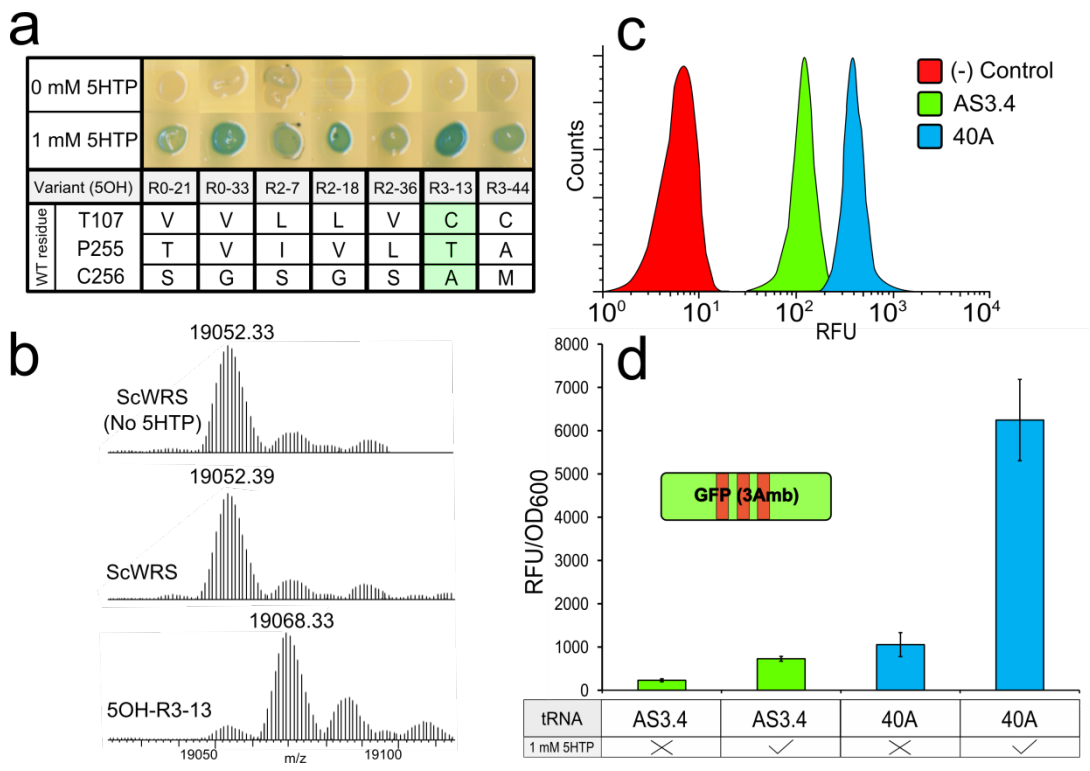


Figure 1.3 CPR evolved 5-hydroxy-L-tryptophan utilizing tRNA synthetase and optimized tRNA.

a, tRNA synthetase variants from rounds of CPR selection were assayed using β-galactosidase containing a single amber codon with and without supplemented 5HTP in the growth media. **b,** Deconvoluted intact whole-protein mass spectra of dihydrofolate reductase (containing an amber codon at position 10) with wild-type *S. cerevisiae* synthetase (ScWRS) which incorporates tryptophan (expected mass: ~19052) or with the evolved 5OH-R3-13 variant which demonstrates the ~16 Da mass shift expected from incorporation of 5HTP. **c,** tRNA amber suppression efficiency was quantitated by the ability to suppress 3 amber codons in GFP in conjunction with the wild-type tRNA synthetase. The parental AS3.4 and CPR-evolved 40A tRNA are compared by flow cytometry, demonstrating the optimized tRNA increases GFP production. **d,** Amber suppression efficiency was quantified via fluorescence by using GFP (3 amber) using the 5OH-R3-13 synthetase with the parental AS3.4 or CPR-evolved 40A tRNA in the presence and absence of supplemented 5HTP. Fluorescence/OD600 ratio reported is the average of three independently grown cultures; error bars represent one standard deviation.

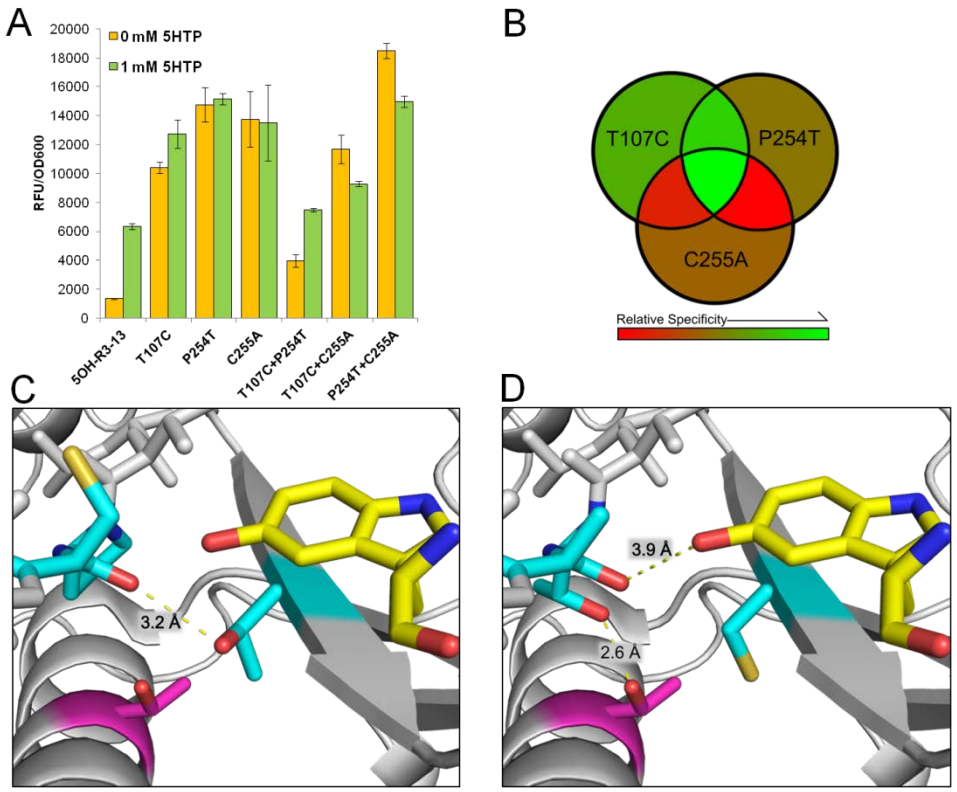


Figure 1.4 : Characterization of 5OH-R3-13 and Specificity Model.

(A) Wild-type S. cerevisiae tryptophanyl tRNA synthetase was mutated to contain single and double mutations of the 5HTP incorporating synthetase variant 5OH-R3-13 (T107C, P254T, C255A). Fluorescence was assayed in E. coli using GFP (1Amb) with and without 1 mM 5-hydroxy-L-tryptophan contained in the media. (B) Bonham projection of relative specificity between single, double and triple mutants of 5OH-R3-13. (C) S. cerevisiae tryptophanyl tRNA synthetase binding pocket with 5HTP (yellow) modeled into the binding pocket, library residues (cyan) and T127 (magenta). (D) Proposed model for binding pocket alterations that lead to 5 hydroxy-L-tryptophan specificity.

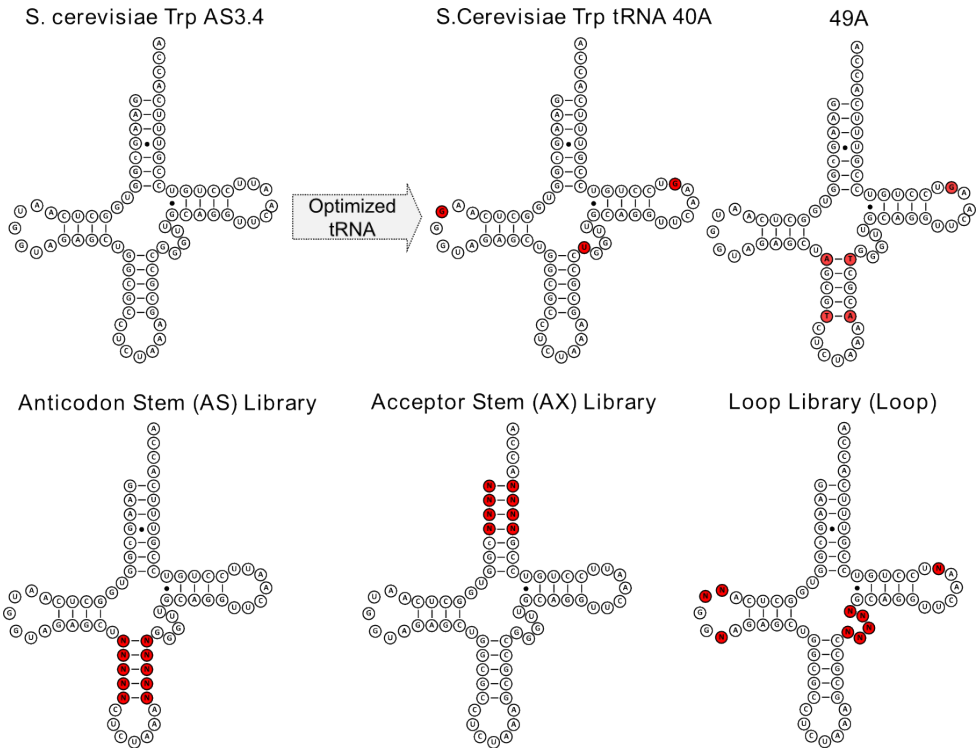


Figure 1.5 : Optimized tRNA sequence and tRNA libraries.

(Top) Cloverleaf structures and sequences of the orthogonal yeast tryptophanyl AS3.4 tRNA and the CPR evolved tRNAs (40A and 49A). (Bottom) tRNA libraries used as input for ten rounds of CPR selection.

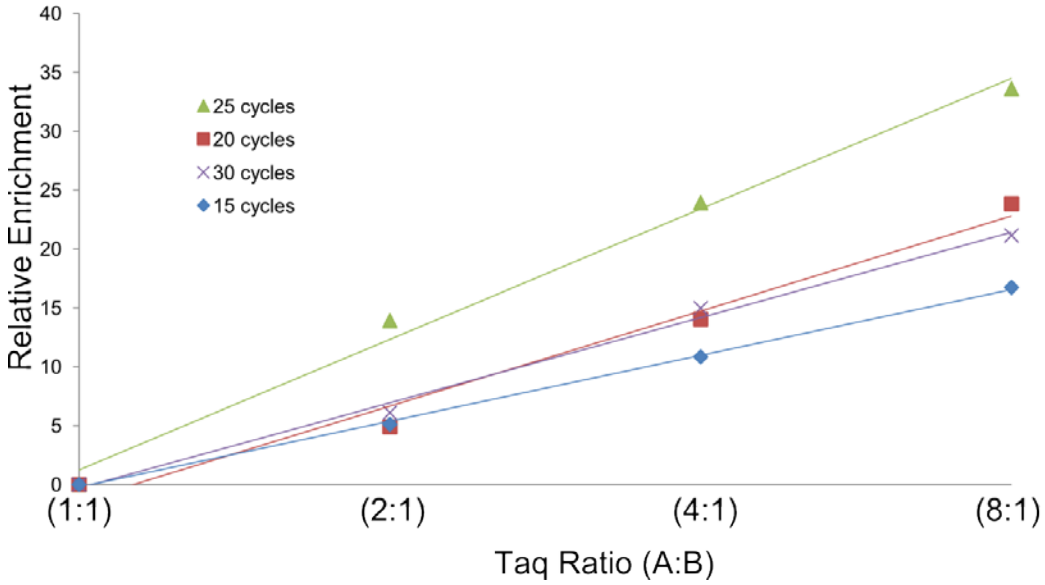


Figure 1.6 : *Taq* DNA polymerase abundance and cycle number influence dynamic range of selection.

A DNA template "A" was emulsified with an abundance of *Taq* DNA polymerase and mixed, in equal parts, with a separate emulsion containing template "B" with decreasing amount of *Taq* DNA polymerase. Emulsion mixes were thermal-cycled with variable numbers of cycles and the enrichment of “A” relative to “B” was determined.

Chapter 2: A Legacy Biosensor System Evolved for Allosteric and Intramolecular Logic

In this chapter, I describe the use of CPR for the directed evolution of biosensors. A single evolutionary "legacy" biosensor from *E. coli*, the tryptophan repressor (TrpR), was evolved and designed to confer a variety of novel responses. The wild-type TrpR regulates expression of the enzymes responsible for L-tryptophan biosynthesis by binding to a palindromic operator sequence upon allosteric activation by L-tryptophan. Using CPR, TrpR was evolved to respond to two novel effectors, 5-bromo-L-tryptophan and 6-bromo-L-tryptophan, in a specific manner. Biosensors possessing these novel regulatory features were then used to evolve novel DNA operator binding specificity, resulting in a total of fifteen unique biosensors that each respond to a novel effector molecule and operator sequence. In addition, higher-order regulatory architectures were designed by the intramolecular fusion of the two monomers that form the TrpR dimer. Tethering unique repressor monomers enabled NAND-gated effector logic and modularity of the DNA operator sequence, potentially enabling construction of dozens of unique repressor phenotypes.

INTRODUCTION

Organisms must dynamically respond to their environment, and intracellular contextual cues such as regulation of metabolic pathways, cell cycle progress, or threats from foreign viral infections (López-Maury et al., 2008; de Nadal et al., 2011). Cellular responses are mounted either by activation of biomolecular complexes or through signal

transduction pathways that will ultimately modulate gene expression in response to stimulus (Kiel et al., 2010). For instance, allosteric transcription factors (or biosensors) fine tune gene expression depending on the concentration of effectors molecule through the interplay between conformational folding states and promoter binding (Hilser et al., 2012; Motlagh et al., 2014). Biosensors have been widely used in synthetic biology for developing complex phenotypes like pattern formation, bacterial population behaviors, and as a tool for improving metabolic flux (Basu et al., 2005; Chen et al., 2015; Raman et al., 2014; Tabor et al., 2009).

Synthetic regulatory networks are often constructed from various evolutionary contexts or are part-mined from genomic databases (Nielsen et al., 2013). However, this may complicate the construction of genetic circuitry. Host context may render parts unusable due to cross-reactivity or auxiliary components being required (Rhodius et al., 2014; Temme et al., 2012). The exact biomolecular functions of parts from metagenomic data may also be unknown, requiring laborious efforts to identify function (Moon et al., 2012; Stanton et al., 2014). In contrast, well characterized and evolutionarily ancient parts that have evolved with host machinery are known to function in a predictable manner, so called “legacy parts.” Evolutionary theory predicts how legacy parts could expand to adopt new phenotypes over time through gene duplication, specialization, and recombination (Hughes, 1994; Ohno et al., 1968). Here we examine whether a single legacy biosensor system, the tryptophan repressor (TrpR), contains the functional and evolutionary plasticity to create diverse regulatory architectures.

The TrpR biosensor is the main regulatory component that governs expression of genes from the tryptophan biosynthetic operon in *E. coli* (Xie et al., 2003). Two TrpR monomers undergo dimerization to form an aporepressor complex that will ultimately bind two palindromic operator sequences (O_{WT} CTAGTAC). Increased cellular concentrations of the small molecule, L-tryptophan, triggers the complex to bind these operators with high affinity, effectively inhibiting gene expression from the P_{TRP} promoter by blocking transcription machinery access (Hurlburt and Yanofsky, 1992; Yang et al., 1996).

To expand the function of the allosteric TrpR biosensor, we utilized CPR which we have previously been shown to be a powerful tool for the directed evolution of individual parts or circuit function (Ellefson et al., 2014; Meyer et al., 2015). In CPR, the *in vivo* function of a circuit is linked to the expression of a thermostable DNA polymerase, Taq, as an output. Individual circuits (transformed into *E. coli* cells) achieve a spectrum of polymerase production based on behavior of library variants. Subsequently, up to 10^9 circuits undergo emulsion PCR – enabling the simultaneous screening of all circuits in the library. The most active circuits express the greatest levels of Taq polymerase and will subsequently achieve greater amplification in the emulsion PCR – allowing *in vitro* selection over a wide range of circuit operations. The CPR circuit was adapted for evolution of the TrpR biosensor to expand its functional capabilites.

RESULTS

Evolving the Trp biosensor for recognition of novel allosteric effectors

Allosteric biosensors are complex molecular machines that respond to signaling molecules by altering conformations upon binding (Popovych et al., 2009; Reichheld et al., 2009; Weaver et al., 2001). Evolving allosteric biomolecules is challenging due to the dynamic nature of binding modes, which can be difficult to fully capture during directed evolution (Doi and Yanagawa, 1999; Tang and Cirino, 2011; Taylor et al., 2015). We chose to alter the binding specificity of the TrpR to respond to halogenated tryptophan analogs, 5- or 6-bromo-L-tryptophan (5BrW and 6BrW, respectively), that have been implicated in the biosynthesis of neuropharmacological, anti-tumorogenic, and other pharmaceutical compounds (Bush et al., 1987; Craig et al., 1997; Efange et al., 1990). A number of TrpR mutations are known to cause non-allosteric binding (termed 'super-aporepressors') (Arvidson et al., 1993) – in order to maintain the dynamic response mechanism of the TrpR during directed evolution, we modified CPR for both positive (binding) and negative (non-binding) selection (Fig. 2.1). Due to the TrpR repressor system having a negative functional output (repression of signal), the positive CPR selection was performed using an inverter circuit by inhibiting expression of the λCI repressor regulating expression of the λ_{PR} promoter (Fig. 2.2a). By including the small molecule analog during the positive selection, and omission during negative selection, repressor variants are enriched that dynamically respond to the effector molecule.

Amino acid residues proximal to the 5- or 6-halogenation site were randomized using site saturation mutagenesis (NNS randomized) at positions V55, I57, V58, E59,

E60 and A50, I57, A77, T81, I82, respectively. Putative allosteric repressor variants were placed under the control of a medium strength inducible promoter P_{LacUV5} in the positive selection and constitutively expressed by promoter P_{CON} during the negative selection to mitigate leaky expression of Taq polymerase. Three cycles of positive CPR selection were performed – interspersed with four cycles of negative CPR selection (outlined in Material and Methods). Small scale screening of repressor variants from both libraries identified biosensors capable of responding to 5BrW and 6BrW, and as expected variants had lost specificity for L-tryptophan, as no binding signal was detected even on tryptophan rich 2xYT media. The 5BrW repressor ($5R_{WT}$) regulated gene expression from the P_{Trp+} promoter (a higher expression Trp promoter) over a 50-fold dynamic range with an EC_{50} of 17 μM and an EC_{50} of 79 μM for the inverted circuit (Fig. 2.2d). The CPR evolved 6BrW repressor ($6R_{WT}$) regulated gene expression over a 45-fold dynamic range with an EC_{50} of 53 μM and an EC_{50} of 171 μM for the inverted circuit. Given the tryptophan analogs 5BrW and 6BrW only vary structurally by a single position on the tryptophan heterocycle, non-specific repressor variants may have been enriched during selection. Surprisingly, small molecule cross-reactivity assays of the biosensor variants revealed both were fully orthogonal for their cognate targets (Fig 2.2c). Testing with another halogen substitution (5- or 6-chlorotryptophan) revealed the biosensors were not specific for bromine, which is unsurprising given the similar atomic properties of halogens (Fig 2.3).

CPR evolution of novel DNA operator:TrpR interactions

After small molecule binding, biosensors alter conformations to promote binding or dissociation from cognate DNA operator sequences. To determine if altered DNA binding specificities are compatible and as facile to evolve as the allosteric binding site, we next attempted to alter the DNA operator specificity of the TrpR biosensors. Upon small molecule effector binding and subsequent dimerization, the TrpR natively binds two palindromic operator sites near the -10 region of the promoter (Otwinowski et al., 1988). Since variation of the DNA operator site will influence promoter strength, we screened several candidate operator sequences and identified four additional operators that did not inactivate the promoter (Fig. 2.4). The identified operator sites (O_1, O_A, O_B, O_D) had higher expression than the native P_{TRP} promoter and were repressed negligibly by the wild-type TrpR. These operators were cloned into the promoter driving the expression of the λCI repressor for positive selection CPR. Randomization of residues directly contacting the promoter were made on the $5R_{WT}$ and $6R_{WT}$ scaffolds at positions K72, G78, I79, A80, T83 for operators O_1 and O_D and Q78, R69, K72, G78,I79, A80, T83 for operators O_A and O_B (Fig. 2.5a).

Three rounds of positive selection CPR were performed, which narrowed the library sufficiently for screening. Sequenced variants showed convergence towards specific amino acid motifs, presumably to allow binding to the new operator sites. For instance, evolution of the biosensors for O_D resulted in a perfectly matching DNA binding domain mutations K72S, G78C, I79C, A80L, and T83R. This commonality was not observed for each of the selections, O_A variants contained many relatively large

differences which may be explained by the 6-bromo repressor containing mutations directly adjacent to the DNA library residues – precluding certain amino acid combinations. Biosensor variants were tested for promoter repression with the 5BrW or 6BrW effector molecules (Fig 2.5b). On their cognate operator sequences, the variants (termed R subscript cognate operator; e.g. "R_1") displayed a range of repression activity, most repressing over 20-fold with addition of the appropriate effector small molecule. In addition, all of the DNA binding motifs from the 5-bromo selections could be transplanted on the wild-type binding pocket except for the motif for operator O_D, which was later selected independently using CPR. In general, the biosensors respond specifically to the their cognate effector. However, the operator variants from the 6BrW responsive biosensors displayed variable levels of activation with 5BrW. This may be due to the amino acid residues comprising the small molecule effector pocket and the amino acids involved in DNA recognition having overlap in the 6BrW biosensor libraries, but not the 5BrW libraries. To overcome this in the future the opposing small molecule could be included in a negative selection CPR step that would eliminate cross reactive variants.

Using CPR we functionally expanded a single starting legacy biosensor to contain fourteen novel phenotypes. To demonstrate their ability to specifically modulate gene expression based on the parameters they were evolved for, all 15 biosensors were tested for orthogonality for both DNA operator binding specificity, as well as, small molecule effector recognition (Fig. 2.5c). Across functional DNA and effector space, the repressors were modular despite not including a negative selection CPR step during their evolution.

Several notable exceptions were the R_A biosensor repressing the O_{WT} (~6-fold), $6R_1$ repression (~7-fold) with 5BrW on its cognate operator, and $6R_A$ repressing (~5-fold) with 5BrW on its cognate operator sequence. These off-target effects could likely be re-tuned, for instance by lowering effector concentrations below 1 mM. Nonetheless, this degree of orthogonality is surprising, especially as the set was evolved from a single evolutionary legacy part and only a handful of mutations are required to drastically change phenotype. However, to equip a single cell with multiple different biosensor phenotypes the molecular assembly of the repressor dimer poses a challenge to maintaining the desired logic functions.

Design of allosteric and intramolecular logic biosensors

CPR based directed evolution of orthogonal biosensors identified a suite of repressor variants, each with different molecular logic functionalities. However, coexpression of repressor subunits would likely result in the breakdown of desired logic – as assembly of the active dimer complex will occur randomly. This logic interference prevents higher order regulatory architectures from being constructed in a single chassis. To overcome this limitation, we developed a system for the allosteric and intramolecular (A.I.) logic of biosensors. Dimers are engineered for orthogonality by redesigning the dimer interface, as well as, covalent linkage (or "tethering") by expressing the two monomers as a fusion protein. The biosensor A.I. logic system, using the TrpR repressor as a scaffold, results in NAND based logic – as repression conditions for both monomers must be met for both half-sites to actively regulate promoter expression.

Initial efforts to develop the biosensor A.I. logic system involved reengineering the dimerization interface of the TrpR protein. The TrpR codons were randomized at positions W19, V23, H35, and L39, which have been implicated as key residues involved in the folding pathway of the TrpR dimer (Miño et al., 2013; Royer et al., 1993; Shao et al., 1997). Three rounds of positive selection CPR were performed to identify dimer regions that still resulted in an active repressor complex (Fig. 2.6). Screening variants from the resulting selection demonstrated that many possibilities exist as viable interfaces. Combinations of interfaces were constructed to guide the correct folding of dimers, with emphasis placed on creating steric clashes between non-cognate dimers. While we found that some interface regions could be programmed, this would be challenging without further mutating additional interface regions. In addition, we reasoned this approach is unlikely to be scalable – as each additional repressor interface would need to be orthogonal.

Based on the crystal structure of the TrpR dimer (PDB: 1RCS), we noticed the C-terminal amino acid was in proximity to the N-terminal of the second monomer. Circular permutation of the two monomers may promote intramolecular folding of proteins and provide a scalable solution to avoiding logic interference (Fig. 2.7). Repressors were tethered by creating a linker sequence between the two monomers (Material and Methods). This approach was combined with our best interface library variants to ensure that cross dimerization did not occur between tethered repressors. Additionally, the DNA sequences of two halves of the tethered biosensors were recoded to limit homology and possible recombination.

Biosensor A.I. logic was validated by using the $5R_{WT}$ and $6R_{WT}$ mutants against the wild-type operator site (O_{WT}). Monomers expressed bicistronically demonstrated logic interference by responding independently to both signaling effector molecules (Fig. 2.7b). The A.I. biosensor, displayed NAND logic – repressing over 50-fold only when both 5BrW and 6BrW were present in the growth media. This demonstrates both halves of the tryptophan repressor must be in a conformationally active state in order to achieve biologically relevant binding affinity to the operator sites and subsequent promoter repression.

Theoretically, A.I. biosensor logic should also extend to the DNA binding specificity of both halves of the tethered dimer. To demonstrate this, tethered combinations of repressors R_1 and R_B were tested for their ability to repress hybrid promoter sequences with operator sequences (O_1:O_1, O_B:O_B, and O_1:O_B). In agreement with the previous finding that A.I. logic was dependent on correct logic for both biosensor halves, the DNA binding variants must match both operator recognition sites in order to bind and represses promoter function (Fig. 2.7c). We found that there was variable repressor activities of the hybrid TrpR variants, ranging from high repression (>50-fold) to modest repression (~8-fold). However, this may due to the differences in promoter strength that resulted from hybrid DNA operator sequences. Despite this, each DNA based A.I. logic biosensor was specific for the given operator site pairs and required no additional engineering to achieve modularity.

