Identification and Characterization of Functional Biomolecules by *In Vitro* Selection

by

Andrew Carl Larsen

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April, 2015 by the
Graduate Supervisory Committee:

John C. Chaput, Chair
Bertram L. Jacobs
Timothy L. Karr

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

In vitro selection technologies allow for the identification of novel biomolecules endowed with desired functions. Successful selection methodologies share the same fundamental requirements. First, they must establish a strong link between the enzymatic function being selected (phenotype) and the genetic information responsible for the function (genotype). Second, they must enable partitioning of active from inactive variants, often capturing only a small number of positive hits from a large population of variants. These principles have been applied to the selection of natural, modified, and even unnatural nucleic acids, peptides, and proteins. The ability to select for and characterize new functional molecules has significant implications for all aspects of research spanning the basic understanding of biomolecules to the development of new therapeutics. Presented here are four projects that highlight the ability to select for and characterize functional biomolecules through *in vitro* selection.

Chapter one outlines the development of a new characterization tool for *in vitro* selected binding peptides. The approach enables rapid screening of peptide candidates in small sample volumes using cell-free translated peptides. This strategy has the potential to accelerate the pace of peptide characterization and help advance the development of peptide-based affinity reagents.

Chapter two details an *in vitro* selection strategy for searching entire genomes for RNA sequences that enhance cap-independent initiation of translation. A pool of sequences derived from the human genome was enriched for members that function to enhance the translation of a downstream coding region. Thousands of translation enhancing elements from the human genome are identified and the function of a subset is validated *in vitro* and in cells.

Chapter three discusses the characterization of a translation enhancing element that promotes rapid and high transgene expression in mammalian cells. Using this ribonucleic acid sequence, a series of full length human proteins is expressed in a matter of only hours. This advance provides a versatile platform for protein synthesis and is espcially useful in situations where prokaryotic and cell-free systems fail to produce protein or when post-translationally modified protein is essential for biological analysis.

Chapter four outlines a new selection strategy for the identification of novel polymerases using emulsion droplet microfluidics technology. With the aid of a fluorescence-based activity assay, libraries of polymerase variants are assayed in picoliter sized droplets to select for variants with improved function. Using this strategy a variant of the 9°N DNA polymerase is identified that displays an enhanced ability to synthesize threose nucleic acid polymers.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. John Chaput, for giving me the time and freedom I needed to pursue projects that interested me the most. I would also like to thank my supervisory committee members, Drs. Bertram Jacobs and Timothy Karr for their advice and guidance throughout this work.

To the members of the Chaput lab. You have all have helped me learn a great many things, both scientific and otherwise. The kindness you have all shown can never be repaid in full and you will always have my appreciation. Thanks especially to Dr. Brian Wellensiek and soon to be Dr. Matthew Dunn for their contributions to the work presented here.

I also owe significant gratitude to our collaborators for supporting our polymerase engineering endeavors. Thanks to Cody Youngbull and Andrew Hatch for providing the opportunity, guidance and advice to tackle a very complex problem.

I would like to thank the Biological Design Graduate Program for the opportunity to study abroad and for supporting the ideology that graduate education should push the boundaries of traditional academic departments.

I would like to thank the National Science and Engineering Research Council of Canada for fellowship support towards my doctoral studies. I would also like to thank the Graduate & Professional Student's Association at Arizona State University for travel awards which significantly enhanced my education.

To my friends and family, thanks for your support and especially for listening when you made the mistake of asking what I was up to or how things were going. I owe all of my success to my best friend and wife, Dr. Sheri Skerget. Thanks for sharing this journey and congratulations on winning this round.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

General Approach for Characterizing *In Vitro* Selected Peptides with Protein Binding Affinity

Publication Note

This research was originally published in Analytical Chemistry. Larsen, A.C., Gillig, A., Shah, P., Sau, S.P., Fenton, K.E., Chaput, J.C. (2014) General approach for characterizing *in vitro* selected peptides with protein binding affinity. Analytical Chemistry 86:7219-7223 © The American Chemical Society.

Introduction

In vitro selection technologies have become indispensible tools for identifying high affinity peptides to proteins of broad medical and biological interest (1–3). However, the technological advances that have made it possible to generate long lists of candidate peptides have far outpaced our ability to characterize the binding properties of individual peptides. This disparity is due, in part, to recent advances in DNA sequencing technology, which have made it possible to generate millions of peptide sequences from a single *in vitro* selection experiment (4, 5). Other factors that have contributed to the rise in peptide sequence discovery include the use of bar coded libraries and liquid handling robots in selection protocols (6, 7).