DISCUSSION

Advantages of A.I. logic gated biosensors

Functional biosensors are of ever increasing interest, as accurate measurements of intramolecular or extracellular signals can be used for real time monitoring of metabolic flux or even the directed evolution of metabolic pathways (Raman et al., 2014; Rogers et al., 2015). This is especially valuable for measuring signals that do not have a visible or growth phenotype, or regulating genes that can dynamically respond to the contextual environment – such as degradation of a contaminant (Chen et al., 2014; Yoshida et al., 2016). Here, we demonstrate complex signaling behaviors can be developed, such as cellular logic based A.I. biosensors. Biosensors that perform logic-gated function intramolecularly, at the protein function level, may have specific advantages over traditional transcription regulated gene networks. In particular, transcriptional networks can be slow to respond – having to cascade a signal through multiple stages especially in low nutrient conditions (Hooshangi et al., 2005; Rosenfeld and Alon, 2003). Some systems have overcome the temporal barrier by relying on protein degradation instead of transcriptional activation which can respond more quickly (Prindle et al., 2014). However, the logic functions of A.I. biosensors respond entirely at the protein level, and signal output can be directly tied to DNA binding affinity (Alonso et al., 2015). The evolved TrpR components may also be useful for studying the *in vivo* dynamics of the TrpR regulatory system, as most studies have relied on amino acid depletion in minimal media which alters cell physiology due to starvation conditions and traces of L-tryptophan will always remain (Klig et al., 1988).

The evolutionary plasticity of a single legacy biosensor

The functional expansion of a single legacy biosensor, TrpR, to contain fifteen unique responses – binding to various small molecule signals and DNA operator sites – speaks to the evolutionary plasticity of regulatory elements. The malleability of this repressor protein demonstrates that organisms can evolve to adapt to a wide variety of environmental signals and mount unique and orthogonal responses. The creation of A.I. biosensors gives a plausible evolutionary path to higher order regulatory architectures through the recombination of gene duplication and recombination of biosensors (Teichmann and Babu, 2004).

MATERIALS AND METHODS

Validation of vectors for biosensor based positive and negative selection CPR

Biosensor based positive and negative selection vectors were constructed using overlapping DNA oligonucleotides (IDT) and standard cloning procedures. The pTrpR expression and library plasmids were constructed from a ColE1 ori, amp^R, and $LacI^Q$ expression plasmid under the control of either a P_{LacUV5} inducible promoter (used for positive selection CPR and functional assays) or a strong constitutive promoter P_{CON} (for negative selection CPR). The selection vectors pPOS and pNEG were constructed from a p15A ori, cam^R, and $LacI^Q$ expression plasmid (pACYC-solo). The pPOS plasmid uses the wild-type Trp promoter to drive expression of the λCI-LVA degradation tagged repressor. The promoter λ_{PR} drives expression of either GFP (for screening) or Taq polymerase (for positive selection CPR), thus creating an inverter circuit for Trp

biosensor function. The pNEG plasmid uses the wild-type Trp promoter to directly drive the expression of GFP (for screening) or Taq polymerase (for negative selection CPR).

Both selection plasmids (pPOS and pNEG) were functionally tested by expression of GFP under active or repressed conditions. To test this, P_{LacUV5} drove expression of either a functional TrpR or an inactive TrpR (truncated by two TAA stop codons and containing a unique internal HindIII restriction site; plasmid pTrpR.2TAA). To test positive or negative selection CPR, these control constructs were co-transformed with the Taq expressing pPOS or pNEG vectors. A single round of CPR selection was performed (detailed below) for both positive and negative selections using various ratios of active or inactive repressor (1:10 to 1:10,000). Under these conditions, positive selection should enrich for active TrpR repression, while negative selection should enrich for inactive TrpR repressors. Following selection, CPR derived amplicons were digested with HindIII for 2 hours to allow digestion of the internal HindIII site present in the inactive TrpR control vector (pTrpR.2TAA). The enrichment factor for positive selection was estimated to be roughly 300-fold per round of CPR, while the negative selection was estimated to be roughly 200-fold per round of CPR. These were calculated using the following formula:

$$\frac{X\ (initial)}{Y\ (initial)} * EF = \frac{X\ (final)}{Y\ (final)}$$

Where ($X_{initial}/Y_{initial}$) is the initial ratio of active or inactive repressor, EF is the enrichment factor, and (X_{final}/Y_{final}) is the ratio of active or inactive repressor - post-CPR selection.

CPR directed evolution of biosensors

TrpR repressor libraries were created through site-saturation mutagenesis using degenerate oligonucleotides with NNS or NDS codon randomization (IDT). Libraries were cloned into an ampicillin resistant; ColE1 origin plasmid that contained the IPTG inducible lacUV5 promoter. Libraries were electroporated into *E. coli* strain JW4356 (a tryptophan repressor knockout strain) pre-transformed with either the positive or negative selection vector. Initial library sizes were above the theoretical diversity threshold (~3.0 x 10^7) and were maintained throughout the CPR selection with a transformation efficiency of at least 10^6, but more typically 10^7-10^8. Overnight library cultures were seeded at a 1:20 ratio into fresh 2xYT media supplemented with 100 µg / mL ampicillin, 34 µg / mL chloramphenicol, and 1 mM of 5-bromo-L-tryptophan or 6-bromo-L-tryptophan, when necessary. Cells were grown for 1 hour at 37°C. Cells were subsequently induced by the addition of 1 mM IPTG and incubated at 37°C for an additional four hours.

Induced cells (200 µL total) were spun in a tabletop centrifuge at 3,000 x g for 8 minutes. The supernatant was discarded and the cell pellet was resuspended in CPR mix: 1x Taq buffer (10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgSO_4$), 260 µM dNTPs, 530 nM forward and reverse CPR selection primers. The resuspended cells were placed into a 2 mL tube with a 1mL rubber syringe plunger and 600 µL of oil mix (73% Tegosoft DEC, 7% AbilWE09 (Evonik), and 20% mineral oil (Sigma-Aldrich)). The emulsion was created by placing the cell and oil mix on a TissueLyser LT (Qiagen) with a program of 42Hz for 4 minutes. The emulsified cells were thermal-cycled with the program: 95°C - 3min, 25x (95°C - 30 sec, 55°C - 30 sec, 72°C - 1 min). Emulsions were

broken by spinning the reaction (10,000 x g - 5 min), removing the top oil phase, adding 150 µL of H_2O and 750 µL chloroform, vortexing vigorously, and finally phase separating in a phase lock tube (5Prime). The aqueous phase was cleaned using a PCR purification column which results in purified DNA, including PCR products as well as plasmid DNA. Subamplification with corresponding outnested recovery primers ensures that only PCR fragments are amplified. Typically, this is achieved by addition of 1/10 the total purified DNA using Accuprime Pfx (ThermoFisher) in a 20 cycle PCR, however challenging rounds of selection could require increasing the amount of template DNA or cycle number to achieve detectable amplification. Resulting DNA products were re-cloned into the selection vector - completing one full round of CPR selection.

Evolution and functional assays of allosteric repressors

Amino acid randomized libraries (NNS codon) were constructed by overlap extension PCR and cloned into the pTrpR.2TAA plasmid for positive selection. For the 5-bromotryptophan library, amino acids: V55, I57, V58, E59, and E60 were randomized. For the 6-bromotryptophan library, amino acids: A50, I57, A77, T81, and I82 were randomized. Libraries underwent positive selection CPR by co-transformation with the pPOS selection vector and supplemented with 1 mM of the corresponding tryptophan analog in the media. Subsequently, variants amplified from round 1 were cloned into the pTrpR.CON.2TAA vector, which utilizes a constitutive promoter to drive the repressor variants during the negative selection. The variants underwent negative selection CPR by co-transformation with the pNEG plasmid. Omitting the tryptophan analogs during the negative selection CPR counter-selects for variants which bind to the Trp promoter

without the analog, thereby repressing Taq expression. The entire selection process was completed in 7 total rounds, in the following order: positive (rounds 1, 4, and 7) and negative (rounds 2, 3, 5, 6).

Tryptophan analog responsive variants, $5BrR_{WT}$ and $6BrR_{WT}$, were transformed into JW4356 *E. coli* cells. Depending on direct or inverted circuit screening architecture, these were co-transformed with either pNEG+.GFP (expressing GFP with Trp+ promoter with enhanced expression) or pPOS.GFP for an inverted signal. Overnight cultures of individual colonies were seeded into a 96-well grow block containing 2xYT supplemented with ampicillin, chloramphenicol, and 1 mM IPTG at a 1:20 dilution ratio. Wells contained the indicated concentration of either 5-bromo-DL-tryptophan (Sigma) or 6-bromo-DL-tryptophan (Gold Biotechnology). Cells were incubated for 5 hours at 37°C to allow expression of repressors and GFP. Prior to fluorescence measurement on a plate reader (Tecan M200, excitation 469 nm; emission 501 nm), cells were centrifuged at 4°C for 20 minutes at 3,000 x g and resuspended in 1x PBS solution. Fluorescence intensities were measured and normalized to the OD_{600} of the culture. Non-fluorescent JW4356 cell fluorescence was subtracted from all samples, which was calculated by the fluorescence divided by OD_{600} for the parental JW4356 strain. Maximal signal output was calculated by transformation of a control vector - with wild-type TrpR (for inverter circuit), or inactivated TrpR (for repression circuit).

To calculate the half maximal effective concentration (EC_{50}) for repression, the dose-response function for either $5BrR_{WT}$ or $6BrR_{WT}$ was fit to the equation:

$$\frac{A}{1+\left(\frac{x}{c}\right)^{b}}$$

where (c) is the EC_{50} of 5- or 6-bromo-DL-tryptophan (mM).

To calculate the EC_{50} for activation (inverter circuit), the dose-response function for either $5BrR_{WT}$ or $6BrR_{WT}$ was fit to the equation:

$$\frac{a*x}{b+x}$$

where (b) is the EC_{50} of 5- or 6-bromo-DL-tryptophan (mM).

Screening of Trp promoters with novel operators

Novel operator sites were designed by mutating palindromic operator sites at conserved binding residues except for the +1 and +2 positions of the (CTAGTAC) wild-type operator sequence. These residues were avoided because they form part of the core of the Trp promoter's -10 region (Pribnow box) and would likely influence promoter activity. Mutant operators focused heavily on mutation of the +3, +4, +5, and +7 positions of the operator due to the phylogenetic conservation of sequences at these positions and the clear contacts that are made with the repressor itself. These mutant operators were placed into the wild-type Trp promoter and tested for activity. GFP expression with a truncated TrpR (2TAA) repressor indicated the total promoter strength of the mutant operators driving the expression of GFP. Operators were also tested for their orthogonality to the wild-type repressor by co-testing with the TrpR under the expression of a high strength tac1 promoter. Candidate operator sites were chosen based on promoter activity and orthogonality to the wild-type repressor.

Evolution and functional assays of novel operator binding

Amino acid randomized libraries (NDS codons) were built on top of the $5BrR_{WT}$ and $6BrR_{WT}$ scaffold sequences. For operator 1 and operator D (O_1 or O_D) libraries, positions K72, G78, I79, A80, and T83 were NDS randomized. For operator A and operator B (O_A or O_B), two libraries were used. The first library was NDS codon randomized at positions Q68, R69, K72, I79, A80, and T83. The second was NDS codon randomized at positions Q68, R69, G78, I79, A80, and T83 with K72S mutation.

The positive selection vector, pPOS, was modified so the Trp promoter driving expression of the λCI repressor contained mutant palindromic operator sites. TrpR variants that most effectively bind the novel DNA operator sequences will down-regulate the λCI, which subsequently increases expression of Taq polymerase. Positive CPR selection was utilized using the previously described methods, and contained 1 mM of either 5BrW or 6BrW. The libraries underwent three (for O_1 and O_D) or four (for O_A and O_B) rounds of positive CPR selection prior to functional screening. To create the L-tryptophan binding repressor with altered DNA operators, the corresponding 5BrR variant mutations were placed into the wild-type binding pocket. This created highly functional repressors, except in the case of O_D. To obtain this variant, an independent CPR selection was done on the wild-type scaffold with K72, I79, A80, and T83 fully (NNS) randomized. After three rounds an O_D binding variant was obtained.

Plasmids with variant repressors were transformed in *E. coli* JW4356 along with plasmids encoding the Trp promoter driving GFP, regulated by the cognate operator sequences. Overnight cultures were seeded into a 96-well grow block containing 2xYT

supplemented with ampicillin, chloramphenicol, and 1 mM IPTG at a 1:20 dilution ratio. If indicated, 1 mM 5BrW or 6BrW was added to the growth media. Cells were grown for 6 hours at 37°C and subsequently processed and measured as previously described. Non-fluorescent JW4356 cell fluorescence from non-GFP expressing JW4356 cells was subtracted from experimental samples. Maximal promoter output was calculated by a control experiment transformed with inactive TrpR (TrpR.2TAA) that expresses a premature stop codon in the coding sequence.

The orthogonality of all repressor and operator combinations was performed as described above. All fifteen repressors were transformed into cells containing each of the five operator (O_{WT}, O_1, O_A, O_B, or O_D) P_{Trp} promoters. These were either grown in conditions with 2xYT media (high concentration tryptophan media), 1 mM 5-bromo-DL-tryptophan, or 1 mM 6-bromo-DL-tryptophan. Fold repression was calculated by dividing the fluorescence of the maximal promoter strength by the fluorescence of the experimental condition.

Construction of biosensor driven allosteric and intramolecular logic

The TrpR interface was engineered by randomization (NNS) of positions between the interface of the dimers, followed by three rounds of positive CPR selection. The positions randomized were: W19, V23, H35, and L39. In the wild-type repressor, positions W19 and V23 of the first monomer directly interact with positions H35 and L39 of the second monomer. Between the dimer, two regions comprise the interface which theoretically can be engineered to prevent monomers from self-dimerization. Screening of TrpR interface variants and sequencing revealed many viable combinations of amino

acids that result in an active repressor. One variant, Int.3 (W19L, V23M, H35D, and L39W), was of particular interest because the large hydrophobic tryptophan residue (which has been shown to be an essential component of the folding pathway of the dimer) had translocated orientations, to the opposite side of the interface. Monomers could be programmed with the two tryptophan residues sterically clashing between mismatched monomers, thus preventing folding and dimerization. This resulted in the following combinations of amino acids W19, V23, D35, W39 for the first monomer and L19, M23, H35 and L39 for the second.

TrpR repressors were tethered by fusion of the two monomers using circular permutation. The last aspartate residue (D108) of the first monomer was deleted and the N-terminal fifteen amino acids of the second monomer was also deleted. The deletion of residues was performed to prevent (1) inadvertent translation initiation at residue M11 and (2) long flexible linkers can be targets for proteolysis. Either of these would result in non-tethered repressors, that could potentially disassemble and reassemble with other repressor monomers. In addition to the deletion of these residues, screening of a small (4 amino acid) linker was performed to identify residues that may promote intramolecular folding of the two monomers. Linker libraries contained 4 NNS codons and were screened for functional repressor activity. The linker region His, Arg, Phe, Asn was used for all fusion proteins.

Functional testing of tethered $6BrR_{WT}$ and $5BrR_{WT}$ repressor constructs was carried out by a GFP assay using repression of the wild-type Trp promoter. Cells were induced for 7 hours with 5 mM of 5- or 6-bromo-DL-tryptophan and 1 mM IPTG. Cells

were handled as described previously before fluorescence measurements. For DNA operator tethering, synthetic Trp promoters were created by combinations of the operators O_1 and O_B to create (O_1:O_1, O_B:O_B, and O_1:O_B). These were co-transformed with tethered DNA repressor variants: R_1-R_1, R_B-R_B, or R_1-R_B. Cells were induced with IPTG for 5 hours prior to fluorescence measurements.

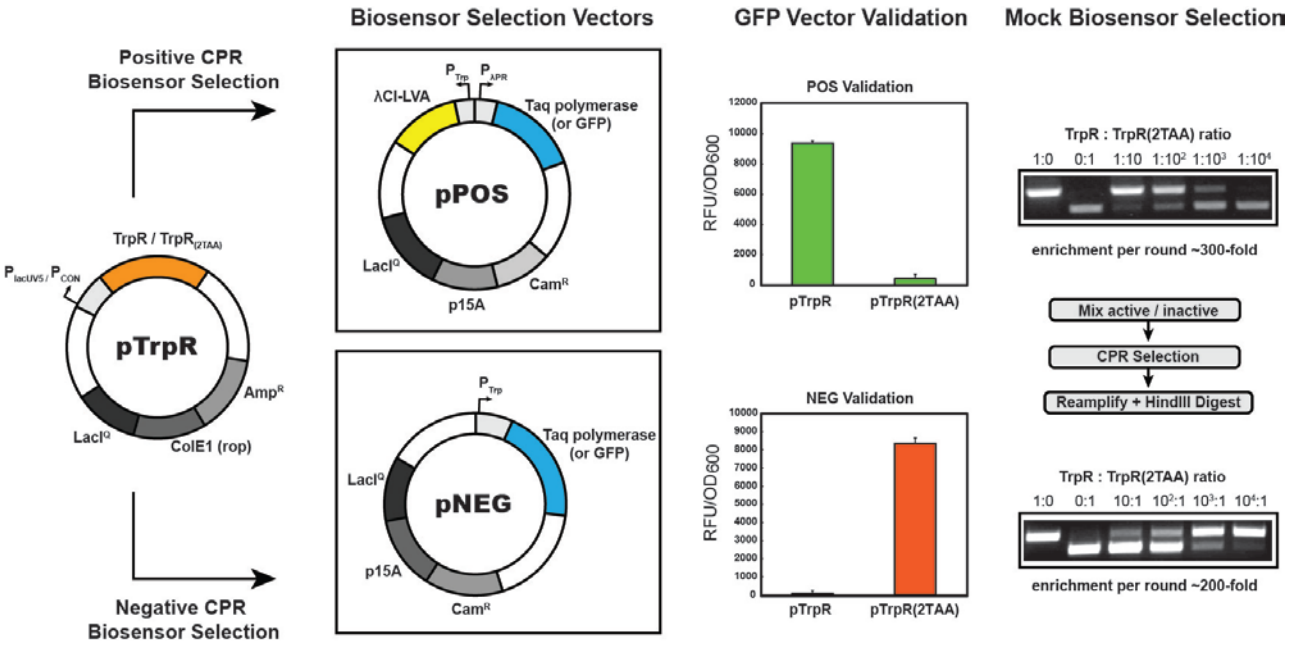


Figure 2.1 : Validation of positive and negative CPR for biosensor evolution.

Circuits for the evolution of biosensors are outlined above. The pTrpR plasmid contains an active or inactive version (2TAA) of wild-type tryptophan repressor (TrpR). This plasmid is co-transformed with biosensor selection vectors, pPOS (for positive selection CPR) or pNEG (for negative selection CPR). Selection vectors were validated with GFP substituted for Taq polymerase, and assayed for fluorescence with an active or inactive TrpR. Mock CPR selections were carried out by mixing active and inactive repressors (at the ratio indicated) and performing a single positive or negative CPR round. Resulting products were size separated and final post-selection ratios were estimated to yield an enrichment factor per single round of CPR.

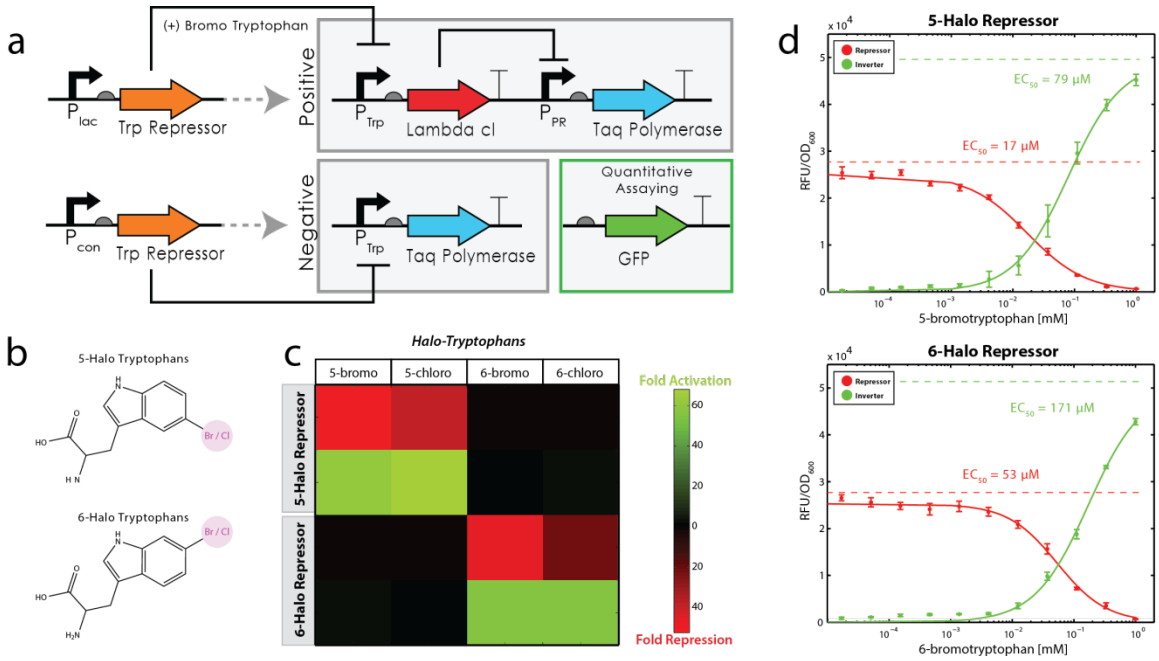


Figure 2.2 : Selection and characterization of a 5- and 6-bromotryptophan repressor.

a, Halogenated tryptophan responsive repressors were selected by performing rounds of biosensor positive or negative selection CPR. In positive selection CPR, the desired small molecule is added to the media and biosensors that actively suppress λCI expression are enriched. In negative selection CPR, the desired small molecule is omitted from the media and biosensors that are non-specific will inhibit Taq polymerase production. **b**, Molecular structure of 5-halo-tryptophan and 6-halo-tryptophan. **c**, Evolved biosensors for 5-halogenated or 6-halogenated tryptophan are orthogonal, however will respond to bromo or chloro modifications indiscriminately. **d**, Dose-response functions of the 5-halo and 6-halo biosensors with either 5-bromo-DL-tryptophan or 6-bromo-DL-tryptophan, in the direct or inverted circuit architecture.

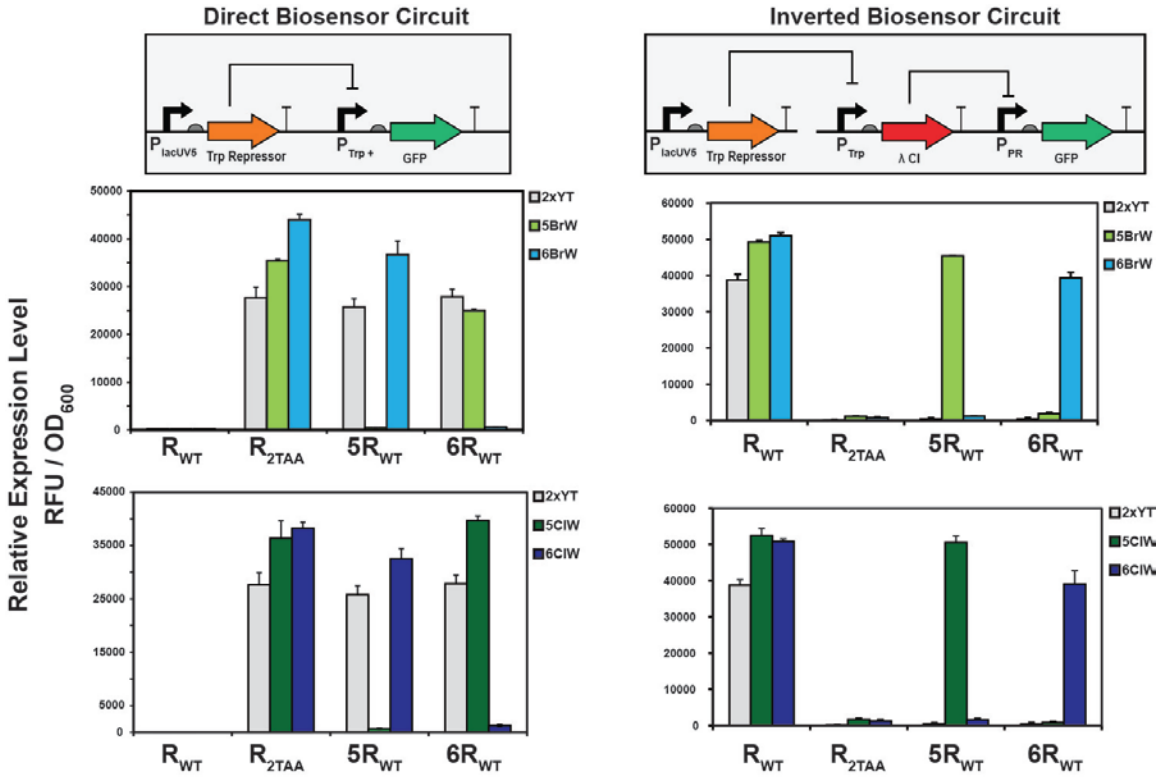


Figure 2.3 : Characterization of evolved biosensors repression and cross-reactivity.

The wild-type TrpR (R_{WT}), inactive TrpR (R_{2TAA}), 5-halo ($5R_{WT}$), and 6-halo ($6R_{WT}$) repressors were characterized for *in vivo* repression with tryptophan (2xYT), 5-bromo (5BrW), 5-chloro (5ClW), 6-bromo (6BrW), or 6-chloro (6ClW) tryptophan analogs (at 1 mM final concentration). Repression efficiency was either measured in a direct biosensor circuit or inverted biosensor circuit architecture by measurement of the relative expression level of GFP.

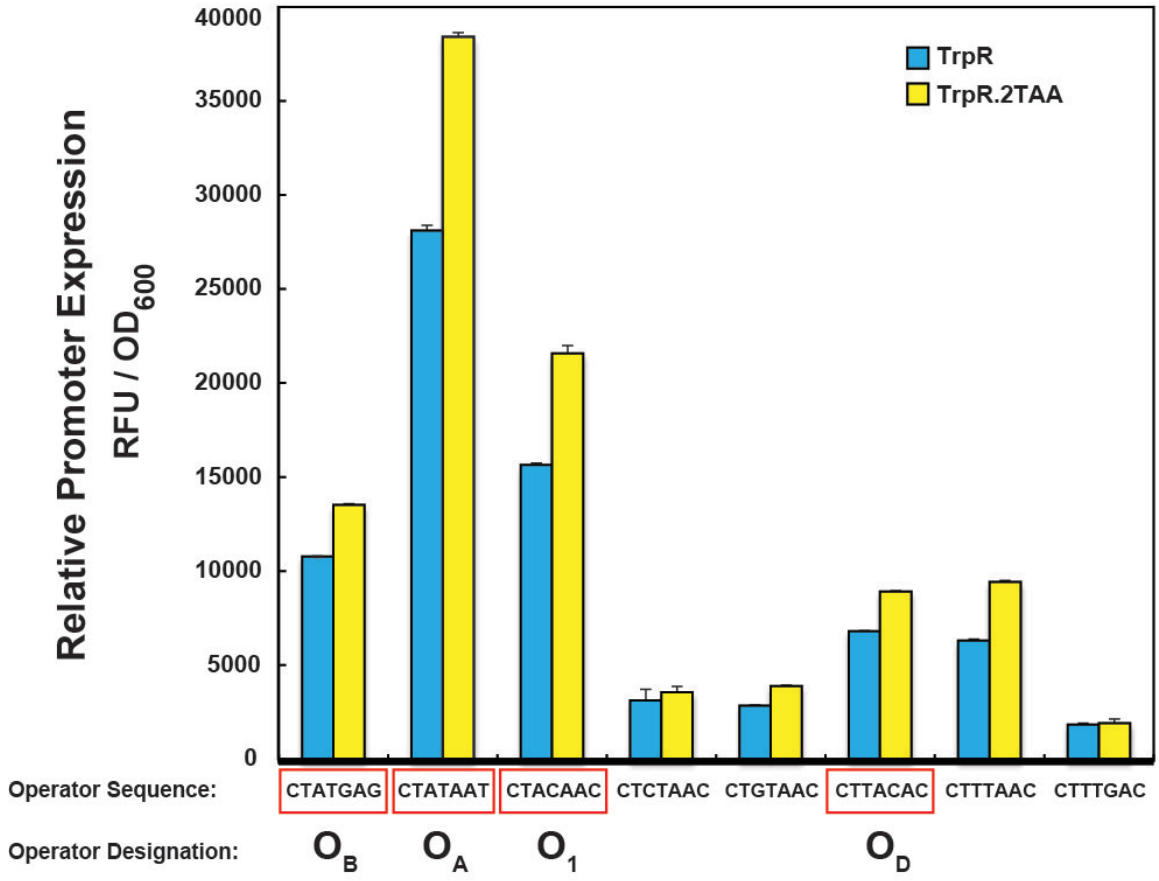


Figure 2.4 : Identification of novel operator sites that are active and orthogonal.

Potential orthogonal operator sites were designed by mutating conserved residues and residues with obvious contacts to the TrpR protein. Additionally, the +1 and +2 sites of the operator were not changed since they constitute part of the Pribnow box. Candidate operator sites were cloned into the pNEG vector construct and tested for promoter expression of GFP. These were tested with both active wild-type and inactive repressors, to identify operators that resulted in strong promoters and were relatively orthogonal to the wild-type TrpR.

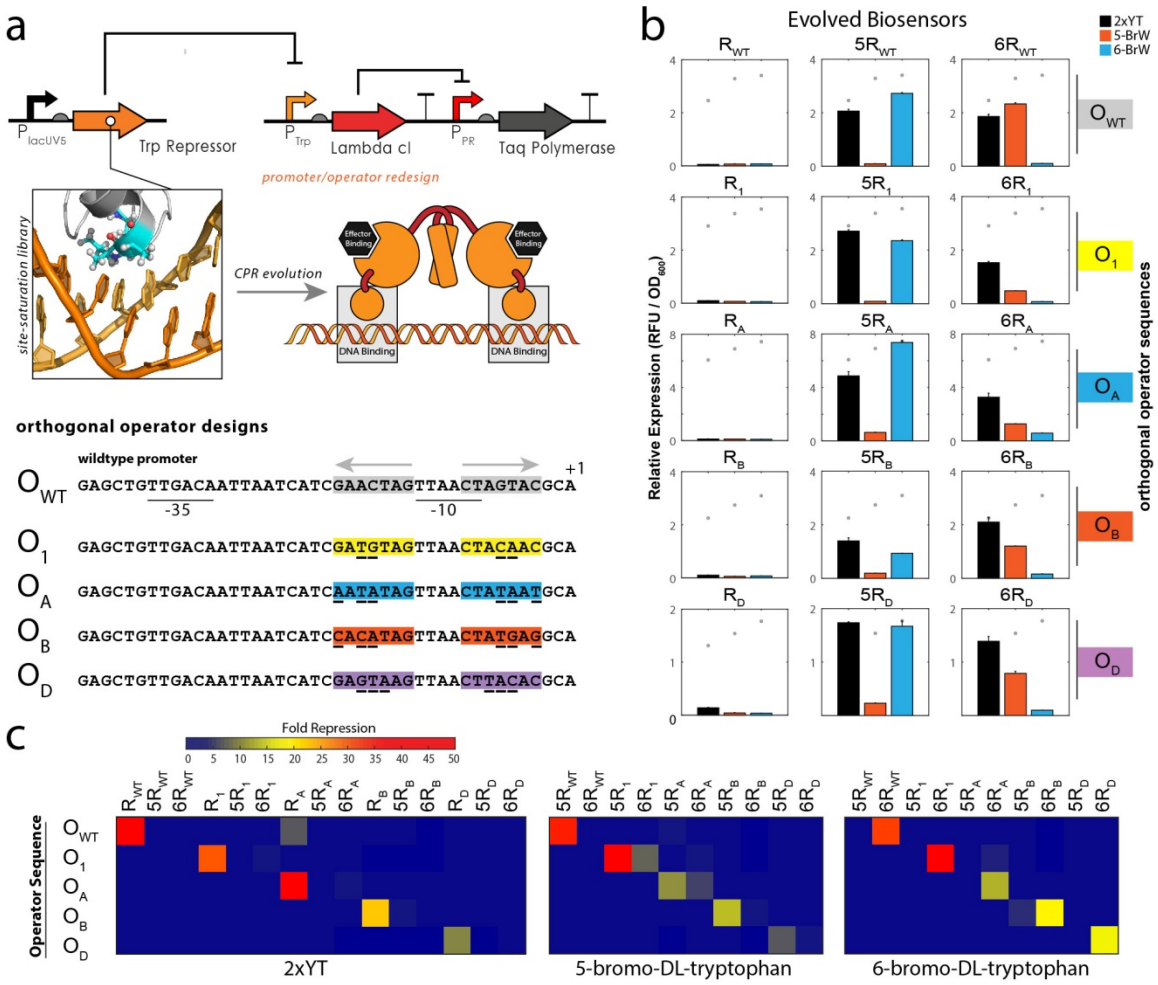


Figure 2.5 : Evolution and characterization of biosensors responsive to novel operators.

a, CPR scheme for the directed evolution of novel operator recognition and repression. Orthogonal operator sites (O_1, O_A, O_B, O_D) are inserted into the positive selection CPR circuit by swapping the promoter driving the expression of the λCI protein. Biosensors capable of binding the operator sites and blocking transcription initiation will produce Taq polymerase more effectively. **b**, Evolved biosensor variants (annotated R_1 for the tryptphan responsive operator 1 binding variant, $5R_1$ for the 5-bromo-tryptophan operator 1 binding variant, etc.) were tested for GFP repression activity on their cognate operator sites with 2xYT, 1 mM 5-bromo-tryptophan, or 1 mM 6-bromo-tryptophan. **c**, All fifteen biosensors were tested for against each operator sequence and tryptophan analog condition. A heatmap of repression activity was generated by calculating the fold repression for each biosensor in each experimental condition.

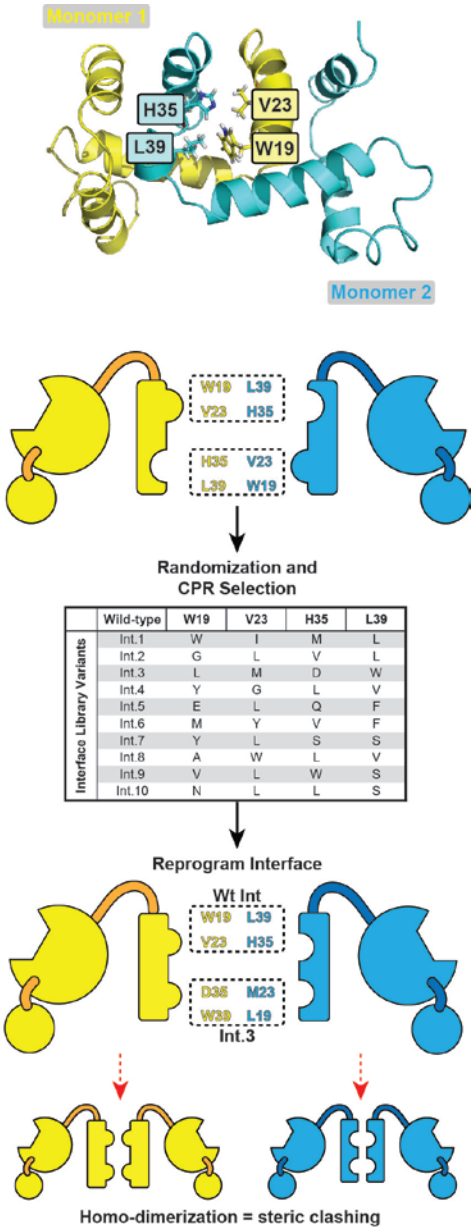


Figure 2.6 : Design and selection of a dimerization interface on the TrpR

In the top panel, a structural model of the TrpR dimer is shown. For clarity, one monomer is labeled yellow and the other blue. Residues involved in a critical region of the dimer are labeled. These residues were randomized and underwent three rounds of positive CPR selection. Novel interface residues were identified by their ability to repress the Trp promoter. Interface residues were reprogrammed to result in steric clashing between homo-dimers, to promoter hetero-dimer formation.

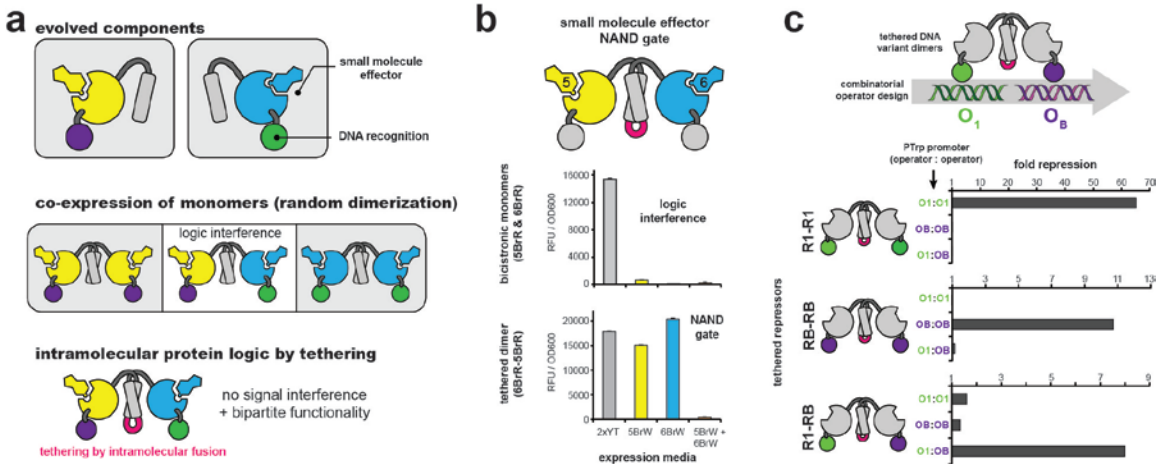


Figure 2.7 : Design of A.I. logic gated biosensors

a, A design method for the higher order assembly of genetic architecture with evolved biosensor components. Coexpression of evolved monomers results in logic interference due to random dimerization and subsequent hybrid phenotypes. Tethering of repressor monomers by intramolecular fusion ensures correct folding of biosensors, and prevents logic interference. **b**, A.I. logic enables protein-level NAND effector logic. Fusion of the $6R_{WT}$ and $5R_{WT}$ biosensors creates a NAND gate, requiring both 5-bromo-tryptophan and 6-bromo-tryptophan to repress expression from the wild-type Trp promoter. **c**, A.I. logic enables modular swapping of DNA recognition domains. Combinations of R_1 and R_B biosensor variants were tested for repression from combinations of O_1 and O_B operator sequences.

Chapter 3: Synthetic Evolutionary Origin of a Proofreading Reverse Transcriptase

While proteins involved in fundamental biological processes are under extraordinary functional constraints, recent work has shown that core components in the central dogma can be engineered to have radically altered properties, such as the creation of expanded genetic alphabets, tethered ribosomes, and rewiring of the genetic code (Lajoie et al., 2013; Malyshev et al., 2014; Orelle et al., 2015). One of the last discovered components of the central dogma, reverse transcriptase (RT), has for the most part been the province of a single family of enzymes of ancient evolutionary origin and are inherently error prone due to their lack of a proofreading (3'- 5' exonuclease) domain. To determine if the lack of proofreading is a historical coincidence or a functional limitation of reverse transcription, we attempted to evolve a high fidelity, thermostable DNA polymerase to efficiently use RNA templates. The Reverse Transcription Xenopolymerase (RTX) is a new, distinct evolutionary lineage that maintains the DNA proofreading activity of its parent with a RNA template, which was shown to greatly improve RT fidelity. Mutations in highly conserved molecular checkpoints render RTX capable of polymerizing either DNA or RNA templates with similar efficiencies, enabling applications such as single enzyme RT-PCR and direct RNA sequencing without cDNA isolation. The creation of RTX confirms proofreading is compatible with reverse transcription and reemphasizes that core molecular machinery that has otherwise been conserved over billions of years can be engineered to attain novel functionality.

INTRODUCTION

The molecular basis for all life rests on the information flow between DNA, RNA, and proteins (Crick, 1970). Early notions of how information was processed was amended after Temin and Baltimore's discovery of the reverse transcriptase (RT) enzyme (Baltimore, 1970; Temin and Mizutani, 1970) - clashing with the generally accepted view of a unidirectional central dogma. Evolutionary studies have elucidated that the RT family has a single ancient evolutionary origin which is supported by conserved amino acid homology and the ubiquitous presence of RT across multiple domains of life (Xiong and Eickbush, 1990). RT is hypothesized to be the catalyst in the transition of the RNA to DNA world by providing an avenue to copy RNA into more stable DNA genomes (Darnell and Doolittle, 1986), but has since found a variety of functional roles in biology including: telomere addition, mitochondrial plasmid replication, transposition, and the proliferation of retroviral genomes (Boeke and Stoye, 1997).

The progenitor of RT is postulated to be an RNA dependent RNA polymerase. Since RNA polymerases generally lack an error-checking 3'-5' exonuclease domain (Nakamura, 1997; Xiong and Eickbush, 1990), proofreading activity is also not present across the RT family, resulting in low fidelity reverse transcription and characteristic quasispecies behavior in organisms that rely upon it for replication (Meyerhans et al., 1989). In contrast to reverse transcriptases, other DNA polymerase families have evolved exquisite proofreading mechanisms to increase DNA synthesis fidelity during genome replication (Kunkel and Bebenek, 2000). Given this dichotomy, it is unclear whether the

lack of proofreading by RTs is due to historical coincidence or whether reverse transcription is mechanistically incompatible with proofreading.

In order to determine whether the evolutionary divide between RTs and DNA polymerases is a matter of history or function, we have attempted to directly evolve a reverse transcription xenopolymerase (RTX; Fig 3.1) from an error correcting DNA polymerase using a modified directed evolution strategy (Ghadessy et al., 2001), called reverse transcription compartmentalized self replication (RT-CSR) (Fig 3.1). RT-CSR enables the simultaneous screening of up to 10^9 polymerase variants for reverse transcriptase activity. Briefly, E. coli cells containing a library of putative RTX variants are physically compartmentalized through a water-in-oil emulsion and subsequently PCR cycled. Primers are included in the emulsion to facilitate self-replication, but contain RNA residues which behave as a template in subsequent cycles of PCR to enforce reverse transcriptase activity.

RESULTS

RT-CSR and Evolution of the Reverse Transcriptase

We chose the Archaeal Family-B DNA polymerases (polB) as a starting point for directed evolution of the RTX as they are monomeric, hyper-thermostable, highly processive, and contain robust proofreading domains – prized tools for modern PCR applications. Previous studies have attempted to rationally design these enzymes to utilize RNA templates with limited success (Jozwiakowski and Connolly, 2011; Moser et al., 2012), and initial experiments confirmed that two common polB enzymes from P. furiosus and T. kodakarensis (KOD) (Lundberg et al., 1991; Takagi et al., 1997) failed to

polymerize across even five template RNA bases. Initial modeling to identify mutations enabling RT activity was considered, but deemed impractical given the extensive contacts these polymerases make with the template (over 50 direct interactions). We initiated evolution using low stringency RT-CSR (10 RNA residues) with a purely random library (1-2 amino acid mutations per gene) of KOD polymerase variants. As polymerases were enriched, we gradually increased the stringency of RT-CSR by the stepwise addition of RNA into primers. By cycle 18, maximum stringency was achieved as primers were entirely composed of RNA - requiring reverse transcription of 176 residues to occur every thermal-cycle in order to maintain exponential amplification in the emulsion PCR.

Following RT-CSR, profiling of variants revealed extensive mutations throughout the polymerases, as over one thousand PCR cycles were performed since the initial library was introduced. One particular variant, B11, contained 37 mutations. As expected RT-CSR enriched for RT activity and B11 was capable of reverse transcription of at least 500 base pairs, however sequencing and testing confirmed inactivation of the proofreading domain (Fig. 3.2). Kinetic analyses established B11 utilizes both DNA and RNA templates with similar efficiencies by greatly lowering the Km on RNA:DNA heteroduplex. We attempted to restore proofreading by transplantation of the wild-type 3'-5' exonuclease which reactivated proofreading capabilities, albeit to barely detectable levels. Encouraged that minimizing extraneous mutations intrinsically derived from the RT-CSR process could restore proofreading, we sought to design polymerases with a minimal core set of mutations.

Identification of Mutations involved in Reverse Transcriptase Activity

To understand how our process reshaped KOD polymerase to utilize RNA templates, we deep sequenced RT-CSR cycles to recapitulate the evolutionary path to RT activity (Fig. 3.3 and Table 3.1). Mutations were identified throughout the template-polymerase interface, but interestingly conserved mutations accumulated precisely as the polymerase encounters residues in the template strand. Polymerases contained mutations that were hypothesized to be molecular checkpoints used to enforce strict DNA template utilization: (1) as the template enters, (2) near the active site, and (3) at the nascent duplex. Given the likely importance of these regions, we used computer modeling to determine the molecular basis for RNA utilization.

The first evolved mutation localized near the template entry site of the polymerase at position R97 (Fig. 3.3). Proximal to this site, native polB scans for uracils (typically caused by cytosine deamination) by flipping template bases into a specialized pocket to halt polymerization until the mutation can be corrected by repair machinery (Fogg et al., 2002; Greagg et al., 1999). Evolved polymerases contained a variety of amino acid mutations to R97, all of which destabilize a salt bridge to the phosphate backbone that presumably regulates base flipping into the pocket.

As template residues near the active site, they encounter the most conserved RT-CSR mutation, Y384H, which prevents Y384 and Y494 from hydrogen bonding to the 2' hydroxyl of template RNA by reorganizing a hydrogen bonding network. Post polymerization, in the thumb domain, the most prevalent mutations (E664K, G711V, and E735K) promote tighter homo and hetero duplex binding in both A and B-form

conformations. Previous studies support this hypothesis and have shown that the E664K mutation alone causes markedly increased binding to RNA:DNA heteroduplexes (Cozens et al., 2012). To further validate we established an optimized set of mutations, we fully randomized (NNS library) several positions and repeated RT-CSR. In support of our modeling, many amino acids solutions were viable at position R97, but other positions (namely, Y384 and E664) had strong preferences for particular amino acids.

Design of the First Proofreading Reverse Transcriptase

Modeling driven designs of several polymerases with favorable RT mutations were synthesized and empirically tested. Our most promising RTX contained less than half the mutations of B11, without sacrificing catalytic efficiency or Km on RNA (Fig. 3.4). Mutations in RTX were demonstrated to not negatively affect desirable properties of parental KOD polymerase. Thermostability was maintained with optimal RT occurring ~70°C and consequently RTX was capable of single enzyme RT-PCR (in which RTX performs both the first-strand RT synthesis and PCR amplification). Across several RNA samples and gene loci, RTX demonstrated remarkable processivity on RNA templates, performing RT-PCR on RNAs over 5 kilobases in length (Fig. 3.5).

RTX is a Proofreads on RNA and DNA Templates

We next tested if RTX retained the proofreading activities of the parental polymerase. Initial testing using dideoxy mismatch primers in PCR demonstrated robust proofreading activity on DNA template (Fig. 3.6), but it was unclear whether the proofreading mechanism was compatible during reverse transcription since RNA:DNA heteroduplexes can adopt A-form helical structures (Wang et al., 1982). Primer extension

with a canonical matched base pair or a 3' deoxy mismatched pair (preventing extension until terminator excision) were tested. As expected, both wild-type KOD and RTX were capable of extending mismatched primers on DNA templates, unlike exonuclease deficient mutants. When tested on a RNA template, KOD's exonuclease was stimulated - which actively degraded the priming oligonucleotide, while RTX could extend the mismatched primer with activity indistinguishable from DNA templated proofreading (Fig. 3.7).

Given RTX is the first polymerase capable of proofreading during reverse transcription, we hypothesized RTX may have increased RT fidelity compared to natural polymerases. Barcoded primers utilized during RT of several human mRNAs allowed multiple reads of a single cDNA during deep sequencing - reducing background sequencing errors by several orders of magnitude (Schmitt et al., 2012). Sequencing analyses revealed the control retroviral RT (MMLV) had an error rate between $4.8x10^{-4}$ to $1.1x10^{-4}$ while RTX had an error rate of $3.5x10^{-5}$ to $3.7x10^{-5}$ (3 to 10 fold lower) (Fig. 3.7). The mutational spectra of RTX favored G to A transitions and G to T transversions, which accounted for nearly half the observed mutations. Inactivating the RTX's proofreading capabilities increased error frequency nearly 3-fold, supporting active proofreading was occurring during RT. Interestingly, inactivating the proofreading of RTX shifted the mutational bias - perhaps due to preferential editing of particular mismatches (Fig. 3.7 and Table 3.2). Given the barcoding error detection limit is identical to the observed error of RTX (Schmitt et al., 2012) (confirmed by measuring fidelity of

wild-type KOD on DNA templates (Table 3.2)), we anticipate the true error rate for RTX to be even lower than reported.

RTX in Nextgen RNAseq Workflows

Due to its unique properties, RTX has great potential to streamline workflow (combining RT and PCR steps) and increase the precision of transcriptomics, reducing biases and errors introduced in the reverse transcription step of RNA-Seq protocols (Ozsolak and Milos, 2011). To demonstrate the immediate utility, we implemented RTX into a commonly used platform for directional RNA sequencing. Analysis revealed nearly identical coverage and expression profiles, suggesting that RTX is compatible with established workflows. In addition, we developed a more streamlined protocol to directly sequence RNA. Using a traditional Sanger sequencing approach (Sanger et al., 1977), a $GATC_5$ RNA repeat was directly sequenced (Fig. 3.8). Direct RNA sequencing should be adaptable to single molecule sequencing platforms, enabling high throughput and high fidelity sequencing of complex RNA samples by eliminating the biases created in cDNA synthesis and subsequent amplification.

Beyond applications in sequencing, the surprising flexibility of the RTX lineage may presage the ability to utilize entirely new chemistries in genetics. Primer extension reactions were performed on a ribose sugar analog (2' O-methyl DNA) that indicated that reverse transcription of RTX polymerases could extend alternative templates but with much lower efficiency, indicating a preference for RNA substrates (Fig. 3.9). However, RTX was still far more efficient at utilizing 2’-OMe DNA than the parental wild-type,

opening the door to further optimization and due to 2'-OMe stability, potential therapeutic applications.

DISCUSSION

RTX as a Second Origin of Reverse Transcription

The directed evolution of RTX marks a second origin for RNA reverse transcriptase function fundamentally distinct from the retroelement lineage. Using the RT-CSR process, the substrate specificity of a high fidelity DNA polymerase was fundamentally altered - highlighting the remarkable plasticity of highly conserved molecular machinery. Ostensibly, the mutations found unlocked molecular checkpoints in the discrimination of DNA and RNA, but surprisingly did not disrupt the proofreading capabilities of the polymerase. This was unexpected especially given RNA:DNA hybrid duplexes often form A-helical structures unlike DNA:DNA duplexes, and may provide insights into the transition from polymerization to editing modes of the polymerase. We found that reverse transcription proofreading increased the fidelity of cDNA synthesis above those found in natural retroviral RTs, which enabled higher precision RNA-Seq.

RTX and the Implications for the Origin of Life and Future Reverse Transcription

Only a handful of mutations were required to impart RT activity, suggesting the evolutionary hurdle for high fidelity reverse transcription is relatively low. Despite this, all retroelements utilize proofreading deficient RTs suggesting that high error rates are either a historical coincidence or an evolutionary strategy to promote diversity. Another possible explanation is that high fidelity was never required simply because RNA

genomes are small due to their inherent instability (Huff et al., 1964). Given the plasticity of these polymerases for modified templates and the adaptability of the RT-CSR framework (as primers are simply programmed to contain modified bases), RTX evolution should be compatible with many base and sugar analogs (Malyshev et al., 2012; Schoning, 2000; Yamashige et al., 2012; Yang et al., 2011). Combination with already evolved XNA polymerases could enable synthesis of genomes entirely composed of artificial nucleic acids (Pinheiro et al., 2012). Further integration with other synthetic machinery such as expanded genetic codes or engineered ribosomes will begin to further increase the gap between synthetic and natural life.

MATERIALS AND METHODS

Initial reverse transcription test for polymerases

30 pmol of 5' fluorescein labeled primer (25FAM) were annealed with 30 pmol of template and 0.4 µg of polymerase by heat denaturation at 90°C for 1 minute and allowing to cool to room temperature. Reactions were initiated by the addition of "start" mix which contained (50 mM Tris-HCl (pH8.4), 10 mM $(NH_4)2SO_4$, 10 mM KCl, 2 mM $MgSO_4$) and 200 µM dNTPs. MMLV polymerase was treated according to manufacturers recommendations (New England Biolabs). Reactions were incubated for 2 minutes at 68°C until terminated by the addition of EDTA to a final concentration of 25 mM. The labeled primer was removed from the template strand by heating sample at 75°C for 5 minutes in 1x dye (47.5% formamide, 0.01% SDS) and 1 nmol of unlabeled BLOCKER oligonucleotide (to competitively bind the template strand). Samples were run on a 20% (7 M urea) acrylamide gel.