Countering these advances is the slow pace at which individual peptides are characterized. In many cases, peptides identified by i*n vitro* selection are produced by solid-phase synthesis, purified by HPLC chromatography, and assayed for function using analytical techniques, such as surface plasmon resonance (SPR), that require large amounts of highly pure peptide or protein. Because this process is both time-consuming and costly, many researchers have turned to column binding assays as a way to quickly screen *in vitro* selected peptides for high affinity binding (8, 9). Although such assays are relatively easy to perform and use only small amounts of

peptide, the data produced is not quantitative. These assays also suffer from high background and problems caused by differential peptide expression, which can make it difficult to compare different peptides analyzed in side-by-side assays (9). Even when high affinity peptides are discovered, additional experiments are needed to obtain quantitative metrics, such as equilibrium binding affinity constants (Kd) that help describe the physical properties of the peptide–protein interaction.

Recognizing the limitations of traditional methods, we sought to develop a new analytical technique to identify high affinity peptides from enriched pools of *in vitro* selected sequences. Our goal was to develop a rapid and inexpensive method that would make it possible to rank selected peptides based on their relative binding affinity and, in a second step, determine the Kd value for the subset of high affinity ligands. The challenge was to design a system that would require minimal amounts of peptide and protein, was amenable to diverse protein classes, and allowed individual assays to be performed in a parallel format.

We envisioned an overall system in which peptides generated by cell-free expression would be brought to equilibrium with their cognate protein, and bound peptide–protein complexes would be separated from the unbound peptide using a double-filter binding assay (Figure 1.1). We felt that this strategy has a number of key advantages over existing methods. First, cell-free peptide synthesis makes it possible to synthesize large numbers of different peptide sequences in a fraction of the time that it would take to obtain the same constructs by solid-phase synthesis (hours vs days) (10). Second, peptides made by cell-free synthesis can be engineered to carry a protein affinity tag, which allows for purification by affinity chromatography. Third, peptides produced by cell-free synthesis can be labeled with ^{35}S-methionine, a radioisotope that allows for accurate detection at low concentrations without altering the physical properties of the peptide. Fourth, filter-

binding assays provide a useful method for determining Kd values, as binding can be measured across a range of protein concentrations (11). Last, the entire process can be performed in parallel, which makes it possible to simultaneously analyze the binding properties of many different peptides.

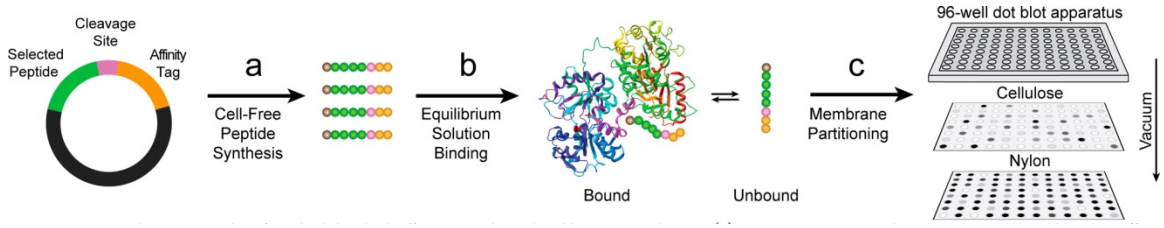


Figure 1.1. General strategy to identify and validate high affinity peptides isolated by *in vitro* selection. (a) DNA sequences encoding peptides with ligand binding affinity are inserted into a custom peptide expression vector, expressed in a coupled cell-free transcription–translation system as ^{35}S-labeled peptides and purified by affinity chromatography. (b) Peptides are equilibrated in solution with their cognate protein target. (c) Bound and unbound peptides are separated in a 96-well dot blot apparatus by passing the mixture through a two-membrane system.