Reverse Transcription CSR (RT-CSR)

KOD polymerase libraries were created through error prone PCR (unless otherwise indicated) to have a mutation rate of ~1-2 amino acid mutations per gene. Libraries were cloned into tetracycline inducible vector and electroporated into DH10B E. coli. Library sizes were maintained with a transformation efficiency of at least 10^6, but more typically 10^7-10^8. Overnight library cultures were seeded at a 1:20 ratio into fresh 2xYT media supplemented with 100 µg / mL ampicillin and grown for 1 hour at 37°C. Cells were subsequently induced by the addition of anhydrotetracycline (typically at a final concentration of 200 ng / mL) and incubated at 37°C for 4 hours. Induced cells (200 µL total) were spun in a tabletop centrifuge at 3,000 x g for 8 minutes. The supernatant was discarded and the cell pellet was resuspended in RTCSR mix: 1x Selection buffer (50 mM Tris-HCl (pH8.4), 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$), 260 µM dNTPs, 530 nM forward and reverse RNA containing primers. The resuspended cells were placed into a 2 mL tube with a 1mL rubber syringe plunger and 600 µL of oil mix (73% Tegosoft DEC, 7% AbilWE09 (Evonik), and 20% mineral oil (Sigma-Aldrich)). The emulsion was created by placing the cell and oil mix on a TissueLyser LT (Qiagen) with a program of 42Hz for 4 minutes. The emulsified cells were thermal-cycled with the program: 95°C - 3min, 20x (95°C- 30 sec, 62°C- 30 sec, 68°C- 2 min). Emulsions were broken by spinning the reaction (10,000 x g - 5 min), removing the top oil phase, adding 150 µL of H2O and 750 µL chloroform, vortexing vigorously, and finally phase separating in a phase lock tube (5Prime). The aqueous phase was cleaned using a PCR purification column which results in purified DNA, including PCR products as well as

plasmid DNA. Subamplification with corresponding outnested recovery primers ensures that only polymerases that reverse transcribed are PCR amplified. Typically this is achieved by addition of 1/10 the total purified emulsion using Accuprime Pfx (ThermoFisher) in a 20 cycle PCR, however challenging rounds of selection could require increasing the input DNA or cycle number to achieve desired amplification.

Molecular Modeling of RT-CSR Mutations

The wild-type structures of the family B polymerases from Thermococcus kodakaraensis (KOD) (PDB: 4K8Z) and Thermococcus gorgonarius (PDB: 2VWJ) were prepared for mutational analyses using the Molecular Operating Environment (MOE.10.2015) software package from Chemical Computing Group. The structures were inspected for anomalies and protonated/charged with the Protonate3D subroutine (341K, pH 8.0, 0.1 M salt). The protonated structures were then lightly tethered to reduce significant deviation from the empirically determined coordinates and minimized using the Amber12:EHT forcefield with Born solvation model to an RMS gradient of 0.1 kcal mol^{-1} $Å^{-1}$. These structures were then used as templates to build homology models of RT-CSR mutations. Homology models of the variants were prepared by creating 25 main chain models with 25 sidechain samples at 341K (625 total) within MOE. Intermediates were refined to an RMS gradient of 1 kcal mol^{-1} $Å^{-1}$, scored with the GB/VI methodology, minimized again to an RMS gradient of 0.5 kcal mol^{-1} $Å^{-1}$, and protonated. The final model for each variant was further refined by placing the protein within a 6 Å water sphere and minimizing the solvent enclosed structure to an RMS gradient of 0.001 kcal mol^{-1} $Å^{-1}$. Models were evaluated by calculating Phi-Psi angles and superimposed

against the reference structures. RNA:DNA A-form duplexes were created and superimposed against the empirically derived coordinates for the DNA:DNA duplex. Models were then minimized and contact energies measured within MOE.

Cloning and purification of polymerase variants

Escherichia coli DH10B and BL21 (DE3) strains were used for cloning and expression, respectively. Strains were maintained on either Superior or 2XYT growth media. Polymerases were cloned into a modified pET21 vector using NdeI and BamHI sites. Overnight cultures of BL21 (DE3) harboring each of the variants were grown overnight in Superior broth at 37°C. Cells were then diluted 1:250, and protein production was induced with 1 mM IPTG during mid-log at 18°C for 20 hrs. Harvested cells were flash-frozen and lysed by sonication in 10 mM phosphate, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 10% glycerol, pH 7 (Buffer A). Cleared cell lysates were heated at 85˚C for 25 min, cooled on ice for 20 minutes, and filtered (0.2 μm). The filtrate was then passed over a DEAE column, immediately applied to an equilibrated heparin column, and eluted along a sodium chloride gradient. Polymerase fractions were collected and dialyzed into Buffer A. Enzymes were further purified using an SP column and again eluted along a salt gradient. Pooled fractions were then applied to a Sephadex 16/60 size exclusion column (GE Healthcare), concentrated, and dialyzed into storage buffer (50 mM Tris-HCl, 50 mM KCl, 0.1 mM EDTA, 1 mM DTT, 0.1% Non-idet P40, 0.1% Tween20, 50% glycerol, pH 8.0). Working stocks were made at 0.2 mg/mL.

PCR Proofreading Assay

50 µL PCR reactions were set up with a final concentration of 1x Assay Buffer (60 mM Tris-HCl (pH8.4), 25 mM $(NH_4)_2SO_4$, 10 mM KCl), 200 µM dNTPs, 2 mM $MgSO_4$, 400 nM or forward and reverse primers, 20 ng of template plasmid and 0.2 µg polymerase. Reactions were thermal-cycled using the following program: 95°C - 1min, 25x (95°C- 30 sec, 55°C- 30 sec, 68°C- 2 min 30 sec).

Primer Extension Assay

10 pmol of 5' fluorescein labeled primer were annealed with 50 pmol of template RNA or DNA and 0.4 µg of polymerase by heat denaturation at 80°C for 1 minute and allowing to cool to room temperature. Reactions were initiated by the addition of "start" mix which contained (1x Assay Buffer, 2 mM $MgSO_4$ and 200 µM dNTPs. Reactions were incubated for 10 minutes at 68°C until terminated by the addition of EDTA to a final concentration of 25 mM. The labeled primer was removed from the template strand by heating sample at 75°C for 5 minutes in 1x dye (47.5% formamide, 0.01% SDS) and 1 nmol of unlabeled blocker oligonucleotide (to competitively bind the template strand). Samples were run on a 20% (7 M urea) acrylamide gel.

Reverse transcriptase fidelity (SSCS)

Templates for SSCS were prepared by first strand reverse transcription or primer extension (plasmid DNA template) with barcoded primer. Polymerization reactions were carried out according to manufacturer's recommendations for recombinant MMLV (New England Biolabs). For experimental polymerases, reverse transcription or primer extension was performed in 1x Assay Buffer, 200 µM dNTPs, 1 mM $MgSO_4$, 400 nM

barcoded reverse primer, 40 units RNasin Plus, 0.2 µg polymerase, and template (1 µg Human heart total RNA or 1 ng plasmid). Reactions were incubated at 68°C for 30 minutes (cDNA synthesis) or 2 minutes for DNA primer extension. Single stranded products were PCR amplified using Accuprime Pfx polymerase (ThermoFisher) with reverse and corresponding indexed forward primer. Samples were submitted for Illumina miseq PE 2x250.

Targeted DNA sequencing reads were aligned and grouped based on unique molecular barcodes tagging individual reverse transcription events using ustacks (v1.35). Using a modified version of the single strand consensus sequence program (SSCS), only groups containing three or more reads were analyzed. From these reads, a consensus sequence was built if more than sixty-six percent of the bases at each position were in agreement, otherwise the base was called as N and disregarded in the remaining analysis. Consensus reads were then aligned to the reference sequence using BWA-MEM (v0.7.7), and single nucleotide variants and indels were identified. The polymerase fidelity was calculated as the sum of indels and erroneous bases as a fraction of the total number of aligned bases.

RTPCR Assay

50 µL reverse transcription PCR (RTPCR) reactions were set up on ice with the following reaction conditions: 1x Assay Buffer, 1 mM $MgSO_4$, 1 M Betaine (Sigma-Aldrich), 200 µM dNTPs, 400 nM reverse primer, 400 nM forward primer, 40 units RNasin Plus (Promega), 0.2 µg polymerase and 1 µg of Total RNA from Jurkat, Human Spleen or E. coli (Ambion). Reactions were thermal-cycled according to the following

parameters: 68°C - 30 min, 25x (95°C- 30 sec, 68°C (63°C for rpoC) - 30 sec, 68°C - 30 s/kb).

Single nucleotide incorporation kinetics

Duplexes (DNA:DNA or DNA:RNA) were assembled by combining equimolar amounts of a DNA 25-mer (5'-CCCTCGCAGCCGTCCAACCAACTCA-3') and DNA or RNA 36-mer (3'-GGGAGCGTCGGCAGGTTGGTTGAGTGCCTCTTGTTT-5') in 10 mM Tris-HCl, 0.1 mM EDTA (pH 8.0). Solutions were heated to 95°C for 5 min, slowly cooled to 60°C for 10 min, and then cooled to room temperature for 15 minutes. Reactions (100 µL) consisting of assay buffer, 1 mM $MgSO_4$, and 500 nM duplex were initiated by variable amounts of α-P32-dCTP (0.003-400µM), which was diluted 1:400 in unlabeled dCTP. Reactions were allowed to proceed 3-14 minutes. 10 µL aliquots were quenched by the addition of EDTA (0.25 M final concentration) in 15-120s intervals. Aliquots (2 µL) were spotted on DE81 filter paper and washed 6 times in 5% NaH_2PO_4 (pH 7), 2 times in ddH_2O and finally in 95% EtOH. Dried filter paper was exposed for 24 hrs and imaged on a STORM scanner. Initial rates were obtained by analysis using FiJi (Image J). Kinetic parameters were determined by non-linear regression using SigmaPlot10.

RNA sequencing and analysis

RNA from U87MG glioblastoma cells (ATCC® HTB-14) were harvested using trizol LS following manufacturer's instructions (10296-028, Thermo fisher scientific). Ribosomal RNAs were then removed from the RNA samples using Ribozero rRNA removal kit (MRZH11124, Epicentre) and cleaned using RNeasy MinElute Cleanup Kit

(Qiagen). rRNA depleted RNAs were fragmented using NEBNext Magnesium RNA Fragmentation Module (E6150S, NEB) to 200-300bp size range followed by kinase treatment to prepare for adaptor ligation. Illumina libraries were prepared using NEBNext Multiplex Small RNA Library Prep kit (E7580, NEB) and size selected to remove adaptor dimers using Ampure XP beads. 6 Illumina libraries were prepared from the same pool of RNA using experimental reverse transcriptases and ProtoScript II Reverse Transcriptase from the library prep kit. RNASeq libraries were sequenced on Illumina HiSeq 2000, 2x100bp by the genome sequencing and analysis facility at the University of Texas at Austin.

The evaluation of RNA-seq quality control metrics was performed via RNA-SeQC (v1.1.8). For transcript abundance analysis, fpkm values were generated through the cufflinks/cuffnorm pipeline (v2.2.1) and transformed both by log2 and to fit the range [-3,3].

RNA Sanger Sequencing

Sanger sequencing reactions were set up by preparing 1X Assay Buffer, 1 mM $MgSO_4$, 10 pmol RT.Probe, 50 pmol SangerGATC Template, 0.4 ug RTXexo- , and 50 µM dNTPs. For the indicated terminator nucleotide, a 25:1 ratio of 3' dideoxy terminator to unmodified NTP was used. Reactions were thermal cycled 6x (68°C - 20sec, 85°C - 5sec). Reactions were terminated by the addition of EDTA to a final concentration of 25 mM. The labeled primer was removed by heating sample at 75°C for 5 minutes in 1x dye (47.5% formamide, 0.01% SDS) and 1 nmol of unlabeled SangerBlocker oligonucleotide.

ACKNOWLEDGMENTS

We would like to thank the University of Texas Genomic Sequencing and Analysis Facility (GSAF) for next-generation sequencing support. This work was supported by DARPA (HR0011-12-2-0001), NSSEFF (FA9550-10-1-0169), and the Welch foundation (F-1654).

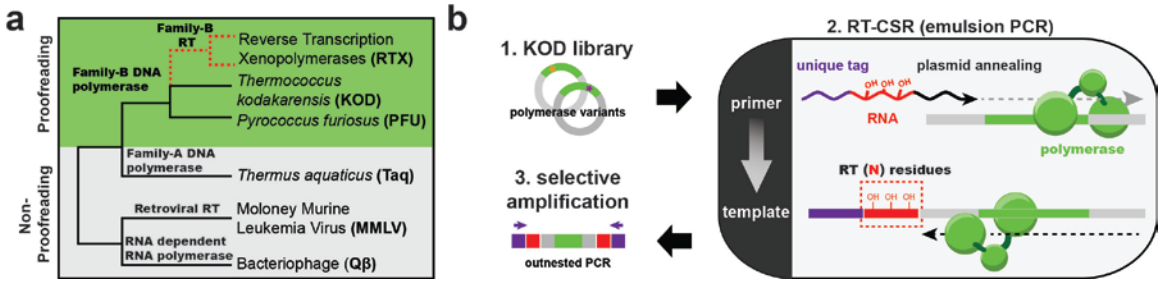
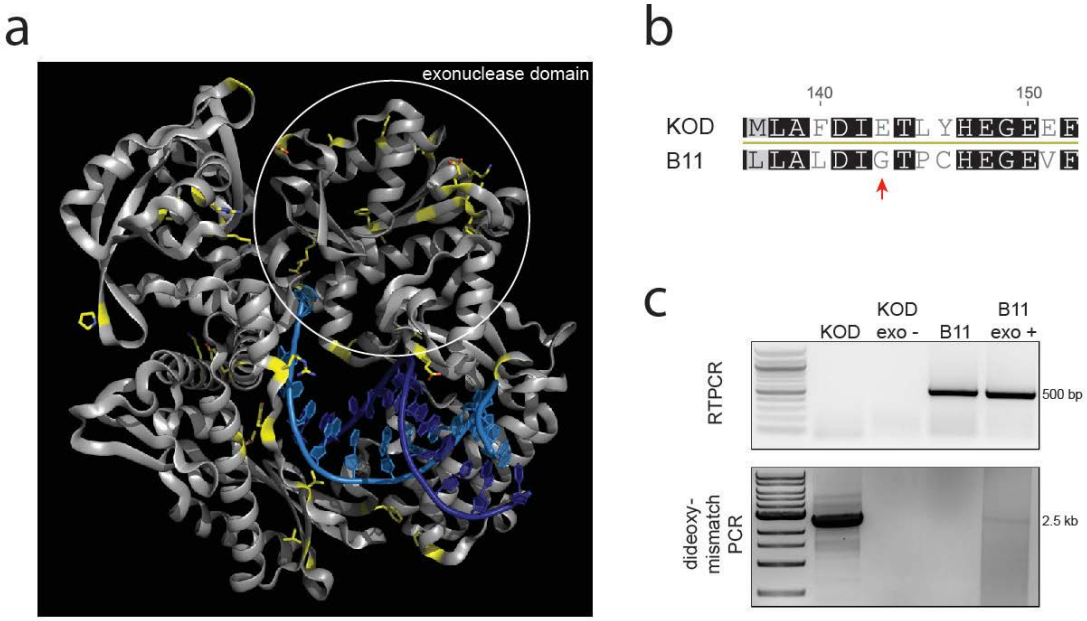


Figure 3.1 : Evolution of a synthetic family of reverse transcriptases by RT-CSR

a, Polymerase phylogeny depicts reverse transcription xenopolymerases (RTX) as a second, evolutionarily distinct, origin of reverse transcriptase. Unlike native reverse transcriptases, Archaeal family-B polymerases are proofreading - allowing RTX the potential for proofreading reverse transcription. **b**, Framework for the directed evolution of hyper-thermostable reverse transcriptase using reverse transcription compartmentalized self replication (RT-CSR). Libraries of polymerase variants are created, expressed in *E. coli*, and *in vitro* compartmentalized. During emulsion PCR, primers flanking the polymerase enable self-replication, but are designed with a variable number of RNA bases separating the plasmid annealing portion from the unique recovery tag. Outnested PCR ensures that only polymerases with reverse transcriptase activity are selective amplified.



d

	DNA			RNA		
Enzyme	***k*cat**	**Km**	***k*cat/Km**	***k*cat**	**Km**	***k*cat/Km**
KOD	160.4	39.7	4.0	nd	nd	nd
B11	48.2	7.9	6.1	54.5	23.1	2.4

Figure 3.2 : Characterization of the B11 Polymerase

a, Mutations in the B11 polymerase (yellow) are mapped onto the KOD polymerase (grey with DNA primer:template duplex in blue). Thirty seven mutations were accumulated, many found in the proofreading domain. **b**, Examination of the 3'-5' exonuclease active site shows a mutation at glutamate 143 to glycine. **c**, Functional assays reveal B11 polymerase is capable of single enzyme RTPCR of a 500 base pair region of the *HSPCB* gene, as well as the B11 with grafted wildtype proofreading domain. Proofreading activity was qualitatively measured in a dideoxy-mismatch PCR, which requires removal of a 3' deoxy mismatch primer before polymerization occurs. **d**, Kinetics of KOD polymerase (D141A, E143A) and B11 polymerase on DNA and RNA templates.

Initial Selection			
Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
97	94.2%	R -> H	58.70%
		R -> S	22.30%
		R -> C	13.20%
587	27.9%	F -> L	15.10%
		F -> L	12.80%
119		R -> H	10.90%

Round 10			
Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
97	97.9%	R -> F	17.70%
		R -> A	11.80%
		Other	68.40%
384		Y -> H	81.20%
210		N -> D	63.70%
389		V -> I	50.20%
587	37.3%	F -> I	14.00%
		F -> L	23.30%
711		G -> S	29.30%
664		E -> K	29.20%
168		A -> T	25.70%
521		I -> L	24.20%
454		G -> D	22.20%
490		A -> T	17.40%
634		G -> D	16.00%
528		I -> L	14.50%
734		E -> K	14.10%
493		Y -> C	13.90%
311		Y -> C	12.10%
292		A -> T	11.80%
137		M -> I	11.30%
677		G -> S	10.90%
440		R -> H	10.80%
144		T -> A	10.80%
171		I -> V	10.60%
748		F -> Y	10.00%

Round 18			
Amino acid Position	Mutation Frequency	Amino Acid Change	Variant Frequency
384		Y -> H	96.00%
97	93.3%	R -> A	20.80%
		R -> F	18.00%
		Other	54.50%
389		V -> I	91.90%
210		N -> D	84.90%
493	83.3%	Y -> C	59.00%
		Y -> L	13.20%
		Y -> F	11.10%
664	82.7%	E -> K	60.40%
		E -> Q	22.30%
711	75.0%	G -> S	46.80%
		G -> V	28.20%
521		I -> L	59.40%
490		A -> T	58.50%
587	55.1%	F -> L	36.80%
		F -> I	18.30%
168		A -> T	36.70%
734		E -> K	34.50%
137	33.9%	M -> I	20.30%
		M -> L	13.60%
748		F -> Y	22.40%
735		N -> K	18.80%
593		K -> N	16.90%
590		T -> A	15.80%
605		T -> I	13.20%
143		E -> G	13.00%
501		R -> H	12.90%
144		T -> A	12.50%
150		E -> D	12.20%
145		L -> P	11.50%
741		V -> A	11.30%
692		K -> R	11.20%
454		G -> D	11.10%

Table 3.1 : Mutations of KOD Polymerase Throughout RT-CSR

Deep sequencing of RT-CSR libraries. Amino acid residues with mutations occuring in 10% of the population are shown in order of frequency. Some positions contained several amino acid possibilities and the sum of frequencies were totaled. Synonymous mutations are not shown.

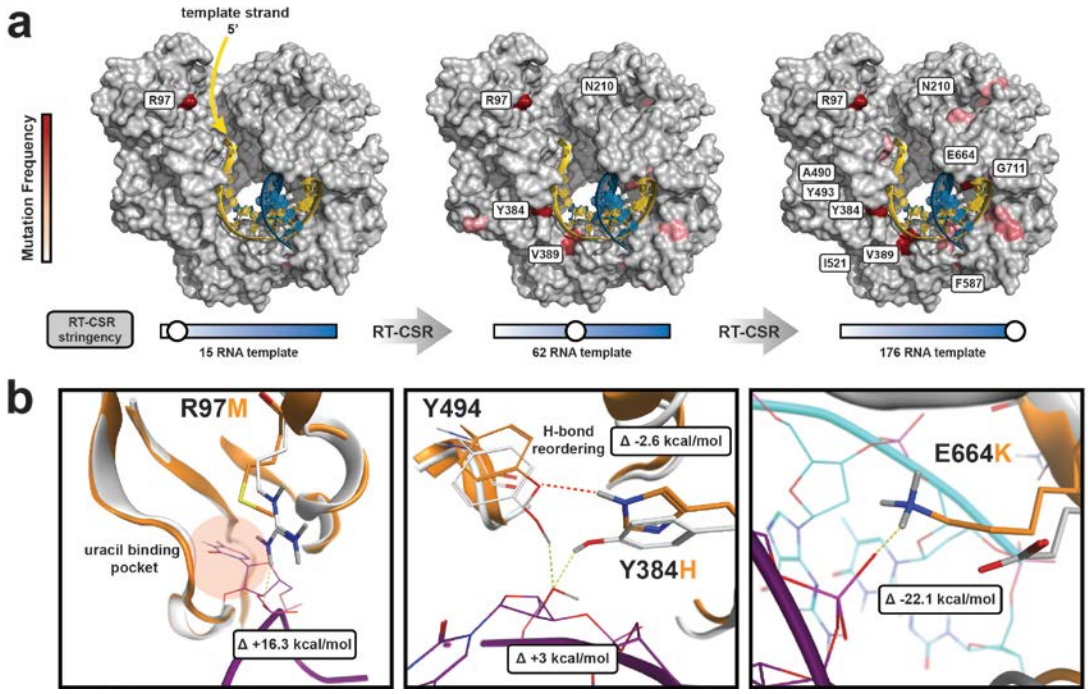
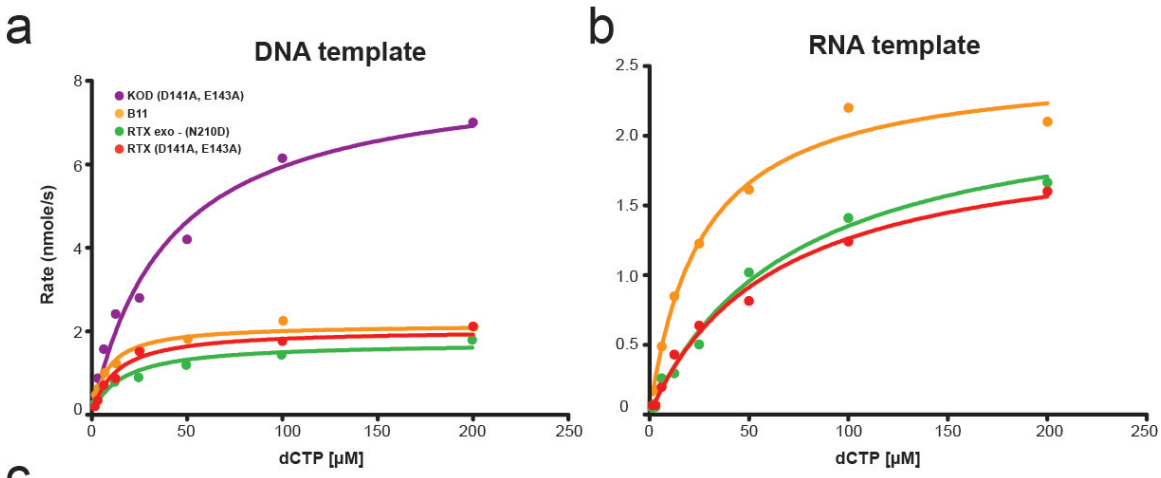


Figure 3.3 : Molecular checkpoints involved in the transition of template recognition

a, Structural heat map of mutated residues over the RT-CSR process found by deep sequencing. Conserved mutations are colored incrementally darker shades of red to indicate frequency in the polymerase pool. Amino acid residues that were mutated in over 50% of the population were labeled. Figure was adapted from KOD structure PDB 4K8Z. **b**, Computer modeling of KOD (grey) and RTX mutations (orange) at checkpoints responsible for DNA and RNA template recognition at R97, Y384, and E664. Free energy changes between wild-type KOD and RTX mutations are inlet displayed.



c

	DNA			RNA		
Enzyme	*k* cat	Km	*k* cat/Km	*k* cat	Km	*k* cat/Km
KOD	160.4	39.7	4.0	nd	nd	nd
B11	48.2	7.9	6.1	54.5	23.1	2.4
RTX	49.1	15.7	3.1	47.8	62.7	0.8
RTX exo-	56.8	12.1	4.7	55.9	71.5	0.8

Figure 3.4 : Kinetic Characterization of Polymerases

Steady-state kinetics of polymerase variants. Initial rates of single nucleotide (dCTP) incorporation by exonuclease deficient polymerases were plotted against the concentration of dCTP using DNA (**a**) or RNA templates (**b**). **c**, Kinetic parameters were estimated by fitting the data to the Michaelis-Menten equation. Nucleotide addition could not be determined (nd) for KOD on RNA templates, due to low activity.

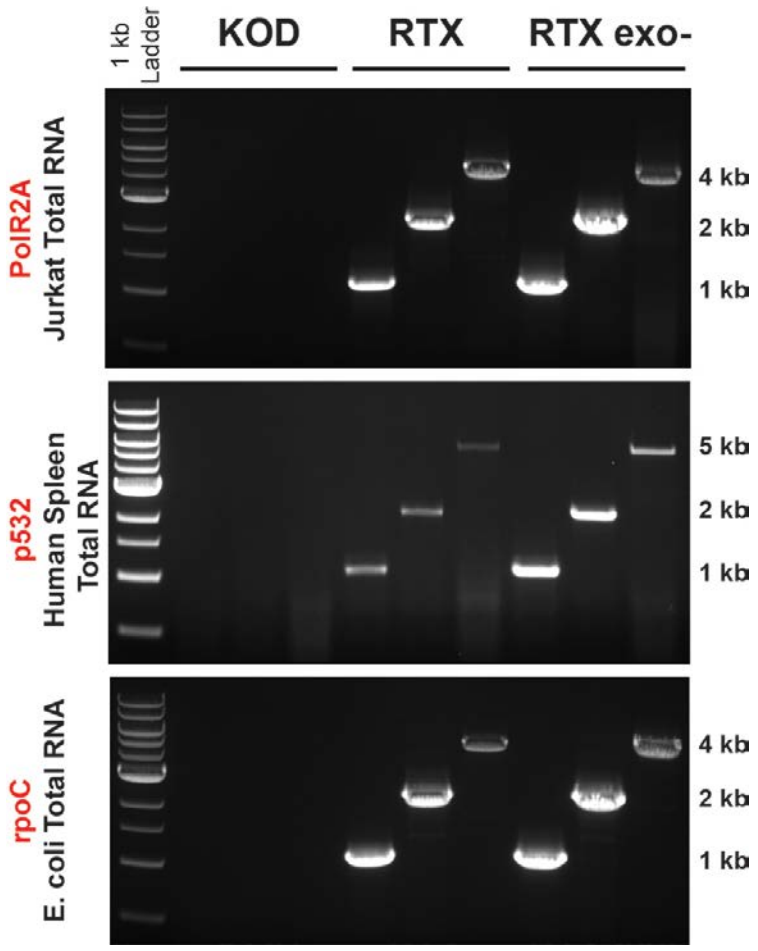


Figure 3.5 : RTPCR Reactions with RTX Polymerase

Reverse transcription PCR (RTPCR) was performed using KOD polymerase, RTX, and the proofreading deficient version of RTX (N210D; exo-). Various genes were amplified (red), two human genes, PolR2A and p532, and rpoC from *E. coli* from various RNA sources. Using gene specific forward and reverse primers, gene regions were amplified, demonstrating efficient single enzyme RTPCR.

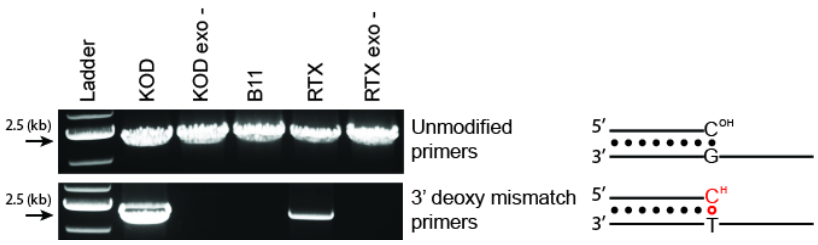


Figure 3.6 : RTX Polymerase Proofreads on a DNA Template

DNA polymerase activity was assessed by PCR using unmodified primers on a 2.5 kilobase fragment. Proofreading (3'-5' exonuclease activity) was assessed by use of 3' deoxy mismatch primers in the PCR. Inactivating the proofreading domain (N210D) in RTX prevents cleavage of the 3' deoxy primer and subsequently PCR.

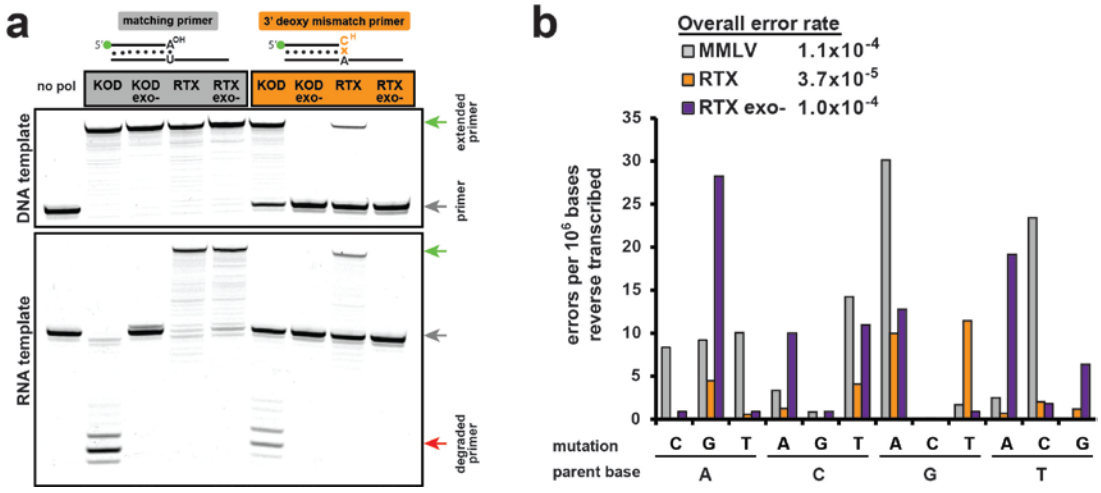


Figure 3.7 : RTX polymerase proofreads during reverse transcription

a, Primer extension reactions of KOD and RTX polymerases and their proofreading deficient counterparts (exo-), on both DNA and RNA templates. Extension reactions were performed with matched 3' primer:templates (grey) or a 3' deoxy mismatch (orange), which must be excised before extension can proceed. The primer is denoted by a gray arrow, extended product in green, and exonuclease degraded primer in red. **b**, Deep sequencing of reverse transcription reaction on HSPCB gene. The overall error rate was determined by dividing the sum of base substitutions and indel formation by the total number of bases sequenced. The error profile of MMLV, RTX, and RTX exo- is shown as number of errors per million bases sequenced.

a

HSPCB reverse transcription

Polymerase	RTX	MMLV	RTX exo-	B11
Total Matches	1.44E+07	1.20E+06	1.10E+06	1.33E+07
Total Mismatch	520	124	102	3136
Total Indel	15	7	11	416
Error Rate	3.71E-05	1.10E-04	1.03E-04	2.66E-04
Base:Mutation	Mutation Frequency			
T to A	1.92%	2.42%	20.59%	23.82%
G to A	27.69%	29.03%	13.73%	6.12%
T to C	5.58%	22.58%	1.96%	2.68%
G to C	0.38%	0.00%	0.00%	0.00%
T to G	3.27%	0.00%	6.86%	8.16%
C to G	0.38%	0.81%	0.98%	2.36%
C to A	3.46%	3.23%	10.78%	6.92%
A to T	1.54%	9.68%	0.98%	1.66%
G to T	31.73%	1.61%	0.98%	4.11%
C to T	11.35%	13.71%	11.76%	10.01%
A to C	0.19%	8.06%	0.98%	0.70%
A to G	12.50%	8.87%	30.39%	33.45%

PolR2A reverse transcription

Polymerase	RTX	MMLV	RTX exo-
Total Matches	1.66E+07	1.12E+06	1.26E+07
Total Mismatch	537	536	4175
Total Indel	54	7	965
Error Rate	3.56E-05	4.86E-04	4.08E-04
Base:Mutation	Mutation Frequency		
T to A	2.61%	0.56%	35.52%
G to A	14.34%	1.68%	2.18%
T to C	13.04%	88.25%	2.68%
G to C	0.74%	0.37%	0.05%
T to G	1.49%	0.00%	1.51%
C to G	0.37%	0.19%	2.35%
C to A	6.89%	0.00%	5.27%
A to T	1.12%	0.19%	2.75%
G to T	34.08%	2.05%	3.83%
C to T	12.66%	2.80%	8.02%
A to C	0.56%	0.75%	1.20%
A to G	12.10%	3.17%	34.63%

b

HSPCB (DNA Template)

Polymerase	RTX	MMLV	RTX exo-	B11	KOD
Total Matches	1.84E+07	2.23E+06	4.65E+06	2.33E+07	1.49E+07
Total Mismatch	1521	297	795	5697	627
Total Indel	305	17	92	852	5
Error Rate	9.93E-05	1.41E-04	1.91E-04	2.80E-04	4.23E-05
Base:Mutation	Mutation Frequency				
T to A	4.67%	5.39%	15.47%	19.89%	2.71%
G to A	13.41%	14.14%	14.97%	9.60%	26.16%
T to C	3.35%	7.74%	3.40%	3.48%	5.74%
G to C	0.13%	0.34%	1.13%	2.42%	0.00%
T to G	0.66%	0.34%	2.14%	3.39%	0.00%
C to G	5.85%	1.35%	11.45%	11.01%	0.32%
C to A	14.73%	10.10%	16.48%	10.88%	34.13%
A to T	1.58%	12.46%	2.52%	6.90%	0.48%
G to T	6.38%	7.41%	2.64%	2.98%	6.54%
C to T	12.29%	8.08%	9.06%	10.67%	8.29%
A to C	0.72%	12.79%	0.75%	1.79%	0.80%
A to G	36.23%	19.87%	20.00%	16.99%	14.83%

Table 3.2 : Fidelity of RTX Polymerases on DNA and RNA Templates

a, Fidelity profile for reverse transcription on two human genes, HSPCB and PolR2A using the SSCS technique. The error rate is calculated by dividing total mutations (mismatch + indel) over the total number of bases sequenced. The frequency of each possible mutation is listed as a percentage of total mutations. **b**, Fidelity profile for DNA template (cloned plasmid DNA) polymerization using cloned HSPCB.

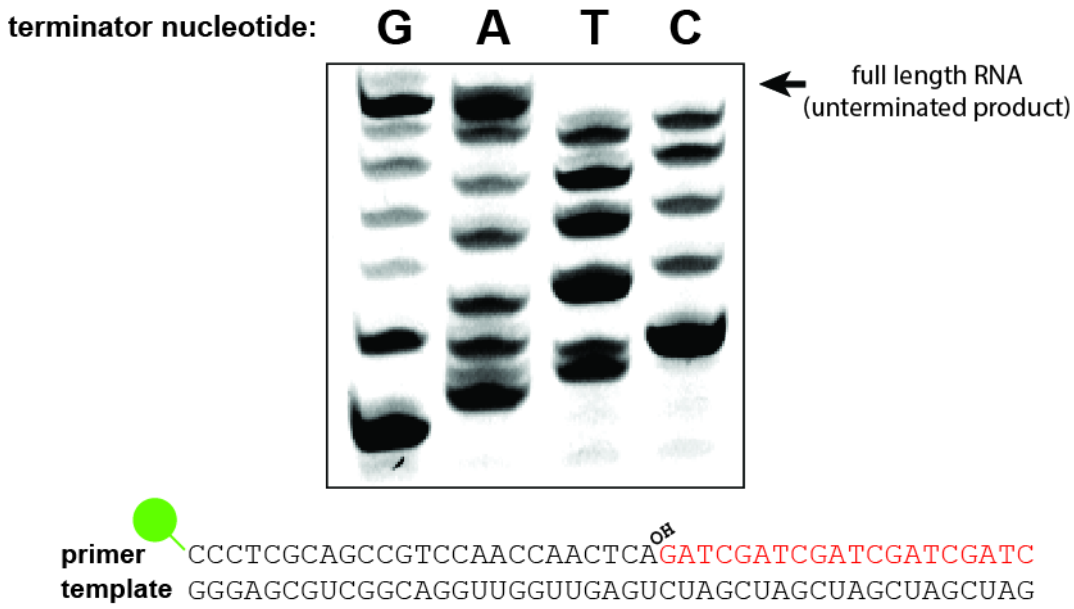


Figure 3.8 : Direct Sanger Sequencing of RNA Templates

Primer extension reactions were carried out with a 5' FAM labeled oligonucleotide with terminator nucleotides (ddGTP, ddATP, ddTTP, ddCTP) at a 25:1 ratio (ddXTP:dXTP). Reactions were performed with RTX exo- to prevent exonuclease cleavage of terminated extension products. The primer:template RNA complex is depicted with the 3' hydroxyl group on the labeled primer. Termination region (sequenced bases) is shown in red.

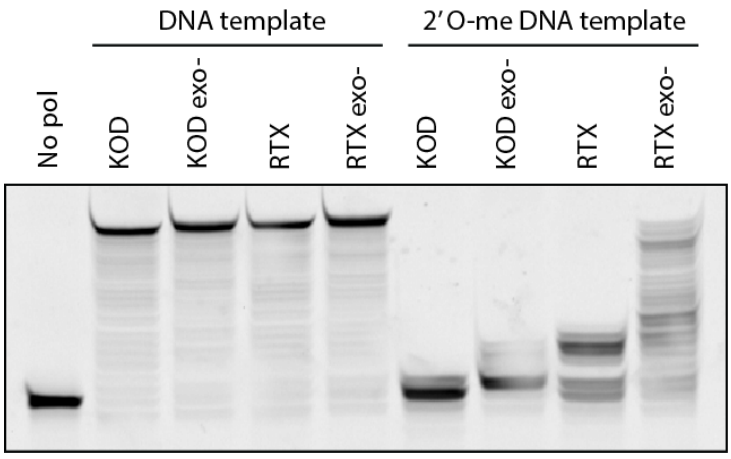


Figure 3.9 : RTX polymerizes across 2' O-methyl DNA Templates

Primer extension reactions on DNA and 2' O-methyl DNA substrates using KOD, KOD exo-, RTX, and RTX exo-. KOD polymerases were not capable of primer extension indicating 2' O-methyl DNA is not a substrate. RTX enzymes could polymerize across 2' O-methyl substrates, but stimulated proofreading preventing fully extended products.

CONCLUSION

The works herein describe the means in which to engineer various aspects of the central dogma of biology. The main role emulsion based technology has played should not be underestimated, as the separation of reactions into individual microreactions allows the selection of large libraries of functional proteins or nucleic acids. This was specifically achieved because genotype-phenotype linkage could be maintained, as well as, the ability to introduce non-native substrates into the emulsion droplets. Using these basic engineering principles several novel systems were developed that enabled the *in vitro* selection of tRNAs, tRNA-synthetases, T7 RNA polymerase, allosteric repressors for a wide range of function, and the creation of the first proofreading reverse transcriptase. In this section, lessons from the selections are discussed that are not simply anecdotal, but are principles that are likely to be true across a wide range of selections for different molecules, especially using the approaches developed here. A discussion about future selections and methods will also be included as perhaps a jumping point for future ideas.

AN UNEXPECTED JOURNEY

In our initial projects in graduate school, Adam Meyer and I had been working on an evolution scheme using an *in vitro* autogene system where a T7 RNA polymerase would copy its own gene, by placing its promoter sequence upstream of the T7 coding sequence (Davidson et al., 2012). This scheme is elegant in its simplicity, which is why it was such an attractive approach as an engineering platform. In the end, this system was not robust enough to allow even the evolution of itself and certainly not in more complex

pathways. After several cycles of selection, the autogene would error catastrophe and be overtaken by molecular parasites that arise during the reverse transcription and PCR recovery of the RNA. Whilst we could speculate in depth about why this system was not evolutionarily stable, the highest level justification was that the selective advantage of the most active polymerases was lost. This could be for many reasons: perhaps each compartment only contained enough resources to make a single protein copy of the T7 gene, perhaps there were not enough RNA nucleotides to allow the self-amplification in any meaningful way, or maybe the nuclease that is often used to treat lysates had residual activity and degraded what little RNA came out of the system in the first place (Pelham and Jackson, 1976). Whatever the problem was, the system would collapse after several cycles of selection and become useless for self evolution, and the fact that it could not even evolve self meant that trying to make the system more complex would be a fruitless endeavor.

Despite the failures of the autogene, it was a proving ground for directed evolution chops and theories. The autogene was a black box and it was hard to peek inside to try and diagnose issues. The rabbit reticulocyte lysate was provided from a commercial source and therefore the treatment procedures were largely unknown and the consequences not understood (Davidson et al., 2012; Ghadessy and Holliger, 2004). Lysates in general only have a limited capacity for transcription and protein production. Unlike a living system, a lysate cannot respond to signals and produce more protein or recycle metabolites. The buildup and accumulation of expended metabolites will inhibit the transcription and translation capacity (Carlson et al., 2012). In addition, it is

conceivable that the procedures involved in the breaking of the emulsion bubbles could introduce RNase which would degrade what little products the autogene made. But despite all the concerns and unknowns, in the end every piece of the autogene puzzle was optimized to allow even the slightest amount of evolution to occur. The techniques used to emulsify transcription/translation reactions could be used for other schemes down the road. Troubleshooting reverse transcription and PCR reactions was also a formative experience. In the end, the autogene was a trial by fire and clued me in on what would the most desirable features be in a directed evolution platform.

What are the features of a powerful directed evolution platform? There are several features which would distinguish itself from the problems faced by the autogene and many other popular directed evolution platforms (Glasscock et al., 2016; Packer and Liu, 2015). The technique should be modular and adaptable. Selections that are specific to the evolution of a single component are limited, and would likely not be a platform technology which could be applied to a wide variety of situations. Also, the platform should give a large selective advantage to the library variants which are the most functional for the desired property. The most functional biomolecule gene variants should undergo a large enrichment in a single cycle of selection. Such that, if 1 in 500 molecules were active (vs. inactive) then following a single round the functional variant would come to the fore. The selective advantage should also be tunable, such that over the course of evolution the selective pressure can be adjusted. This is sometimes necessary as the desired function cannot be mutationally reached with a single point mutation, and the

biomolecule must be gently guided by modulating the stringency over time - allowing the accumulation of beneficial mutations.

The selection should be able to sample large library sizes, unlike a traditional screen where point mutations are tested individually or in a multiwall plate format. Oftentimes, an entire active site of a protein or RNA will need to be completely randomized (driven mostly by the fact that we have no real way of understanding the active site, or modeling one). Even the randomization of five amino acids of a protein will result in a large library ($\sim 3 \times 10^7$), if amino acids are fully randomized to allow every possible combination of the twenty amino acids. Sifting through a large sequence space of possibilities is often driven out of necessity. The selection should facilitate troubleshooting, unlike the issues faced previously with the autogene. Being able to examine the background biology using fairly standard techniques, such as checking expression levels from a promoter with a fluorescent protein, is oftentimes required to hone in on a set of conditions that facilitates *in vitro* selection. A sometimes overlooked, but logistically important, issue is how time consuming and practical the selection process is. In my experience, certain *in vitro* selections can take many cycles to generate the desired function. A complex phenotype may require iteration of the directed evolution cycle dozens of rounds (Johnston et al., 2001). Turning the evolutionary crank takes trial and error, and being motivated to turn that crank is where simplicity is important. Fortunately, I stumbled upon a system which meets these criteria.

The conception of compartmentalized partnered replication (CPR) was inadvertent. Collaborations with Michael Hammerling and Jeffrey Barrick about the

evolutionary implications of an expanded genetic code left me thinking about systems in which I could explore this question (Hammerling et al., 2014). Initially, we had been pushing forward the notion of evolution of entire organisms with an expanded genetic code, but this posed some large hurdles. When studying whole genomes, it is often hard to know the effect of individual mutations that occurred, unless they have previously been characterized. This is expounded by the fact that genomic evolution studies can generate a pile of data. When evolving phages, this was certainly the case. Without huge efforts to biochemically characterize the effect of nonstandard amino acids (NAAs) that appeared across the proteome, we essentially had no idea what they were doing on a structural and functional basis. For instance, two highly conserved amber codons (by proxy the unnatural amino acid 3-Iodotyrosine) appeared in the phage lysis protein, holin, and one in the T7 RNA polymerase gene. It is clear based on the very high enrichment during evolution that the NAAs in these proteins was giving the phage a selective advantage, perhaps due to increased function of the protein, but how? Because of the complexities studying organisms with expanded genetic codes, I wanted to focus on what impact NAAS would have at the individual protein level. Using a model directed evolution system that utilized a well studied protein would enable me to probe this question. Are there generalizable functionalities that a specific NAA could impart on proteins? For instance, a NAA could possibly promote better folding of a polypeptide (for instance if it was highly charged (Lawrence et al., 2007)), or could enhance the active site of a conserved protease, or increase the thermal-stability of a protein (which was later demonstrated (Ohtake et al., 2015)). There needed to be a general platform for studying

this question. As I was exploring the options for which platform might lend itself to this problem, I inadvertently drew the first diagram of CPR .

CPR was born from the brilliant directed evolution scheme, compartmentalized self replication (CSR) that was developed by Philipp Holliger (Ghadessy et al., 2001). Exploring all turnkey directed evolution schemes that were available led to drawing the CSR scheme (polymerase evolution) with the NAA machinery (Fig. C.1). While this was initially intended to be a platform for the evolution of the polymerase, it was immediately obvious that this could also be a directed evolution platform for the NAA machinery, and more generally, anything that could influence the expression of Taq polymerase (or any PCR polymerase). For instance, we could go back and evolve T7 RNA polymerase which had previously failed with the autogene selection by simply driving the expression of the Taq polymerase with the T7 promoter. In order to select for different genes, primers could simply be reprogrammed to not flank the polymerase gene (like in CSR), but flank whatever gene or gene circuit was influencing the expression of the polymerase.

Another important distinction exists between CSR and CPR. In the CSR setup, each polymerase library variant (or cell) produces an equal amount of polymerase, and the differential replication of variants is driven by the functionality of the polymerase in the given selection parameters. In contrast, the selective advantage of library variants in CPR is driven entirely by the differential expression of the polymerase, such that there is a spectrum of Taq expression levels between variants. Based on how PCR works, it would stand to reason that the more polymerase that was present in a cell should greatly influence the efficiency of replication. This basic principle was demonstrated in the

manuscript (Ellefson et al., 2014); however there is an interesting caveat where each compartment has a finite amount of resources (for instance primers or dNTPs). If overcycled, compartments with less polymerase can catch up to compartments with more simply because resources were expended more quickly in the compartment with more polymerase. However, this issue is easily overcome by not overcycling the reactions during CPR.

CPR Advantages Over Existing Directed Evolution Technologies

Fortuitously, CPR has a number of advantages over preexisting directed evolution technologies – making it a strong contender as a platform technology for the directed evolution of a wide range of biomolecules (Packer and Liu, 2015). This is largely due to the unique combination of the selection being a combined *in vivo* and *in vitro* approach, with the biological circuit functioning during cell growth and the partnered replication happening inside of the emulsion bubble. Most existing methods for directed evolution of a wide array of functions (for instance based on gene expression) can be generalized to several categories with severe flaws. A classic example is the tying of gene function to expression of an antibiotic resistance marker. Gene libraries are cloned such that when functional will express a resistance gene and allow a given cell to survive and develop a colony on a plate or growth in liquid media. While in theory these selections should work marvelously, and have worked for the evolution of many biomolecules (Porcar, 2010), they are extremely prone to cheat. Since such a strong selective pressure (survival) is used, a cell will capitalize by any means necessary in order to survive. For instance, low levels of cloning artifacts while preparing the library or recombination during growth can

allow a cell to survive in such a way that it is not influenced by the desired gene. Oftentimes these artifacts will actually have a growth advantage over even a functional member, because they may more directly make the resistance gene or do not need to express the protein of interest which can sometimes be toxic.

A continuum of selective advantage is available during CPR. Gene or circuit function is directly associated with the *in vivo* production of more and more Taq polymerase, which subsequently will more effectively PCR amplify itself in the subsequent step. This is in contrast to a technique such as fluorescence based cell sorting (FACS), where sorting gates are used. While multiple gates can be set up, there is a clear line of being good enough. Antibiotic selections can have the same disadvantages, as oftentimes the dynamic range can be quite limited as long as the cell can produce enough of the resistance gene; it will survive a wide range of antibiotic concentration in the media.

Unlike most selection schemes, a large sequence space can be covered with relative ease. A single emulsion reaction will simultaneously select roughly 10^9 cells during the time frame of a single PCR reaction, without any additional manipulation. The bottleneck for library size is the *E. coli* transformation efficiency, which can typically allow up to 10^8 unique transformation events. While an antibiotic selection should be able to cover the same library size, the physical plating of the cells can be tricky such that single colonies can be isolated or to not allow growth of cells on top of a lawn of dead cells. While groups have reported FACS selection of libraries this size, this is clearly a monumental effort that cannot be routinely performed. Oftentimes to achieve optimal

sorting efficiencies, the number of events over time must be lowered, as faster sorting is often associated with accidental non-active cells being sorted. Even at its best, sorting is typically capped at ~10^6 events per hour which means that sorting a library of 10^7 would take ten hours. This is problematic not just as a time constraint, but cells awaiting sorting will often sit in a saline buffer solution which can have unknown consequences.

CPR is conducted in discrete non-continuous cycles. While this may limit the continuous march through sequence space, the limiting of selection parasites is of utmost importance. For instance, the most publicized continuous evolution scheme, phage assisted continuous evolution (PACE) (Esvelt et al., 2011), is prone to parasitic contamination (personal communication). The system uses M13 bacteriophage which harbors a gene that will influence the expression of an essential coat protein (G3P), which is located on an *E. coli* plasmid. A "lagoon" is set up in which fresh *E. coli* cells are pumped in and expended cells and M13 virus are sloughed off. This allows genes in the M13 to optimize over time, by iterative replication and mutations, to influence the expression of the G3P. However, the most common solution the phage will find towards increased production of G3P is to overexpress the gene influencing it. This is oftentimes the easiest way to improve function (is to simply have more of it), but does not result in proteins of increased function. The discontinuous nature of CPR allows the genes to be re-cloned into fresh plasmids, such that promoter strength and genetic context is the same between rounds. In addition, certain functionalities cannot be obtained by single point mutations (which is how phage explore sequence space) and require multiple mutations to happen simultaneously (Romero and Arnold, 2009).

FUTURE DIRECTIONS OF CPR BASED EVOLUTION

CPR relies on the differential expression of thermostable DNA polymerase, based on the functionality of a genetic part or circuit. There are many ways to directly tie the expression of Taq to certain cellular processes, as we have demonstrated for both transcription and translation. However, there are many enzymatic or protein functions that cannot directly be tied to gene expression. For instance, metabolic pathways that produce small molecules may not have a visible phenotype or affect cell growth. Metabolic engineering efforts, to produce valuable products using engineered organisms such as artemisinin (Paddon and Keasling, 2014), are often a great challenge because no visible phenotype is assocated with the small molecule of interest. However, life has found ways to alter gene expression based on the presence of small molecules by evolving regulatory genes that will specifically respond to them. For instance, *E. coli* regulates internal metabolites by placing genes into an operon structure which is often regulated by multiple components (McAdams et al., 2004). A common method for regulation is by an allosteric transcription factor, which conformationally responds to the small molecule of interest and regulates the pathway accordingly. By hijacking and rewiring allosteric transcription factors, gene expression can be influenced by the production (or degradation) of small molecules – allowing CPR to be used (Fig. C.2). Due to how allosteric transcription factors function, a dose-response curve is associated which can alter gene expression to varying degrees based on the concentration of the small molecule effector inside the cell. The dose-response of certain molecules can vary widely, ranging from a digital type response (“all or nothing”) or a graded response

("analog" response). The difference in response is largely transcription factor dependent, and can be influenced heavily by factors such as cooperative binding (Giorgetti et al., 2010). The dose-response properties of biological circuits can be tuned heavily based on design principles of synthetic biology (Daniel et al., 2013). For instance, redesigning the ribosomal binding site, operator sites, or even negative feedback on itself can influence the allosteric transcription factor dope-response. These parameters are critical for tuning a CPR circuit, to promote the most gene expression from the most functional variants in a library.

Given that most biotechnology or industrially relevant small molecules do not have an allosteric transcription factor that can be found in nature, it is almost certain that one will have to be engineered from scratch. Fortunately, CPR can be utilized for this as well. As a proof of concept, the tryptophan repressor was chosen to gain insights on the evolution of allosteric transcription factors (discussed in Chapter 2). The tryptophan repressor (TrpR) was chosen because it is very well studied and structures exist, which heavily guide library design. Secondly, it was chosen because I had previously evolved an aminoacyl tRNA synthetase that can incorporate 5-hydroxy-L-tryptophan in response to amber codons. My main objective was to create a nonstandard amino acid operon that would biosynthetically produce, regulate, and incorporate an unnatural amino acid. However, this did not work quite as expected. The first goal of the evolution of the TrpR was to alter the binding pocket to allosterically respond to new small molecule ligands – as engineering allosteric binding pockets to a wide variety of small molecules would be

necessary if CPR is to be adapted for metabolic engineering. Changing the pocket proved more challenging than originally conceived.