Results

To test our strategy, we designed a custom peptide expression vectorthat would encode our peptide of interest followed by a TEV protease cleavage site and the amino acid sequence for the streptavidin binding peptide (SBP) (12, 13). We anticipated that the SBP tag would pro-vide a convenient positive control for subsequent binding assays, since the SBP-streptavidin interaction is well known (14). In our peptide expres-sion assays, we found that peptides generated in rabbit reticulocyte lysate could bepurified from the crude lysate by affinity capture on streptavidin-coated agarose beads. We developed two elution strategies to recover the peptide from the beads. The first strategy involved eluting the beads with deionized water, which allowed us to obtain the peptide of interest as anSBP fusion peptide for control assays with streptavidin. The second strategy involved eluting the

peptide of interest as a free peptide by incubating the beads with TEV protease, which separated the peptide of interest from the SBP portion of the fusion.

While the synthesis and purification of SBP-tagged peptides proceeded without problem, developing a two-membrane system that could efficiently partition bound peptide–protein complexes from the unbound peptide proved more challenging. Although filter-binding assays represent an established method for studying the binding properties of protein–DNA interactions (13, 15, 16) and more recently have been extended to include protein–XNA complexes (17), such systems have not been developed for protein–peptide interactions. This is presumably due to the fact that nitrocellulose has a general nonspecific affinity for amino acids, which precludes its ability to distinguish peptides from proteins.

To identify a suitable membrane pair, we evaluated the binding properties of several common laboratory membranes with different surface compositions and pore sizes in a dot blot apparatus. We tested nitrocellulose, PVDF, nylon, and cellulose membranes with various pore sizes. However, reproducible results were only observed using a double filter membrane setup with a top layer composed of regenerated cellulose and a bottom layer composed of nylon. Using this membrane configuration, peptide–protein complexes were retained on the top cellulose membrane and unbound peptides that passed through the top membrane were captured on the lower nylon membrane.

We tested the reproducibility of the cellulose–nylon membrane system by performing a 96-well dot-blot assay using SBP and streptavidin to represent a model peptide–protein complex. In this binding assay, SBP-tagged peptides labeled with ^{35}S-methionine were equilibrated in phosphate buffered saline (PBS, pH 7) solutions that either lack or contain streptavidin (50 nM). After 1 h of incubation at 25 °C, the solutions were loaded into the dot blot apparatus and passed through cellulose and

nylon membranes. The membranes were removed, dried, and quantified by phosphorimaging (Figure 1.2a). Analysis of the individual spots allowed us to quantify the amount of SBP peptide present on the cellulose and nylon membranes. We found that 89 ± 2% of the ^{35}S-labeled peptide was retained on the cellulose membrane when streptavidin was present in the PBS buffer. By contrast, only 19 ± 6% of the ^{35}S-labeled peptide remained on the cellulose membrane when streptavidin is absent from the buffer. While some variability was observed across the members, this result suggested to us that the cellulose–nylon double-filter system should be sufficient to distinguish the binding properties of different peptides.