The CPR selection circuit was set up around positive and negative selection cycles. Binding would be selected for in the positive round, and not binding would be selected for in the negative round. This is essential because allosteric transcrpiton factors can typically evolve to become non responsive to the small molecule target and become always active irrespective of the small molecule (termed super-aporepressors). By cycling between rounds of positive and negative selection, a dynamic transcription factor can be selected that will only respond when the desired small molecule is present. In total, over half a dozen small molecules were used for directed evolution of the binding pocket. However, various results were achieved. Amino acid residues in the binding pocket of tryptophan would become fixed throughout the course of evolution (suggesting honing in on the best combination of amino acids), but the response varied widely. Some small molecules (including 5-hydroxy-L-trypophan) would only cause a several fold change of promoter regulation even at high concentrations (~1 mM). In the end, two molecules worked well, 5 and 6-bromo-L-tryptophan, and could alter gene expression by up to 50-fold.

This poses a problem, especially if transcription factors are going to be a modular scaffold for evolution of recognition of a wide variety of small molecule targets. Instead, focus was put into driving the evolution of other functions of the allosteric regulator, such as altering the operator binding site to respond to new DNA sequences. In the end, fifteen unique repressors were evolved that responded to tryptophan, 5-bromo-tryotophan, or 6-

bromo-tryptophan across five orthogonal DNA operators. The ability to take a single regulatory protein and evolve so many functions was very interesting from an evolutionary standpoint. It shows the molecular plasticity of proteins in general, but especially the ability for regulatory proteins to respond to new signals and regulate novel operator sites. This reveals the potential for a simple regulatory protein to develop complex regulatory architectures. However, there is a limitation for stacking the molecular logic of the tryptophan repressor by the expression of multiple repressors in a single cell.

The tryptophan repressor folds into a homodimer prior to binding to tryptophan and before binding the operator site. The interface region (which dictates dimer folding) is shared among all the variants created by directed evolution. If multiple species of repressors were expressed in the cell at the same time, then repressors would randomly dimerize and each would have a non-defined logic (or logic interference). To overcome this problem, libraries were made of a region of the interface that is known to prevent proper folding of the dimer. New binding pairs were identified, and were found to have various levels of success. In addition to this approach, it also seemed feasible to covalently attach the C-terminal residues to the N-terminal of the next dimer, creating a fusion (or tethered) repressor protein. This should promote intramolecular folding with the second half of the dimer, instead of dimerization with another repressor, and allow the expression of multiple repressors each with unique logics without interference between repressors. Using this approach the tethered repressors displayed NAND based logic; only binding and repressing gene expression when both conditions were met. For

instance, if the 5-bromo variant was tethered to the 6-bromo variant then repression would only occur when both small molecules are present.

Despite the relative successes of engineering the tryptophan repressor, there were some severe limitations that would prevent this from being a platform for engineering more industrially relevant metabolic pathways. The initial attempts to change the allosteric effector specificity led to many failures, and exhaustive attempts were made at finding tryptophan analogs that would work. My fear, even before beginning work on the tryptophan repressor was that the binding pocket looked rigid. Based on how the dimer folds, there are many structurally important residues in or near the pocket, making major modifications a concern. In addition, the back wall of the pocket is an alpha helix, which limits the types of molecules that could possibly fit into the pocket. Some of the most adaptable binding pockets are generally less structured and play negligible roles in the folding of the protein (Dellus-Gur et al., 2013; Tóth-Petróczy and Tawfik, 2014). If designing the perfect platform transcription factor, one that could bind a wide array of small molecules, a different scaffold besides the TrpR would have to be used. This could be a more widely used scaffold such as the Lys-R type transcription factor family, which are the most broadly used regulatory family in prokaryotes. Perhaps a transcription factor could even be designed from scratch, by the fusion of a DNA binding domain to a highly engineerable binding pocket, like a single chain antibody.

POLYMERASE EVOLUTION OF FUNCTION

Before beginning to lay the foundations of CPR, I attempted to amplify a portion of the Taq polymerase gene using CSR. The demonstration was beautiful - a single clear

PCR product was visible directly after the emulsified CSR reaction was broken. Although focused primarily on pushing CPR forward, the notion of polymerase engineering stuck with me. The first thought was to try and gene shuffle two very utilized family-B polymerases for high fidelity PCR reactions, the replicative polymerase from *Pyrococcus furiousus* (PFU) and *Thermococcus kodakarensis* (KOD). Both of these polymerases share a high degree of sequence homology (making gene shuffling feasible). Despite having nearly identical sequences, there are functional differences such as KOD having higher processivity or PFU being higher fidelity. In a CSR selection, libraries of gene shuffled variants could be parsed to find polymerases that may contain the best features of both polymerases. A rotation student, Daniel Garry (and later Thomas Wall), began this project and turned several rounds of CSR. Shuffled polymerase variants were screened and identified. These displayed promising characteristics, such as the ability to generate more PCR amplification (which CSR often enriches for). In fact, the Ellington lab as a whole still uses a variant of these polymerases, A12, which is useful mostly as an error prone polymerase. Although publications were not generated as a result of this work, it laid the groundwork for what would later be the evolution of the first proofreading reverse transcriptase.

Since CSR selections became a routine process in the lab, there was general curiosity if we could also engineer another DNA polymerase, the reverse transcriptase. However, there were some serious hurdles to using a CSR type evolution scheme with common reverse transcriptases such as MMLV. The CSR process uses high heat as an initial step to break open cells that contain the polymerase variants, which would heat

denature common retroviral RTs. Secondly, the primers used to flank the polymerase gene in CSR would have to be redesigned to specifically bind RNA, which in an organism like *E. coli* can be difficult because the DNA sequence is identical. Primers would indiscriminately bind the RNA or DNA and template, causing amplification even if reverse transcriptase activity was not present. Exponential amplification would also be impossible, again given the lack of thermostability that is required to drive a PCR reaction (plus after the first cycle of PCR, the RNA template would become irrelevant). It took years for these mental hurdles to be overcome.

To formulate the concepts behind the first proofreading reverse transcriptase, two ideas needed to come together. If primers were designed with RNA bases that would flank the polymerase in the CSR, then during the synthesis of the second strand the primer would serve as a template. Designing a unique capture sequence at the ends of the primer ensured that only polymerases that could reverse transcribe across the RNA challenge region would be amplified. This was a great solution for two main reasons, the polymerase would have to reverse transcribe every cycle of PCR (in contrast to reverse transcribing its own gene a single time) and the stringency of selection could be easily tuned by the stepwise addition of longer RNA stretches in the primer region. In fact, as more and more RNAs are added to the primer, eventually the entire sequence will be RNA - requiring efficient reverse transcription every cycle of PCR to maintain exponential amplification in the reaction. The second idea, was perhaps boldness or ignorance. A thermostable (capable of PCR) polymerase needed to be used in this pipeline. KOD polymerase was an obvious choice (it was my favorite polymerase)

because it is extremely processive and has a highly active proofreading domain. However, KOD polymerase does not have any initial reverse transcriptase activity, but I reasoned it could be evolved given the chemical differences in DNA and RNA templating should be quite small. I thought there must be amino acid residues that enforce strict DNA utilization and that these could be mutated to residues that would be accepting of RNA templates.

As rounds of selection, termed reverse transcription compartmentalized self replication (RT-CSR), were turned the notion that a polymerase like KOD could be a reverse transcriptase became less hypothetical. In subsequent cycles, the polymerases were able to survive on selective pressure where in previous rounds it was not – suggesting a gradual transition to reverse transcriptase activity. Sequencing of the variants in early rounds of RT-CSR showed two very prominent mutations, R97X (X = multiple possibilities) and N210D. The R97 mutation is certainly very intriguing, as it sits next to the uracil binding pocket. This pocket, in the native polymerase, scans the template for uracil residues which should never exist in template DNA unless cytosine deamination occurs, and stalls - giving time for repair machinery to correct the mistake. Interestingly, when the structure of this complex was resolved researchers made a mutation to V93Q (Fogg et al., 2002), which sits inside of the uracil binding pocket. The RT-CSR process (and actually normal CSR) consistently identified R97 as the key residue for inactivating this pocket, presumably due to disruption of an ionic bond to the backbone phosphate. Mutation of the uracil binding pocket makes perfect sense, as RNA is chock full of template uracils, but the intriguing part is that the selection process

identified a residue that presumably accomplishes what V93Q does, but better. This is interesting, and suggests that other mutations commonly used in the polymerase may not be optimal residues. For instance, the Y409G and A485L mutations are commonly used for incorporation of dNTP analog substrates, but may not be truly optimal since they were identified from small screens as a result of a crystal structure (Gardner and Jack, 1999). The other position, N210D, has been shown previously to inactivate the proofreading domain of the polymerase (Nishioka et al., 2001). While initially it was thought that this mutation might be essential to the evolution of RT activity (thereby preventing a proofreading reverse transcriptase), other experiments confirm that this is a general phenomenon of CSR selection itself - as polymerases attempt to maximize their self replication (and inactivation of the proofreading domain is a simple strategy to achieve this).

Throughout the eighteen rounds of selection the polymerase accumulated many mutations. Each cycle of RT-CSR generally will require over 60 cycles of PCR for reamplification, recovery, and the self replication during the selection (which likely introduces the most mutations given the proofreading domain is mutated). By the final rounds of RT-CSR, polymerases contained roughly forty mutations. Many of these were simply due to the accumulation of neutral mutations given the heavy mutational load of the system, and sequence conservation and location served as guidelines for the mutations that were most important for RT activity. In addition, experiments where the wild-type 3'-5' exonuclease domain was transplanted on the mutant RT suggested that proofreading could be restored by eliminating spurious mutations. By designing

polymerases that only contained the core set of mutations required for RT activity, the KOD based reverse transcriptase recovered its proofreading activity which was found to act both on RNA or DNA template.

Since this is the first demonstration of a proofreading reverse transcriptase, what were the effects on the fidelity of the RT reaction? The effect of the proofreading domain on DNA templated polymerization is well known to increase the fidelity of genome synthesis, but would this translate to reverse transcription? Most tools to measure the fidelity of polymerases are inherently flawed because they use methods relying on cloning of a genetic or phenotypic marker and count cells based on presence or absence of the trait screened/selected for. The basic equation for fidelity consists of the fraction of positive by the negative phenotypes with added assumptions including the number of doublings or the probability that a given mutation will destroy function of the gene. The inability to detect neutral mutations or cloning artifacts has had consequences in the consistency of data about polymerase fidelity from lab to lab, sometimes ranging several orders of magnitude (Moser et al., 2012; Takagi et al., 1997). Due to the traditional flaws of fidelity measurements and that they do not precisely measure the types of mutations, a new assay was developed. The assay is based around the idea of single stranded consensus sequencing (SSCS), which was modified slightly to contain unique barcodes in the reverse transcription primer. This allows a single reverse transcription event to have an appended barcode sequence which can be used during next-gen sequencing to negate errors that were not derived during the RT by creating a consensus sequence of all reads. While the SSCS method also has a limited threshold for error detection ($\sim 3 \times 10^{-5}$) due to

error that occur during PCR amplification of the single stranded cDNA, it should give a more accurate and consistent account of the fidelity of a polymerase.

In addition, this relative fidelity approaches the error rate of transcription itself which without a further methodological advance will cap the possible error detection (regardless of using a genetic based marker or a sequencing approach). However, other research groups have developed clever methods for precisely measuring the fidelity of transcription (Gout et al., 2013; Traverse and Ochman, 2016). If these techniques were coupled with our approaches, it could enable exact error detection of polymerization. The ability to measure fidelity in a high throughput fashion would not only be interesting for understanding the RTX polymerase developed in Chapter 3, but will also aide in understanding how other replicative polymerase complexes make errors during genome replication.

THE FUTURE OF POLYMERASE EVOLUTION

A growing field of interest is the use of synthetic nucleotide analogs, which have been shown to increase nucleic acid drug efficacy, increase stability, and even be used as a third base pair in a living organism (Malyshev and Romesberg, 2015). Unfortunately these fields have been stifled because polymerases do not use these bases readily, and often need to be engineered or have altered buffer conditions to polymerize effectively. In addition, highly modified nucleotides often need to be converted back to DNA in order to complete *in vitro* selection cycles which requires a reverse transcriptase that can make that conversion. With the evolution of RTX, RNA bases were programmed into the primer used during the selection, but other base analogs can be used in place of RNA.

This effectively shifts the selection pressure to enforce altered template utilization in the reverse transcription. The resulting polymerases gives a starting point for the evolution of biomolecules with nucleotide analogs in a more straightforward way, as selection approaches have gotten more clever at overcoming this problem (Kimoto et al., 2013; Sczepanski and Joyce, 2014). In addition, this may be a starting point for the generation of polymerases that utilize altered substrates throughout an entire PCR reaction.

Polymerases that especially adapted for use with the unnatural base pair (UBP) systems (Fig. C.3) (Malyshev et al., 2014; Thyer and Ellefson, 2014), may enable a more stable propagation *in vivo*. This has limited productivity in the field (personal communication) because they are quickly lost after the course of several replication events. This hinders further development of the biotechnology and the evolution using these systems. Tailor made polymerases that are friendlier to replication of UBPs may be achieved by using the RT-CSR technique. This may promote UBP use in diagnostic technologies and, although RT-CSR may not be capable of evolving the *E. coli* DNA polymerase, structural identification of residues which are more favorable to UBPs may be obtained.

CONSIDERATIONS OF SELECTIONS

Through advancing these projects, critical parameters about how emulsions bubbles are formulated changed. In the initial CPR selections, the classic technique of using a stir bar spinning in oil - whilst aqueous was slowly added was used. This technique is very time consuming – especially for multiplexing selections, and was highly subject to variability depending on the volume of aqueous phage added over time.

This would result in variability between people: with better, more uniform, bubbles. Revisiting how emulsions were set up, we found that the emulsion bubbles used in the original CPR paper resulted in multiple *E. coli* inside each compartment (Fig. C.4). This should in theory greatly reduce the selection efficiency by allowing non-functional individuals to "tag along" with functional counterparts. Interestingly, the selections still worked. More rounds were possibly involved; as initially roughly eight rounds were turned for both the T7 RNA polymerase and tRNA selections, but they still worked and honed in on what were likely the best solutions. The process of setting up emulsions was greatly improved through the use of a tissuelyser (Qiagen) which shakes at specified frequencies. Whilst shaking, the emulsification is facilitated by the use of a nucleating structure - which in our case a 1 mL syringe plunger was most handy. These alterations made emulsion bubbles more uniform and most bubbles now contained only a single *E. coli* cell, as well as, making the process more highly throughput and consistent across individuals in the lab. While some speculation is required, this improved the selection efficiency such that in some cases large libraries (~10^7) could be selected in as little as three rounds of CPR.

Another important consideration about selections and library design was discovered during the repressor selection in Chapter 2. Whilst selecting for the O_D binding repressor, the best variant contained a one amino acid deletion. This is surprising, as the winner of the selection contained an incredibly rare event that should technically not have happened. During oligonucleotide synthesis, -1 base deletions are somewhat common due to incomplete base incorporation. Within the oligonucleotide pool that was

use to construct the library (NNS randomization), three consecutive -1 deletions occurred within a short stretch, resulting in the deletion of an entire codon, in addition to containing the correct amino acid combination that would result in a tight binding repressor. The chances of this happening seem infinitesimally small and perhaps were selected because of pure luck. I think this speaks to a general misappreciation for how libraries are generally designed for directed evolution. The common practice for library design is to simply choose amino acids that are within proximity to the region of interest and do randomization of amino acids. Almost certainly as important as the amino acids themselves, is the positioning of residues around a target site by positioning of the backbone angles. The ability to not only randomize amino acids but also give more freedom to the angle at which they can interact with the target molecule is perhaps equally important. But how might this be remedied? Painstaking library design could in principle achieve this, by designing oligonucleotides that contain both randomization but also codon insertions and deletions. However this is problematic, as it greatly increases library design and complexity of the pool, as well as, financially being more expensive to construct. There are possible solutions to the practical synthesis of such complex libraries, which would not only randomize amino acids but positions. Oligonucletide libraries that are constructed on chips, either through electrochemical or photochemical synthesis protocols (Pike et al., 2002; Singh-Gasson et al., 1999), can create complex libraries not easily constructed from more traditional synthesis approaches. This is possible because each synthesized oligonucleotide is specified by inputting a file of sequences. Each synthesis location (~100,000) can be fully programmed to create an

oligonucleotide of exact sequence. While this is much more limited than traditional randomization, which can cover libraries roughly 1,000-fold bigger, it enables more complex libraries (such as indel libraries) to be created. Computer modeling and design of libraries can eliminate amino acids that will almost certainly not work in most binding pockets (e.g. tryptophan or proline) and redundancy of amino acid combinations found in NNS randomization.

ALTERING THE FUNCTIONALITY OF THE FINAL COMPONENTS OF THE CENTRAL DOGMA

Some of the core features of the central dogma have been engineered as presented in this dissertation (Fig. C.5). In Chapter 1, the genetic code was altered not only by reassignment of codons (shifting the amber codon from a stop codon to a tryptophan or tryptophan analog), but also by the incorporation of a nonstandard amino acid, 5-hydroxy-L-tryptophan. In both Chapters 1 and 2, the regulation of transcription was engineered by selecting the T7 RNA polymerase, as well as, the tryptophan repressor. In Chapter 3, the first proofreading reverse transcriptase was created from an Archaeal DNA polymerase. This polymerase can convert RNA to DNA, DNA to DNA, and even O-methyl RNA to DNA.

These efforts are not only useful research tools for other scientists in the field, but also demonstrate a more profound concept – even the oldest and most evolutionarily conserved cellular processes can be vastly altered given the right evolutionary paradigm. The components of the central dogma are plastic and can be altered to give wildly different functions. This may be surprising or unsurprising – of course biological

complexes could have different function! Researchers have shown this time and time again, with a wide array of molecules. Even molecules without much chemical diversity, like nucleotide polymers, can be coaxed into performing all sorts of functions. Perhaps the ability to engineer the central dogma is surprising simply because it has been quite difficult. It is so deeply entrenched in the biology, that finding selection conditions to allow probing of its components is especially tricky.

While headway has been made herein this dissertation to engineering the central dogma, several components were not covered. While T7 RNA polymerase was engineered to recognize different promoter sequences, this would not technically be an alteration of the genetic code. A true modification of the genetic code relies on altering the informational flow by changing the molecular composition of the RNA molecule. While there are a number of known modifications known to alter T7 RNA polymerase substrate specificity, the high throughput search through sequence space has been hampered by the lack of an *in vitro* or *in vivo* selection for incorporation of nucleotide triphosphate analogs. RNA polymerases that can incorporate modified nucleotides efficiently would have broad implications, as they are useful for therapeutic drug development. But more interestingly, they do not require a primer to initiate polymerization (like the discussed DNA polymerases in Chapter 3) which gives them the unique ability to operate *in vivo* given a cognate promoter sequence.

The last feature of the central dogma, and likely the toughest challenge, is the reengineering of the ribosome. The ribosome is so complicated to engineer because, unlike other components, it is incredibly integrated. The biosynthesis pathway is highly

regulated at every level: post-transcriptional modifications to RNA, many associated and integral proteins, translation being tightly regulated by a host of enzymes, etc. (Kaczanowska and Ryden-Aulin, 2007). Minor alterations can have significant effects, as so many parts come together to form the whole. Despite this complexity, some headway has been made towards engineering a ribosome that can utilize alternative chemistries. Prokaryotic ribosomes have been engineered with various anti-Shine Dalgarno sequences, allowing binding and translation initiation of orthogonal RBS sequences (Rackham and Chin, 2005). In recent work, a "tethered" ribosome was created by a fusion of the 16S and 23S subunits – making the large and small subunits a single contiguous transcript (Orelle et al., 2015). This is significant because unlike orthogonal ribosomes (altered 16S anti-RBS) this allows modifications to the large subunit which contains the peptidyl transferase active site. Mutations to the tRNA entry site or the peptidyl transferase would almost certainly be lethal to *E. coli* cells, but the ability to have an orthogonal ribosome working in parallel with the host ribosome would allow modifications and probing of the function of the ribosome. This could enable mutations that increase unnatural amino acid incorporation, or even allow the polymerization of functional groups other than carboxyl-amine bonds.

As integral parts of the central dogma are reengineered with novel functionalities, and tools are developed in parallel to further modify them – there will be a push towards creating life forms that have an entirely orthogonal central dogma. The hereditary polymer, the messenger, and the protein components will be distinct from life as it currently is. The chemically orthogonal nature of these molecules, and perhaps

organisms, can allow the creation of diagnostic and therapeutic tools that are beyond the current scope of biology.

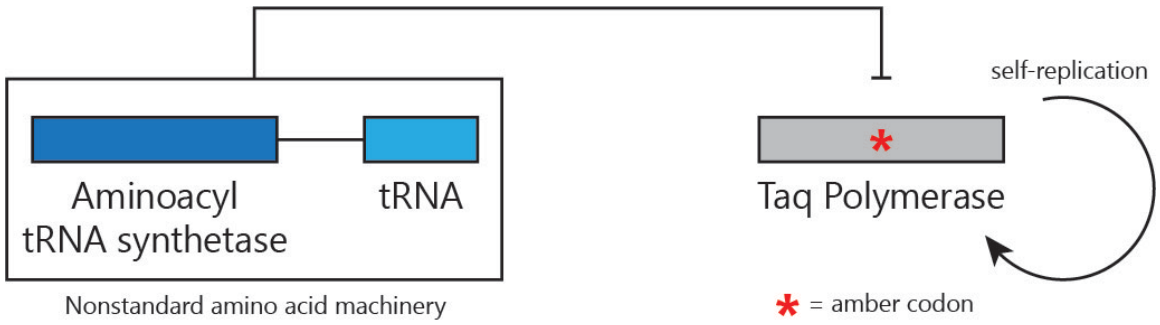


Figure C.1 : First representation of the CPR scheme

The CPR scheme was originally intended to explore the functional consequences of an expanded genetic code on polymerase evolution. However, the diagram also depicted a scheme for the general directed evolution of biomolecules influence the polymerase expression.

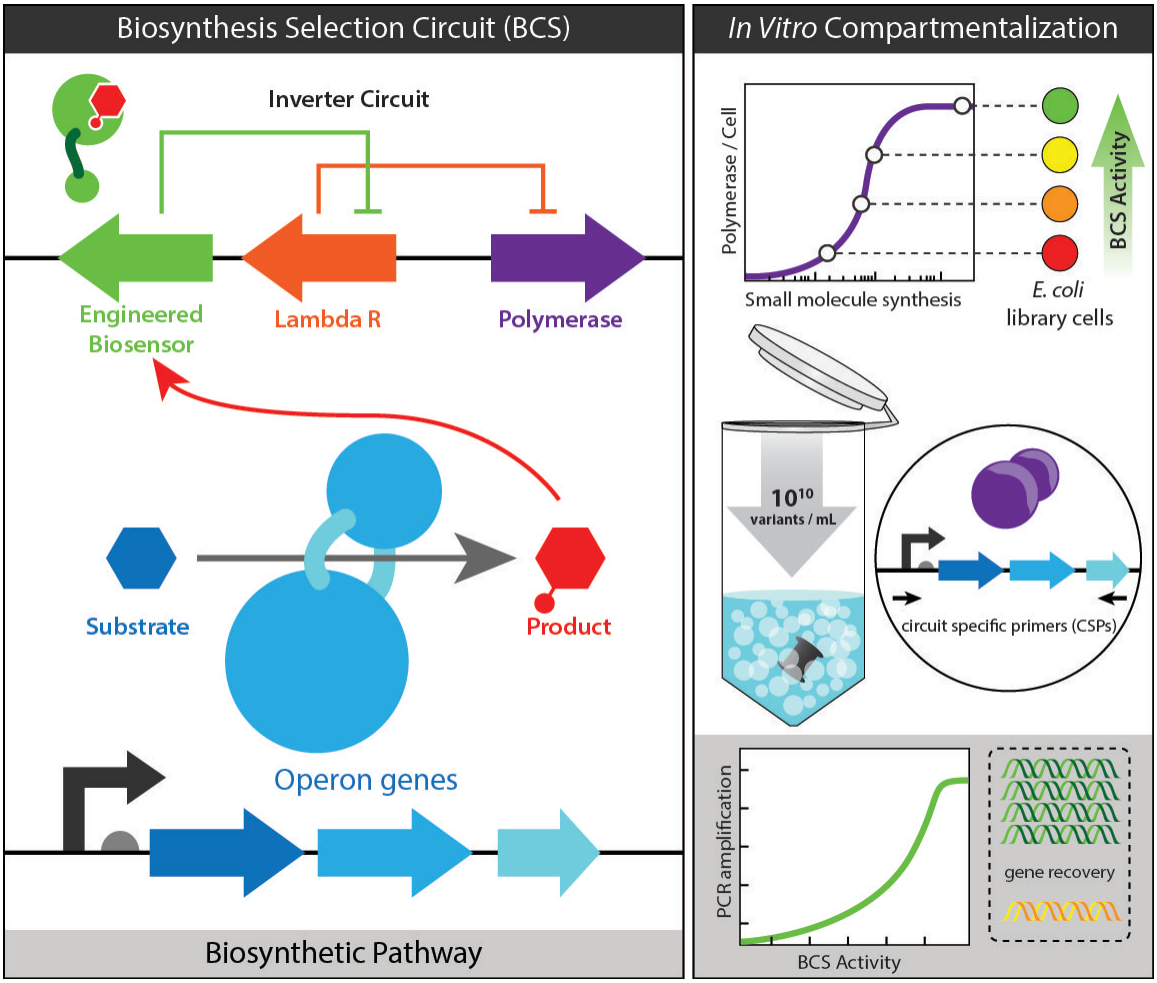


Figure C.2 : Evolved biosensors can be used for the CPR evolution of metabolism

The directed evolution of a biosensor enables CPR selection based on the intracellular concentration of the signal molecule. This can be used for the evolution of single enzymes or entire biosynthetic operons.

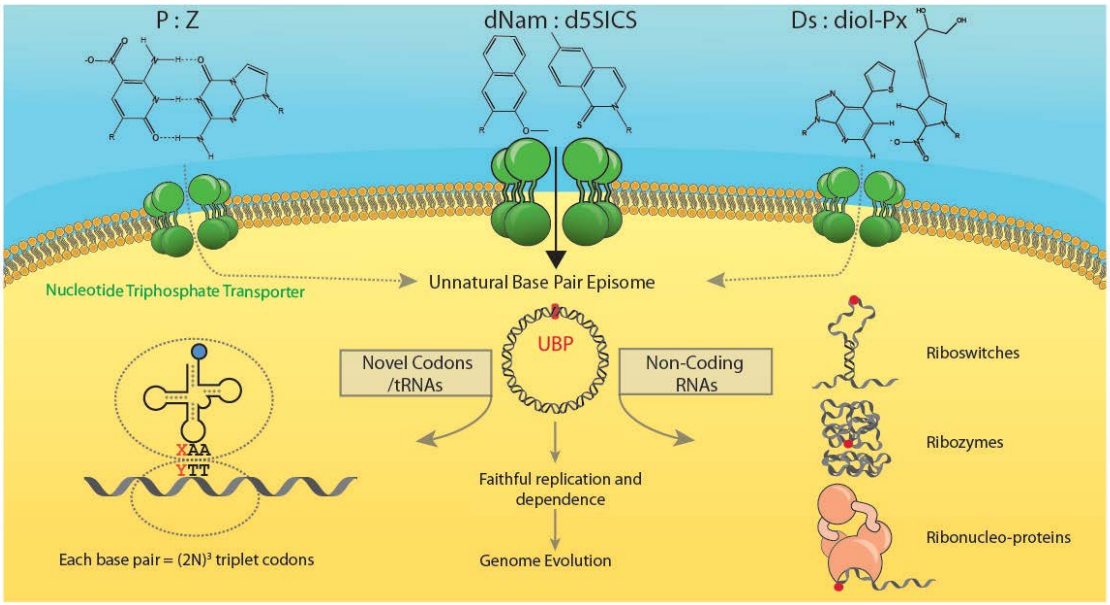


Figure C.3 : Unnatural base pairs expand cellular functions

Unnatural base pair (UBP) systems have the potential to expand a number of cellular systems. Several UBP systems have been described which can potentially be transported into the cell and replicated via endogenous machinery. Transcription of the UBPs could enable a number of expanded functions such as novel tRNA:codon pairs or ribozymes.

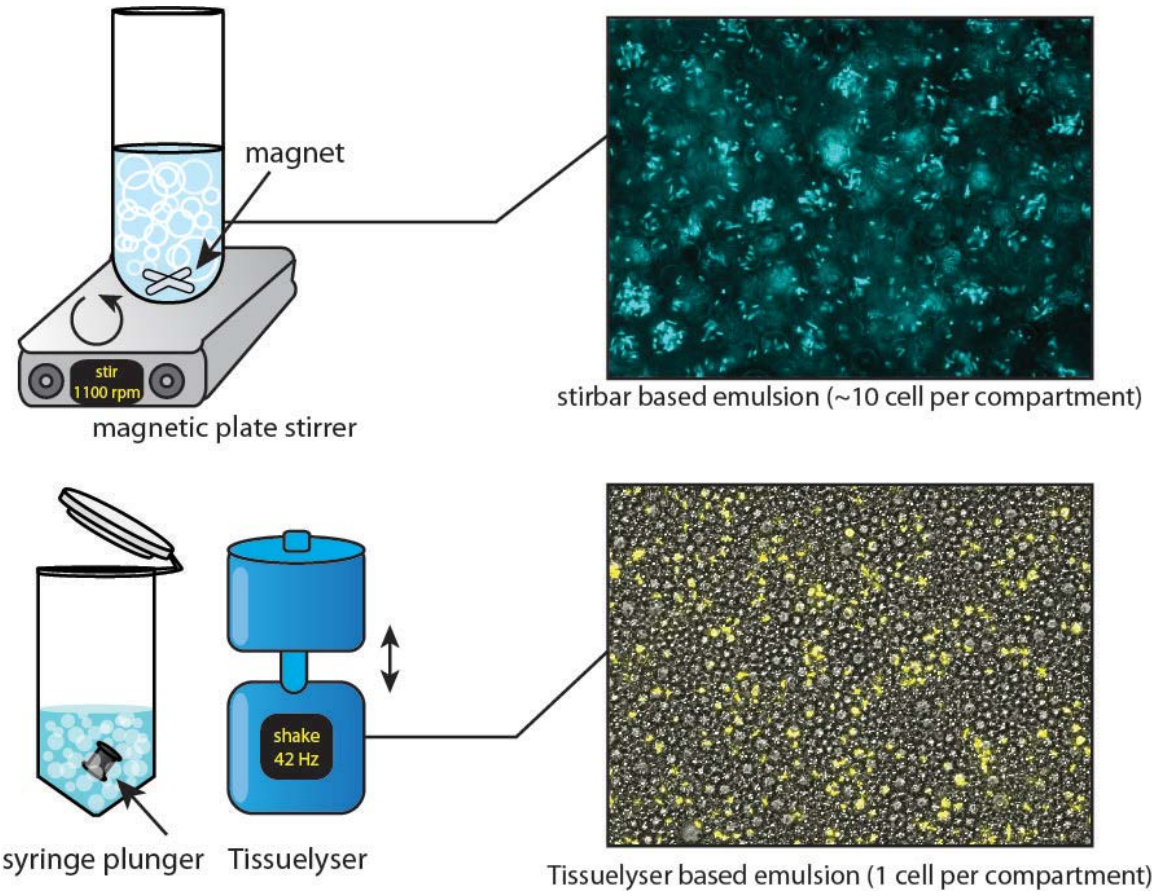


Figure C.4 : Optimization of emulsion setup

Initial emulsions were created by the shear forces generated by a spinning magnetic bar. These resulted in emulsion bubbles of various sizes which contained a number of bacteria. The emulsion parameters were altered to use a tissuelyser, which improved homogeneity and selective enrichment.

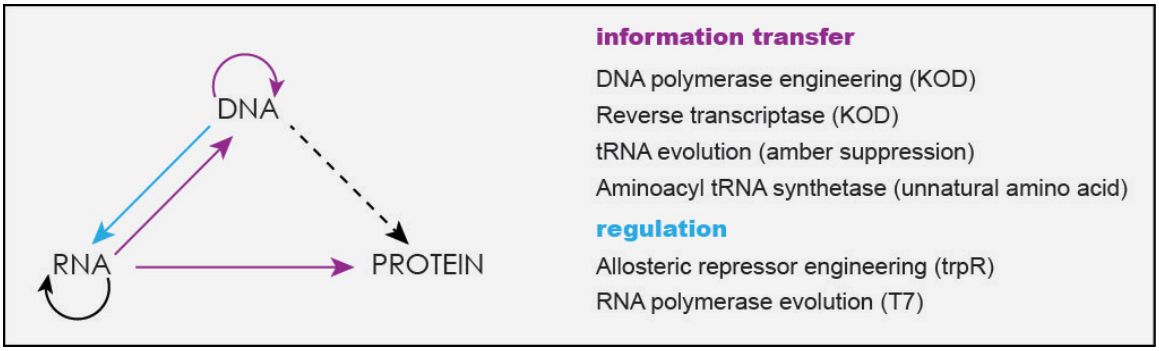


Figure C.5 : Evolution of the central dogma using emulsion based directed evolution

The methods and techniques developed in this dissertation have allowed the engineering of many of the information transfer steps in the central dogma.

References

d'Abbadie, M., Hofreiter, M., Vaisman, A., Loakes, D., Gasparutto, D., Cadet, J., Woodgate, R., Pääbo, S., and Holliger, P. (2007). Molecular breeding of polymerases for amplification of ancient DNA. Nat. Biotechnol. *25*, 939–943.

Agresti, J.J., Kelly, B.T., Jaschke, A., and Griffiths, A.D. (2005). Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using in vitro compartmentalization. Proc. Natl. Acad. Sci. *102*, 16170–16175.

Alonso, N., Guillen, R., Chambers, J.W., and Leng, F. (2015). A rapid and sensitive high-throughput screening method to identify compounds targeting protein-nucleic acids interactions. Nucleic Acids Res. *43*, e52–e52.

Arvidson, D.N., Pfau, J., Hatt, J.K., Shapiro, M., Pecoraro, F.S., and Youderian, P. (1993). Tryptophan super-repressors with alanine 77 changes. J. Biol. Chem. *268*, 4362–4369.

Attwater, J., Wochner, A., Pinheiro, V.B., Coulson, A., and Holliger, P. (2010). Ice as a protocellular medium for RNA replication. Nat. Commun. *1*, 1–8.

Attwater, J., Wochner, A., and Holliger, P. (2013). In-ice evolution of RNA polymerase ribozyme activity. Nat. Chem. *5*, 1011–1018.

Baar, C., d'Abbadie, M., Vaisman, A., Arana, M.E., Hofreiter, M., Woodgate, R., Kunkel, T.A., and Holliger, P. (2011). Molecular breeding of polymerases for resistance to environmental inhibitors. Nucleic Acids Res. *39*, e51.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature *226*, 1209–1211.

Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H., and Weiss, R. (2005). A synthetic multicellular system for programmed pattern formation. Nature *434*, 1130–1134.

Berger, M., Wu, Y., Ogawa, A.K., McMinn, D.L., Schultz, P.G., and Romesberg, F.E. (2000). Universal bases for hybridization, replication and chain termination. Nucleic Acids Res. *28*, 2911–2914.

Bernath, K., Hai, M., Mastrobattista, E., Griffiths, A.D., Magdassi, S., and Tawfik, D.S. (2004). In vitro compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. Anal. Biochem. *325*, 151–157.

Boeke, J.D., and Stoye, J.P. (1997). Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. In Retroviruses, J.M. Coffin, S.H. Hughes, and H.E. Varmus, eds. (Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press),.

Breaker, R.R., and Joyce, G.F. (1994a). Emergence of a replicating species from an in vitro RNA evolution reaction. Proc. Natl. Acad. Sci. U. S. A. *91*, 6093–6097.

Breaker, R.R., and Joyce, G.F. (1994b). Emergence of a replicating species from an in vitro RNA evolution reaction. Proc. Natl. Acad. Sci. U. S. A. *91*, 6093–6097.

Brieba, L.G., and Sousa, R. (2000). Roles of histidine 784 and tyrosine 639 in ribose discrimination by T7 RNA polymerase. Biochemistry (Mosc.) *39*, 919–923.

Bull, J.J., and Pease, C.M. (1995). Why Is the Polymerase Chain Reaction Resistant to In Vitro Evolution? J. Mol. Evol. 1160–1164.

Bush, J.A., Long, B.H., Catino, J.J., Bradner, W.T., and Tomita, K. (1987). Production and biological activity of rebeccamycin, a novel antitumor agent. J. Antibiot. (Tokyo) *40*, 668–678.

Cahill, P., Foster, K., and Mahan, D.E. (1991). Polymerase chain reaction and Q beta replicase amplification. Clin. Chem. *37*, 1482–1485.

Carlson, E.D., Gan, R., Hodgman, C.E., and Jewett, M.C. (2012). Cell-free protein synthesis: Applications come of age. Biotechnol. Adv. *30*, 1185–1194.

Cech, T.R. (1990). Self-Splicing of Group I Introns. Annu. Rev. Biochem. *59*, 543–568.

Chamberlin, M., McGrath, J., and Waskell, L. (1970). New RNA polymerase from Escherichia coli infected with bacteriophage T7. Nature *228*, 227–231.

Chatterjee, A., Xiao, H., Yang, P.-Y., Soundararajan, G., and Schultz, P.G. (2013). A Tryptophanyl-tRNA Synthetase/tRNA Pair for Unnatural Amino Acid Mutagenesis in E. coli. Angew. Chem. *52*, 5106–5109.

Cheetham, G.M., Jeruzalmi, D., and Steitz, T.A. (1999a). Structural basis for initiation of transcription from an RNA polymerase-promoter complex. Nature *399*, 80–83.

Cheetham, G.M., Jeruzalmi, D., and Steitz, T.A. (1999b). Structural basis for initiation of transcription from an RNA polymerase-promoter complex. Nature *399*, 80–83.

Chelliserrykattil, J., and Ellington, A.D. (2004). Evolution of a T7 RNA polymerase variant that transcribes 2'-O-methyl RNA. Nat. Biotechnol. *22*, 1155–1160.

Chelliserrykattil, J., Cai, G., and Ellington, A.D. (2001). A combined in vitro/in vivo selection for polymerases with novel promoter specificities. BMC Biotechnol. *1*, 13.

Chen, T., and Romesberg, F.E. (2014). Directed polymerase evolution. FEBS Lett. *588*, 219–229.

Chen, J., Sun, S., Li, C.-Z., Zhu, Y.-G., and Rosen, B.P. (2014). Biosensor for organoarsenical herbicides and growth promoters. Environ. Sci. Technol. *48*, 1141–1147.

Chen, Y., Kim, J.K., Hirning, A.J., Josi, K., and Bennett, M.R. (2015). Emergent genetic oscillations in a synthetic microbial consortium. Science *349*, 986–989.

Chin, J.W., Cropp, T.A., Anderson, J.C., Mukherji, M., Zhang, Z., and Schultz, P.G. (2003). An expanded eukaryotic genetic code. Science *301*, 964–967.

Cohen, H.M., Tawfik, D.S., and Griffiths, A.D. (2004). Altering the sequence specificity of HaeIII methyltransferase by directed evolution using in vitro compartmentalization. Protein Eng. Des. Sel. PEDS *17*, 3–11.

Cormack, B.P., Valdivia, R.H., and Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). Gene *173*, 33–38.

Cozens, C., Pinheiro, V.B., Vaisman, A., Woodgate, R., and Holliger, P. (2012). A short adaptive path from DNA to RNA polymerases. Proc. Natl. Acad. Sci. *109*, 8067–8072.

Craig, A.G., Jimenez, E.C., Dykert, J., Nielsen, D.B., Gulyas, J., Abogadie, F.C., Porter, J., Rivier, J.E., Cruz, L.J., Olivera, B.M., et al. (1997). A Novel Post-translational Modification Involving Bromination of Tryptophan IDENTIFICATION OF THE RESIDUE, L-6-BROMOTRYPTOPHAN, IN PEPTIDES FROM Conus imperialis AND Conus radiatus VENOM. J. Biol. Chem. *272*, 4689–4698.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563.

Daniel, R., Rubens, J.R., Sarpeshkar, R., and Lu, T.K. (2013). Synthetic analog computation in living cells. Nature *497*, 619–623.

Darnell, J.E., and Doolittle, W.F. (1986). Speculations on the early course of evolution. Proc. Natl. Acad. Sci. U. S. A. *83*, 1271–1275.

Davidson, E.A., Meyer, A.J., Ellefson, J.W., Levy, M., and Ellington, A.D. (2012a). An *in vitro* Autogene. ACS Synth. Biol. *1*, 190–196.

Davidson, E.A., Meyer, A.J., Ellefson, J.W., Levy, M., and Ellington, A.D. (2012b). An in vitro Autogene. ACS Synth. Biol. *1*, 190–196.

Deamer, D.W., and Dworkin, J.P. (2005). Chemistry and Physics of Primitive Membranes. In Prebiotic Chemistry, P. Walde, ed. (Berlin/Heidelberg: Springer-Verlag), pp. 1–27.

Dedkova, L.M., Fahmi, N.E., Golovine, S.Y., and Hecht, S.M. (2003). Enhanced D-amino acid incorporation into protein by modified ribosomes. J. Am. Chem. Soc. *125*, 6616–6617.

Deiters, A., and Schultz, P.G. (2005). In vivo incorporation of an alkyne into proteins in Escherichia coli. Bioorg. Med. Chem. Lett. *15*, 1521–1524.

DeKosky, B.J., Ippolito, G.C., Deschner, R.P., Lavinder, J.J., Wine, Y., Rawlings, B.M., Varadarajan, N., Giesecke, C., Dörner, T., Andrews, S.F., et al. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat. Biotechnol. *31*, 166–169.

DeKosky, B.J., Kojima, T., Rodin, A., Charab, W., Ippolito, G.C., Ellington, A.D., and Georgiou, G. (2015). In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat. Med. *21*, 86–91.

Dellus-Gur, E., Toth-Petroczy, A., Elias, M., and Tawfik, D.S. (2013). What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. J. Mol. Biol. *425*, 2609–2621.

Dickinson, B.C., Leconte, A.M., Allen, B., Esvelt, K.M., and Liu, D.R. (2013). Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. Proc. Natl. Acad. Sci. U. S. A.

Doi, N., and Yanagawa, H. (1999). Design of generic biosensors based on green fluorescent proteins with allosteric sites by directed evolution. FEBS Lett. *453*, 305–307.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc. Natl. Acad. Sci. *100*, 8817–8822.

Edwards, H., and Schimmel, P. (1990). A bacterial amber suppressor in Saccharomyces cerevisiae is selectively recognized by a bacterial aminoacyl-tRNA synthetase. Mol. Cell. Biol. *10*, 1633–1641.

Efange, S.M.N., Michelson, R.H., Remmel, R.P., Boudreau, R.J., Dutta, A.K., and Freshler, A. (1990). Flexible N-methyl-4-phenyl-1, 2, 3, 6-tetrahydropyridine analog: synthesis and monoamine oxidase catalyzed bioactivation. J. Med. Chem. *33*, 3133–3138.

Ellefson, J.W., Meyer, A.J., Hughes, R.A., Cannon, J.R., Brodbelt, J.S., and Ellington, A.D. (2014). Directed evolution of genetic parts and circuits by compartmentalized partnered replication. Nat. Biotechnol. *32*, 97–101.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011a). A system for the continuous directed evolution of biomolecules. Nature *472*, 499–503.

Esvelt, K.M., Carlson, J.C., and Liu, D.R. (2011b). A system for the continuous directed evolution of biomolecules. Nature *472*, 499–503.

Fa, M., Radeghieri, A., Henry, A.A., and Romesberg, F.E. (2004). Expanding the Substrate Repertoire of a DNA Polymerase by Directed Evolution. J. Am. Chem. Soc. *126*, 1748–1754.

Fidalgo da Silva, E., and Reha-Krantz, L.J. (2007). DNA polymerase proofreading: active site switching catalyzed by the bacteriophage T4 DNA polymerase. Nucleic Acids Res. *35*, 5452–5463.

Fogg, M.J., Pearl, L.H., and Connolly, B.A. (2002). Structural basis for uracil recognition by archaeal family B DNA polymerases. Nat. Struct. Biol. *9*, 922–927.

Gardner, A.F., and Jack, W.E. (1999). Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. Nucleic Acids Res. *27*, 2545–2553.

Ghadessy, F.J., and Holliger, P. (2004). A novel emulsion mixture for in vitro compartmentalization of transcription and translation in the rabbit reticulocyte system. Protein Eng. Des. Sel. PEDS *17*, 201–204.

Ghadessy, F.J., and Holliger, P. (2007). Compartmentalized self-replication: a novel method for the directed evolution of polymerases and other enzymes. Methods Mol. Biol. *352*, 237–248.

Ghadessy, F.J., Ong, J.L., and Holliger, P. (2001a). Directed evolution of polymerase function by compartmentalized self-replication. Proc. Natl. Acad. Sci. *98*, 4552–4557.

Ghadessy, F.J., Ong, J.L., and Holliger, P. (2001b). Directed evolution of polymerase function by compartmentalized self-replication. Proc. Natl. Acad. Sci. U. S. A. *98*, 4552–4557.

Ghadessy, F.J., Ramsay, N., Boudsocq, F., Loakes, D., Brown, A., Iwai, S., Vaisman, A., Woodgate, R., and Holliger, P. (2004). Generic expansion of the substrate spectrum of a DNA polymerase by directed evolution. Nat. Biotechnol. *22*, 755–759.

Gibson, D.G. (2011). Enzymatic assembly of overlapping DNA fragments. Methods Enzymol. *498*, 349–361.

Giorgetti, L., Siggers, T., Tiana, G., Caprara, G., Notarbartolo, S., Corona, T., Pasparakis, M., Milani, P., Bulyk, M.L., and Natoli, G. (2010). Noncooperative Interactions between Transcription Factors and Clustered DNA Binding Sites Enable Graded Transcriptional Responses to Environmental Inputs. Mol. Cell *37*, 418–428.

Glasscock, C.J., Lucks, J.B., and DeLisa, M.P. (2016). Engineered Protein Machines: Emergent Tools for Synthetic Biology. Cell Chem. Biol. *23*, 45–56.

Goldsmith, M., and Tawfik, D.S. (2009). Potential role of phenotypic mutations in the evolution of protein expression and stability. Proc. Natl. Acad. Sci. U. S. A. *106*, 6197–6202.

Gout, J.-F., Thomas, W.K., Smith, Z., Okamoto, K., and Lynch, M. (2013). Large-scale detection of in vivo transcription errors. Proc. Natl. Acad. Sci. *110*, 18584–18589.

Greagg, M.A., Fogg, M.J., Panayotou, G., Evans, S.J., Connolly, B.A., and Pearl, L.H. (1999). A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. Proc. Natl. Acad. Sci. U. S. A. *96*, 9045–9050.

Greiss, S., and Chin, J.W. (2011). Expanding the Genetic Code of an Animal. J. Am. Chem. Soc. *133*, 14196–14199.

Griffiths, A.D., and Tawfik, D.S. (2003). Directed evolution of an extremely fast phosphotriesterase by in vitro compartmentalization. EMBO J. *22*, 24–35.

Hammerling, M.J., Ellefson, J.W., Boutz, D.R., Marcotte, E.M., Ellington, A.D., and Barrick, J.E. (2014). Bacteriophages use an expanded genetic code on evolutionary paths to higher fitness. Nat. Chem. Biol. *10*, 178–180.

Haruna, I., Nozu, K., Ohtaka, Y., and Spiegelman, S. (1963). AN RNA “REPLICASE” INDUCED BY AND SELECTIVE FOR A VIRAL RNA: ISOLATION AND PROPERTIES. Proc. Natl. Acad. Sci. U. S. A. *50*, 905–911.

Hilser, V.J., Wrabl, J.O., and Motlagh, H.N. (2012). Structural and Energetic Basis of Allostery. Annu. Rev. Biophys. *41*, 585–609.

Hino, N., Okazaki, Y., Kobayashi, T., Hayashi, A., Sakamoto, K., and Yokoyama, S. (2005). Protein photo-cross-linking in mammalian cells by site-specific incorporation of a photoreactive amino acid. Nat. Methods *2*, 201–206.

Hirao, I., Kimoto, M., Mitsui, T., Fujiwara, T., Kawai, R., Sato, A., Harada, Y., and Yokoyama, S. (2006). An unnatural hydrophobic base pair system: site-specific incorporation of nucleotide analogs into DNA and RNA. Nat. Methods *3*, 729–735.

Holmberg, R.C., Henry, A.A., and Romesberg, F.E. (2005). Directed evolution of novel polymerases. Biomol. Eng. *22*, 39–49.

Hooshangi, S., Thiberge, S., and Weiss, R. (2005). Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. Proc. Natl. Acad. Sci. *102*, 3581–3586.

Huff, J.W., Sastry, K.S., Gordon, M.P., and Wacker, W.E.C. (1964). The Action of Metal Ions on Tobacco Mosaic Virus Ribonucleic Acid *. Biochemistry (Mosc.) *3*, 501–506.

Hughes, A.L. (1994). The Evolution of Functionally Novel Proteins after Gene Duplication. Proc. R. Soc. B Biol. Sci. *256*, 119–124.

Hughes, R.A., and Ellington, A.D. (2010). Rational design of an orthogonal tryptophanyl nonsense suppressor tRNA. Nucleic Acids Res. *38*, 6813–6830.

Hughes, R.A., Miklos, A.E., and Ellington, A.D. (2011). Gene synthesis: methods and applications. Methods Enzymol. *498*, 277–309.

Hurlburt, B.K., and Yanofsky, C. (1992). trp repressor/trp operator interaction. Equilibrium and kinetic analysis of complex formation and stability. J. Biol. Chem. *267*, 16783–16789.

Ichida, J.K. (2005). High fidelity TNA synthesis by Therminator polymerase. Nucleic Acids Res. *33*, 5219–5225.

Ichihashi, N., Usui, K., Kazuta, Y., Sunami, T., Matsuura, T., and Yomo, T. (2013). Darwinian evolution in a translation-coupled RNA replication system within a cell-like compartment. Nat. Commun. *4*, 2494.

Isaacs, F.J., Carr, P.A., Wang, H.H., Lajoie, M.J., Sterling, B., Kraal, L., Tolonen, A.C., Gianoulis, T.A., Goodman, D.B., Reppas, N.B., et al. (2011). Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. Science *333*, 348–353.

Jestin, J.L., Kristensen, P., and Winter, G. (1999). A method for the selection of catalytic activity using phage display and proximity coupling. Angew. Chem. Int. Ed Engl. *38*, 1124–1127.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science *337*, 816–821.

Johnson, D.B.F., Xu, J., Shen, Z., Takimoto, J.K., Schultz, M.D., Schmitz, R.J., Xiang, Z., Ecker, J.R., Briggs, S.P., and Wang, L. (2011a). RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. Nat. Chem. Biol. *7*, 779–786.

Johnson, D.B.F., Xu, J., Shen, Z., Takimoto, J.K., Schultz, M.D., Schmitz, R.J., Xiang, Z., Ecker, J.R., Briggs, S.P., and Wang, L. (2011b). RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. Nat. Chem. Biol. *7*, 779–786.

Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasner, M.E., and Bartel, D.P. (2001). RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science *292*, 1319–1325.

Joyce, C.M., and Steitz, T.A. (1995). Polymerase structures and function: variations on a theme? J. Bacteriol. *177*, 6321–6329.

Jozwiakowski, S.K., and Connolly, B.A. (2011). A Modified Family-B Archaeal DNA Polymerase with Reverse Transcriptase Activity. ChemBioChem *12*, 35–37.

Kaczanowska, M., and Ryden-Aulin, M. (2007). Ribosome Biogenesis and the Translation Process in Escherichia coli. Microbiol. Mol. Biol. Rev. *71*, 477–494.

Kast, P., and Hennecke, H. (1991). Amino acid substrate specificity of Escherichia coli phenylalanyl-tRNA synthetase altered by distinct mutations. J. Mol. Biol. *222*, 99–124.

Kiel, C., Yus, E., and Serrano, L. (2010). Engineering Signal Transduction Pathways. Cell *140*, 33–47.

Kimoto, M., Yamashige, R., Matsunaga, K., Yokoyama, S., and Hirao, I. (2013). Generation of high-affinity DNA aptamers using an expanded genetic alphabet. Nat. Biotechnol. *31*, 453–457.

Kitabayashi, M., Nishiya, Y., Esaka, M., Itakura, M., and Imanaka, T. (2002). Gene Cloning and Polymerase Chain Reaction with Proliferating Cell Nuclear Antigen from *Thermococcus kodakaraensis* KOD1. Biosci. Biotechnol. Biochem. *66*, 2194–2200.

Klig, L.S., Carey, J., and Yanofsky, C. (1988). trp Repressor interactions with the trparoH and trpR operators. J. Mol. Biol. *202*, 769–777.

Kramer, E.B., and Farabaugh, P.J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA N. Y. N *13*, 87–96.

Kunkel, T.A., and Bebenek, K. (2000). DNA Replication Fidelity *. Annu. Rev. Biochem. *69*, 497–529.

Lajoie, M.J., Rovner, A.J., Goodman, D.B., Aerni, H.-R., Haimovich, A.D., Kuznetsov, G., Mercer, J.A., Wang, H.H., Carr, P.A., Mosberg, J.A., et al. (2013). Genomically Recoded Organisms Expand Biological Functions. Science *342*, 357–360.

Lavergne, T., Degardin, M., Malyshev, D.A., Quach, H.T., Dhami, K., Ordoukhanian, P., and Romesberg, F.E. (2013). Expanding the Scope of Replicable Unnatural DNA: Stepwise Optimization of a Predominantly Hydrophobic Base Pair. J. Am. Chem. Soc. *135*, 5408–5419.

Lawrence, M.S., Phillips, K.J., and Liu, D.R. (2007). Supercharging Proteins Can Impart Unusual Resilience. J. Am. Chem. Soc. *129*, 10110–10112.

Lee, Y.-F., Tawfik, D.S., and Griffiths, A.D. (2002). Investigating the target recognition of DNA cytosine-5 methyltransferase HhaI by library selection using in vitro compartmentalisation. Nucleic Acids Res. *30*, 4937–4944.

Lehman, I.R., Bessman, M.J., Simms, E.S., and Kornberg, A. (1958). Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. J. Biol. Chem. *233*, 163–170.

Leman, A.R., and Noguchi, E. (2013). The replication fork: understanding the eukaryotic replication machinery and the challenges to genome duplication. Genes *4*, 1–32.

Levy, M., and Ellington, A.D. (2008). Directed evolution of streptavidin variants using in vitro compartmentalization. Chem. Biol. *15*, 979–989.

Levy, M., Griswold, K.E., and Ellington, A.D. (2005). Direct selection of trans-acting ligase ribozymes by in vitro compartmentalization. RNA N. Y. N *11*, 1555–1562.

Lincoln, T.A., and Joyce, G.F. (2009). Self-Sustained Replication of an RNA Enzyme. Science *323*, 1229–1232.

Liu, C.C., and Schultz, P.G. (2010). Adding New Chemistries to the Genetic Code. Annu. Rev. Biochem. *79*, 413–444.

Liu, W., Brock, A., Chen, S., Chen, S., and Schultz, P.G. (2007). Genetic incorporation of unnatural amino acids into proteins in mammalian cells. Nat. Methods *4*, 239–244.

Loakes, D., Gallego, J., Pinheiro, V.B., Kool, E.T., and Holliger, P. (2009). Evolving a polymerase for hydrophobic base analogues. J. Am. Chem. Soc. *131*, 14827–14837.

Loftfield, R.B., and Vanderjagt, D. (1972). The frequency of errors in protein biosynthesis. Biochem. J. *128*, 1353–1356.

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. Nat. Rev. Genet. *9*, 583–593.

Lu, W.-C., and Ellington, A.D. (2014). Design and selection of a synthetic operon. ACS Synth. Biol. *3*, 410–415.

Lundberg, K.S., Shoemaker, D.D., Adams, M.W.W., Short, J.M., Sorge, J.A., and Mathur, E.J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. Gene *108*, 1–6.

Lutz, S., Burgstaller, P., and Benner, S.A. (1999). An in vitro screening technique for DNA polymerases that can incorporate modified nucleotides. Pseudo-thymidine as a substrate for thermostable polymerases. Nucleic Acids Res. *27*, 2792–2798.

Madan Babu, M. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res. *31*, 1234–1244.

Malyshev, D.A., and Romesberg, F.E. (2015). The expanded genetic alphabet. Angew. Chem. Int. Ed Engl. *54*, 11930–11944.

Malyshev, D.A., Dhami, K., Quach, H.T., Lavergne, T., Ordoukhanian, P., Torkamani, A., and Romesberg, F.E. (2012). Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. Proc. Natl. Acad. Sci. *109*, 12005–12010.

Malyshev, D.A., Dhami, K., Lavergne, T., Chen, T., Dai, N., Foster, J.M., Corrêa, I.R., and Romesberg, F.E. (2014). A semi-synthetic organism with an expanded genetic alphabet. Nature *509*, 385–388.

Mandell, D.J., Lajoie, M.J., Mee, M.T., Takeuchi, R., Kuznetsov, G., Norville, J.E., Gregg, C.J., Stoddard, B.L., and Church, G.M. (2015). Biocontainment of genetically modified organisms by synthetic protein design. Nature *518*, 55–60.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature.

McAdams, H.H., Srinivasan, B., and Arkin, A.P. (2004). The evolution of genetic regulatory systems in bacteria. Nat. Rev. Genet. *5*, 169–178.

Meselson, M., and Stahl, F.W. (1958). THE REPLICATION OF DNA IN ESCHERICHIA COLI. Proc. Natl. Acad. Sci. U. S. A. *44*, 671–682.

Meyer, A.J., Ellefson, J.W., and Ellington, A.D. (2012). Abiotic Self-Replication. Acc. Chem. Res. *45*, 2097–2105.

Meyer, A.J., Ellefson, J.W., and Ellington, A.D. (2015). Directed Evolution of a Panel of Orthogonal T7 RNA Polymerase Variants for in Vivo or in Vitro Synthetic Circuitry. ACS Synth. Biol. *4*, 1070–1076.

Meyerhans, A., Cheynier, R., Albert, J., Seth, M., Kwok, S., Sninsky, J., Morfeldt-Månson, L., Asjö, B., and Wain-Hobson, S. (1989). Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. Cell *58*, 901–910.

Milligan, J.F., Groebe, D.R., Witherell, G.W., and Uhlenbeck, O.C. (1987). Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates. Nucleic Acids Res. *15*, 8783–8798.

Mills, D.R., Peterson, R.L., and Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc. Natl. Acad. Sci. U. S. A. *58*, 217–224.

Miño, G., Baez, M., and Gutierrez, G. (2013). Effect of mutation at the interface of Trp-repressor dimeric protein: a steered molecular dynamics simulation. Eur. Biophys. J. *42*, 683–690.

Moon, T.S., Lou, C., Tamsir, A., Stanton, B.C., and Voigt, C.A. (2012). Genetic programs constructed from layered logic gates in single cells. Nature *491*, 249–253.

Moser, M.J., DiFrancesco, R.A., Gowda, K., Klingele, A.J., Sugar, D.R., Stocki, S., Mead, D.A., and Schoenfeld, T.W. (2012). Thermostable DNA Polymerase from a Viral Metagenome Is a Potent RT-PCR Enzyme. PLoS ONE *7*, e38371.

Motlagh, H.N., Wrabl, J.O., Li, J., and Hilser, V.J. (2014). The ensemble nature of allostery. Nature *508*, 331–339.

Mukai, T., Hayashi, A., Iraha, F., Sato, A., Ohtake, K., Yokoyama, S., and Sakamoto, K. (2010). Codon reassignment in the Escherichia coli genetic code. Nucleic Acids Res. *38*, 8188–8195.

Mullis, K.B., and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol. *155*, 335–350.

de Nadal, E., Ammerer, G., and Posas, F. (2011). Controlling gene expression in response to stress. Nat. Rev. Genet.

Nakamura, T.M. (1997). Telomerase Catalytic Subunit Homologs from Fission Yeast and Human. Science *277*, 955–959.

Nakano, M., Nakai, N., Kurita, H., Komatsu, J., Takashima, K., Katsura, S., and Mizuno, A. (2005). Single-molecule reverse transcription polymerase chain reaction using water-in-oil emulsion. J. Biosci. Bioeng. *99*, 293–295.

Neumann, H., Hancock, S.M., Buning, R., Routh, A., Chapman, L., Somers, J., Owen-Hughes, T., van Noort, J., Rhodes, D., and Chin, J.W. (2009). A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. Mol. Cell *36*, 153–163.

Nielsen, A.A., Segall-Shapiro, T.H., and Voigt, C.A. (2013). Advances in genetic circuit design: novel biochemistries, deep part mining, and precision gene expression. Curr. Opin. Chem. Biol. *17*, 878–892.

Nishioka, M., Mizuguchi, H., Fujiwara, S., Komatsubara, S., Kitabayashi, M., Uemura, H., Takagi, M., and Imanaka, T. (2001). Long and accurate PCR with a mixture of KOD DNA polymerase and its exonuclease deficient mutant enzyme. J. Biotechnol. *88*, 141–149.

Ohno, S., Wolf, U., and Atkin, N.B. (1968). Evolution from fish to mammals by gene duplication. Hereditas *59*, 169–187.

Ohtake, K., Yamaguchi, A., Mukai, T., Kashimura, H., Hirano, N., Haruki, M., Kohashi, S., Yamagishi, K., Murayama, K., Tomabechi, Y., et al. (2015). Protein stabilization utilizing a redefined codon. Sci. Rep. *5*, 9762.

Ong, J.L., Loakes, D., Jaroslawski, S., Too, K., and Holliger, P. (2006). Directed evolution of DNA polymerase, RNA polymerase and reverse transcriptase activity in a single polypeptide. J. Mol. Biol. *361*, 537–550.

Orelle, C., Carlson, E.D., Szal, T., Florin, T., Jewett, M.C., and Mankin, A.S. (2015). Protein synthesis by ribosomes with tethered subunits. Nature *524*, 119–124.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335*, 321–329.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Packer, M.S., and Liu, D.R. (2015). Methods for the directed evolution of proteins. Nat. Rev. Genet. *16*, 379–394.

Paddon, C.J., and Keasling, J.D. (2014). Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. Nat. Rev. Microbiol. *12*, 355–367.

Paige, J.S., Wu, K.Y., and Jaffrey, S.R. (2011). RNA Mimics of Green Fluorescent Protein. Science *333*, 642–646.

Pelham, H.R., and Jackson, R.J. (1976). An efficient mRNA-dependent translation system from reticulocyte lysates. Eur. J. Biochem. FEBS *67*, 247–256.

Pike, A.R., Lie, L.H., Eagling, R.A., Ryder, L.C., Patole, S.N., Connolly, B.A., Horrocks, B.R., and Houlton, A. (2002). DNA On Silicon Devices: On-Chip Synthesis, Hybridization, and Charge Transfer. Angew. Chem. Int. Ed. *41*, 615–617.

Pinheiro, V.B., Taylor, A.I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J.C., Wengel, J., Peak-Chew, S.-Y., McLaughlin, S.H., et al. (2012). Synthetic Genetic Polymers Capable of Heredity and Evolution. Science *336*, 341–344.

Popovych, N., Tzeng, S.-R., Tonelli, M., Ebright, R.H., and Kalodimos, C.G. (2009). Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. Proc. Natl. Acad. Sci. *106*, 6927–6932.

Porcar, M. (2010). Beyond directed evolution: Darwinian selection as a tool for synthetic biology. Syst. Synth. Biol. *4*, 1–6.

Prindle, A., Selimkhanov, J., Li, H., Razinkov, I., Tsimring, L.S., and Hasty, J. (2014). Rapid and tunable post-translational coupling of genetic circuits. Nature *508*, 387–391.

Rackham, O., and Chin, J.W. (2005). A network of orthogonal ribosome x mRNA pairs. Nat. Chem. Biol. *1*, 159–166.

Ramakrishnan, V. (2002). Ribosome Structure and the Mechanism of Translation. Cell *108*, 557–572.

Raman, S., Rogers, J.K., Taylor, N.D., and Church, G.M. (2014). Evolution-guided optimization of biosynthetic pathways. Proc. Natl. Acad. Sci. U. S. A. *111*, 17803–17808.

Ramsay, N., Jemth, A.-S., Brown, A., Crampton, N., Dear, P., and Holliger, P. (2010). CyDNA: synthesis and replication of highly Cy-dye substituted DNA by an evolved polymerase. J. Am. Chem. Soc. *132*, 5096–5104.

Raskin, C.A., Diaz, G.A., and McAllister, W.T. (1993). T7 RNA polymerase mutants with altered promoter specificities. Proc. Natl. Acad. Sci. U. S. A. *90*, 3147–3151.

Reichheld, S.E., Yu, Z., and Davidson, A.R. (2009). The induction of folding cooperativity by ligand binding drives the allosteric response of tetracycline repressor. Proc. Natl. Acad. Sci. *106*, 22263–22268.

Rhodius, V.A., Segall-Shapiro, T.H., Sharon, B.D., Ghodasara, A., Orlova, E., Tabakh, H., Burkhardt, D.H., Clancy, K., Peterson, T.C., Gross, C.A., et al. (2014). Design of orthogonal genetic switches based on a crosstalk map of s, anti- s, and promoters. Mol. Syst. Biol. *9*, 702–702.

Rogers, J.K., Guzman, C.D., Taylor, N.D., Raman, S., Anderson, K., and Church, G.M. (2015). Synthetic biosensors for precise gene control and real-time monitoring of metabolites. Nucleic Acids Res. *43*, 7648–7660.

Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. *10*, 866–876.

Rosenfeld, N., and Alon, U. (2003). Response delays and the structure of transcription networks. J. Mol. Biol. *329*, 645–654.

Royer, C.A., Mann, C.J., and Matthews, C.R. (1993). Resolution of the fluorescence equilibrium unfolding profile of trp aporepressor using single tryptophan mutants. Protein Sci. *2*, 1844–1852.

Rydén, S.M., and Isaksson, L.A. (1984). A temperature-sensitive mutant of Escherichia coli that shows enhanced misreading of UAG/A and increased efficiency for some tRNA nonsense suppressors. Mol. Gen. Genet. MGG *193*, 38–45.

Saiki, R., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., Mullis, K., and Erlich, H. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science *239*, 487–491.

Sakamoto, K., Murayama, K., Oki, K., Iraha, F., Kato-Murayama, M., Takahashi, M., Ohtake, K., Kobayashi, T., Kuramitsu, S., Shirouzu, M., et al. (2009). Genetic Encoding of 3-Iodo-l-Tyrosine in Escherichia coli for Single-Wavelength Anomalous Dispersion Phasing in Protein Crystallography. Structure *17*, 335–344.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. *109*, 14508–14513.

Schoning, K.-U. (2000). Chemical Etiology of Nucleic Acid Structure: The alpha -Threofuranosyl-(3’rightarrow 2’) Oligonucleotide System. Science *290*, 1347–1351.

Schrader, J.M., Chapman, S.J., and Uhlenbeck, O.C. (2011). Tuning the affinity of aminoacyl-tRNA to elongation factor Tu for optimal decoding. Proc. Natl. Acad. Sci. U. S. A. *108*, 5215–5220.

Schweitzer, B.A., and Kool, E.T. (1995). Hydrophobic, Non-Hydrogen-Bonding Bases and Base Pairs in DNA. J. Am. Chem. Soc. *117*, 1863–1872.

Sczepanski, J.T., and Joyce, G.F. (2014). A cross-chiral RNA polymerase ribozyme. Nature *515*, 440–442.

Sepp, A., Tawfik, D.S., and Griffiths, A.D. (2002). Microbead display by in vitro compartmentalisation: selection for binding using flow cytometry. FEBS Lett. *532*, 455–458.

Shao, X., Hensley, P., and Matthews, C.R. (1997). Construction and characterization of monomeric tryptophan repressor: a model for an early intermediate in the folding of a dimeric protein. Biochemistry (Mosc.) *36*, 9941–9949.

Shaw, J.B., Li, W., Holden, D.D., Zhang, Y., Griep-Raming, J., Fellers, R.T., Early, B.P., Thomas, P.M., Kelleher, N.L., Brodbelt, J.S. (2013). Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. J. Am. Chem. Soc.

Shendure, J. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science *309*, 1728–1732.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. (2001). Cell-free translation reconstituted with purified components. Nat. Biotechnol. *19*, 751–755.

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat. Biotechnol. *17*, 974–978.

Söll, D. (1990). The accuracy of aminoacylation--ensuring the fidelity of the genetic code. Experientia *46*, 1089–1096.

Stanton, B.C., Nielsen, A.A.K., Tamsir, A., Clancy, K., Peterson, T., and Voigt, C.A. (2014). Genomic mining of prokaryotic repressors for orthogonal logic gates. Nat. Chem. Biol. *10*, 99–105.

Steer, B.A., and Schimmel, P. (1999). Domain-domain communication in a miniature archaebacterial tRNA synthetase. Proc. Natl. Acad. Sci. U. S. A. *96*, 13644–13649.

Steitz, T.A. (2008). A structural understanding of the dynamic ribosome machine. Nat. Rev. Mol. Cell Biol. *9*, 242–253.

Szathmáry, E., and Demeter, L. (1987). Group selection of early replicators and the origin of life. J. Theor. Biol. *128*, 463–486.

Tabor, J.J., Salis, H.M., Simpson, Z.B., Chevalier, A.A., Levskaya, A., Marcotte, E.M., Voigt, C.A., and Ellington, A.D. (2009). A synthetic genetic edge detection program. Cell *137*, 1272–1281.

Tack, D.S., Ellefson, J.W., Thyer, R., Wang, B., Gollihar, J., Forster, M.T., and Ellington, A.D. (2016). Addicting diverse bacteria to a noncanonical amino acid. Nat. Chem. Biol. *12*, 138–140.

Takagi, M., Nishioka, M., Kakihara, H., Kitabayashi, M., Inoue, H., Kawakami, B., Oka, M., and Imanaka, T. (1997). Characterization of DNA polymerase from Pyrococcus sp. strain KOD1 and its application to PCR. Appl. Environ. Microbiol. *63*, 4504–4510.

Tan, C., Marguet, P., and You, L. (2009). Emergent bistability by a growth-modulating positive feedback circuit. Nat. Chem. Biol. *5*, 842–848.

Tang, S.-Y., and Cirino, P.C. (2011). Design and Application of a Mevalonate-Responsive Regulatory Protein. Angew. Chem. Int. Ed. *50*, 1084–1086.

Tawfik, D.S., and Griffiths, A.D. (1998). Man-made cell-like compartments for molecular evolution. Nat. Biotechnol. *16*, 652–656.

Taylor, A.I., Pinheiro, V.B., Smola, M.J., Morgunov, A.S., Peak-Chew, S., Cozens, C., Weeks, K.M., Herdewijn, P., and Holliger, P. (2015a). Catalysts from synthetic genetic polymers. Nature *518*, 427–430.

Taylor, N.D., Garruss, A.S., Moretti, R., Chan, S., Arbing, M.A., Cascio, D., Rogers, J.K., Isaacs, F.J., Kosuri, S., Baker, D., et al. (2015b). Engineering an allosteric transcription factor to respond to new ligands. Nat. Methods *13*, 177–183.

Teichmann, S.A., and Babu, M.M. (2004). Gene regulatory network growth by duplication. Nat. Genet. *36*, 492–496.

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211–1213.

Temme, K., Zhao, D., and Voigt, C.A. (2012a). Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. Proc. Natl. Acad. Sci. U. S. A. *109*, 7085–7090.

Temme, K., Hill, R., Segall-Shapiro, T.H., Moser, F., and Voigt, C.A. (2012b). Modular control of multiple pathways using engineered orthogonal T7 polymerases. Nucleic Acids Res. *40*, 8773–8781.

Temme, K., Hill, R., Segall-Shapiro, T.H., Moser, F., and Voigt, C.A. (2012c). Modular control of multiple pathways using engineered orthogonal T7 polymerases. Nucleic Acids Res. *40*, 8773–8781.

Thyer, R., and Ellefson, J. (2014). Synthetic biology: New letters for life's alphabet. Nature *509*, 291–292.

Thyer, R., Filipovska, A., and Rackham, O. (2013). Engineered rRNA Enhances the Efficiency of Selenocysteine Incorporation during Translation. J. Am. Chem. Soc.

Thyer, R., Robotham, S.A., Brodbelt, J.S., and Ellington, A.D. (2015). Evolving tRNA(Sec) for efficient canonical incorporation of selenocysteine. J. Am. Chem. Soc. *137*, 46–49.

Tóth-Petróczy, A., and Tawfik, D.S. (2014). The robustness and innovability of protein folds. Curr. Opin. Struct. Biol. *26*, 131–138.

Traverse, C.C., and Ochman, H. (2016). Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. Proc. Natl. Acad. Sci. U. S. A. *113*, 3311–3316.

Vaidya, N., Manapat, M.L., Chen, I.A., Xulvi-Brunet, R., Hayden, E.J., and Lehman, N. (2012). Spontaneous network formation among cooperative RNA replicators. Nature *491*, 72–77.

Vander Horn, P.B., Davis, M.C., Cunniff, J.J., Ruan, C., McArdle, B.F., Samols, S.B., Szasz, J., Hu, G., Hujer, K.M., Domke, S.T., et al. (1997). Thermo Sequenase DNA polymerase and T. acidophilum pyrophosphatase: new thermostable enzymes for DNA sequencing. BioTechniques *22*, 758–762, 764–765.

Walter, N.G., and Engelke, D.R. (2002). Ribozymes: catalytic RNAs that cut things, make things, and do odd and useful jobs. Biol. Lond. Engl. *49*, 199–203.

Wang, A.H.-J., Fujii, S., van Boom, J.H., van der Marel, G.A., van Boeckel, S.A.A., and Rich, A. (1982). Molecular structure of r(GCG)d(TATACGC): a DNA–RNA hybrid helix joined to double helical DNA. Nature *299*, 601–604.

Wang, H.H., Isaacs, F.J., Carr, P. a, Sun, Z.Z., Xu, G., Forest, C.R., and Church, G.M. (2009). Programming cells by multiplex genome engineering and accelerated evolution. Nature *460*, 894–898.

Wang, K., Neumann, H., Peak-Chew, S.Y., and Chin, J.W. (2007). Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. Nat. Biotechnol. *25*, 770–777.

Wang, L., Magliery, T.J., Liu, D.R., and Schultz, P.G. (2000). A New Functional Suppressor tRNA/Aminoacyl−tRNA Synthetase Pair for the in Vivo Incorporation of Unnatural Amino Acids into Proteins. J. Am. Chem. Soc. *122*, 5010–5011.

Wang, L., Brock, A., Herberich, B., and Schultz, P.G. (2001a). Expanding the genetic code of Escherichia coli. Science *292*, 498–500.

Wang, L., Brock, A., Herberich, B., and Schultz, P.G. (2001b). Expanding the genetic code of Escherichia coli. Science *292*, 498–500.

Wang, L., Zhang, Z., Brock, A., and Schultz, P.G. (2003). Addition of the keto functional group to the genetic code of Escherichia coli. Proc. Natl. Acad. Sci. *100*, 56–61.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weaver, L.H., Kwon, K., Beckett, D., and Matthews, B.W. (2001). Corepressor-induced organization and assembly of the biotin repressor: a model for allosteric activation of a transcriptional regulator. Proc. Natl. Acad. Sci. U. S. A. *98*, 6045–6050.

Will, C.L., and Luhrmann, R. (2011). Spliceosome Structure and Function. Cold Spring Harb. Perspect. Biol. *3*, a003707–a003707.

Wochner, A., Attwater, J., Coulson, A., and Holliger, P. (2011). Ribozyme-catalyzed transcription of an active ribozyme. Science *332*, 209–212.

Woese, C. (1998). The universal ancestor. Proc. Natl. Acad. Sci. U. S. A. *95*, 6854–6859.

Wong, J.T.-F. (1983). Membership Mutation of the Genetic Code: Loss of Fitness by Tryptophan. Proc. Natl. Acad. Sci. U. S. A. *80*, 6303–6306.

Wong, T.-Y., Fernandes, S., Sankhon, N., Leong, P.P., Kuo, J., and Liu, J.-K. (2008). Role of premature stop codons in bacterial evolution. J. Bacteriol. *190*, 6718–6725.

Xia, G., Chen, L., Sera, T., Fa, M., Schultz, P.G., and Romesberg, F.E. (2002). Directed evolution of novel polymerase activities: Mutation of a DNA polymerase into an efficient RNA polymerase. Proc. Natl. Acad. Sci. *99*, 6597–6602.

Xie, J., and Schultz, P.G. (2006). A chemical toolkit for proteins — an expanded genetic code. Nat. Rev. Mol. Cell Biol. *7*, 775–782.

Xie, G., Keyhani, N.O., Bonner, C.A., and Jensen, R.A. (2003). Ancient Origin of the Tryptophan Operon and the Dynamics of Evolutionary Change. Microbiol. Mol. Biol. Rev. *67*, 303–342.

Xiong, Y., and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. *9*, 3353–3362.

Yamashige, R., Kimoto, M., Takezawa, Y., Sato, A., Mitsui, T., Yokoyama, S., and Hirao, I. (2012). Highly specific unnatural base pair systems as a third base pair for PCR amplification. Nucleic Acids Res. *40*, 2793–2806.

Yang, J., Gunasekera, A., Lavoie, T.A., Jin, L., Lewis, D.E., and Carey, J. (1996). In vivoandin vitroStudies of TrpR-DNA Interactions. J. Mol. Biol. *258*, 37–52.

Yang, Z., Hutter, D., Sheng, P., Sismour, A.M., and Benner, S.A. (2006). Artificially expanded genetic information system: a new base pair with an alternative hydrogen bonding pattern. Nucleic Acids Res. *34*, 6095–6101.

Yang, Z., Chen, F., Alvarado, J.B., and Benner, S.A. (2011). Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System. J. Am. Chem. Soc. *133*, 15105–15112.

Yoshida, S., Hiraga, K., Takehana, T., Taniguchi, I., Yamaji, H., Maeda, Y., Toyohara, K., Miyamoto, K., Kimura, Y., and Oda, K. (2016). A bacterium that degrades and assimilates poly(ethylene terephthalate). Science *351*, 1196–1199.

Young, T.S., Ahmad, I., Yin, J.A., and Schultz, P.G. (2010). An enhanced system for unnatural amino acid mutagenesis in E. coli. J. Mol. Biol. *395*, 361–374.

Zaher, H.S., and Unrau, P.J. (2007). Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. RNA N. Y. N *13*, 1017–1026.

Zhang, Z., Wang, L., Brock, A., and Schultz, P.G. (2002). The selective incorporation of alkenes into proteins in Escherichia coli. Angew. Chem. Int. Ed Engl. *41*, 2840–2842.

Zhou, M., Dong, X., Shen, N., Zhong, C., and Ding, J. (2010). Crystal structures of Saccharomyces cerevisiae tryptophanyl-tRNA synthetase: new insights into the mechanism of tryptophan activation and implications for anti-fungal drug design. Nucleic Acids Res. *38*, 3399–3413.