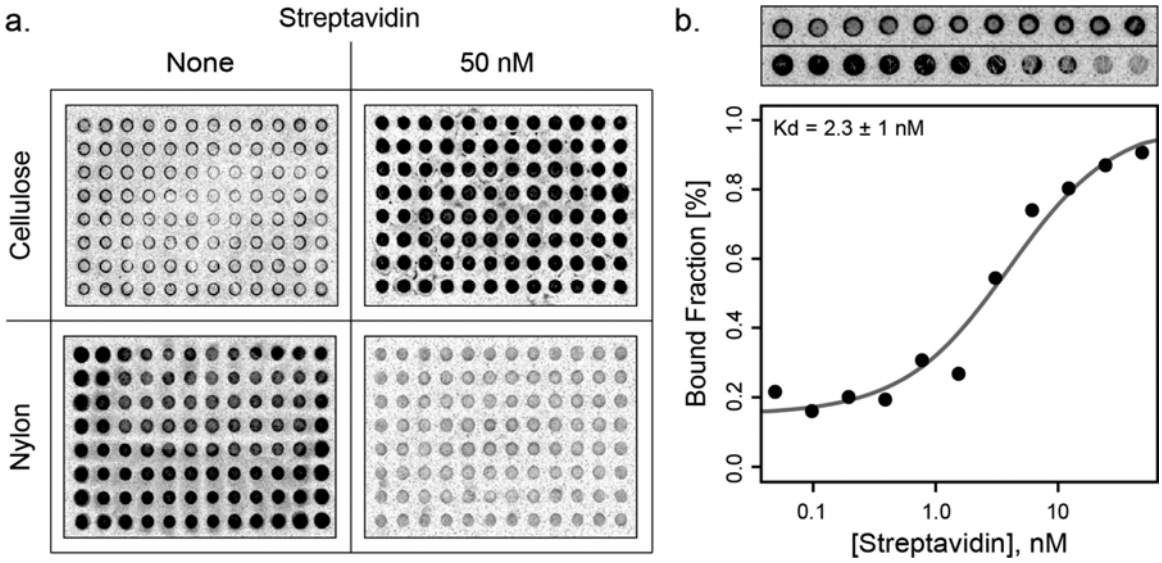


Figure 1.2. Two-membrane double-filtration system for separating bound and unbound peptide–protein complexes. (a) Analysis of the streptavidin binding peptide (SBP) on cellulose and nylon membranes in the absence and presence of streptavidin protein. (b) Equilibrium dissociation plot measuring the binding interaction of SBP with streptavidin.

For the cellulose–nylon system to function as an accurate predictor of peptide binding affinity, it was necessary to confirm that equilibrium was maintained during the filtration step. If equilibrium were disrupted as the peptide–protein complexes passed through the membranes, then the observed binding values would underestimate the true binding affinity of the peptide. To explore this possibility, we

measured the binding affinity constant of the SBP–streptavidin interaction using the double-filter assay. SBP-tagged peptides labeled with ^{35}S-methionine were equilibrated for 1 h in PBS solutions that contained a range of streptavidin protein concentrations. The solutions were loaded into the dot blot apparatus and passed through the cellulose–nylon membranes (Figure 1.2b). Analysis of the bound fraction at each SBP concentration yielded a binding isotherm with a Kd of 2.3 ± 1 nM, which is consistent with the literature value of 2.4 nM (12). On the basis of this result, we concluded that equilibrium is maintained for high affinity peptide–protein complexes.

One of the problems facing those that attempt to identify high affinity peptides to proteins of medical or biological interest is the challenge of distinguishing the highest affinity peptides from a list of *in vitro* selected peptide sequences. In many cases, the highest affinity peptides are not the most abundant sequences or even the sequences that share a common motif (4, 5). Recognizing this problem, we wondered if our cellulose–nylon membrane system could be used to identify and characterize high affinity peptides from a set of *in vitro* selected sequences. To explore this possibility, we identified an mRNA display selection in which a random library of 10^{11} different mRNA–peptide fusions was used to isolate peptides that could bind to human α-thrombin (18). The authors reported 45 sequences that remained in the pool after 10 successive rounds of *in vitro* selection and amplification. Using a column binding assay, two peptides (T10.39 and T10.11) were identified with high affinity to the protein target. Both peptides were synthesized and tested for binding by surface plasmon resonance. T10.39 was found to bind human α-thrombin with a Kd of 166 nM, while T10.11 bound with a Kd of 520 nM (18).

Considering the possibility that some high affinity peptides may have been overlooked due to the limitations of the original column binding assay, we decided to test 24 of the 45 sequences in our cellulose-nylon membrane system. The set of 24

peptides were randomly chosen (Figure 1.3a), inserted into our custom peptide expression vector, expressed in rabbit reticulocyte lysate as ^{35}S-labeled peptide fusions, purified on streptavidin coated agarose beads, and eluted by TEV protease cleavage of the fusion peptide. Three of the peptides (T10.35, T10.46, and T10.57) did not express well and were discarded. Coincidentally, these three peptides also have high hydrophic values, indicating that our screen could be an indicator of peptide solubility. The remaining peptides were each separately incubated with 250 and 500 nM human α-thrombin for 1 hour at room temperature and analyzed in parallel by passing the solutions through the cellulose–nylon membrane system. Control samples lacking thrombin were used to define the level of background binding to the membrane and the fraction of bound peptide was compared to T10.39, a high affinity thrombin-binding peptide (Figure 1.3b).

From our filter-binding assay, we discovered five previously uncharacterized sequences (T10.06, T10.13, T10.25, T10.30, and T10.37) that exhibit at least 25% binding to human α-thrombin when the protein was poised at a concentration of 250 nM. The remaining sequences showed little or no binding, indicating that these sequences are all weak affinity ligands. Of the high affinity peptides, three contain a conserved DPGR motif that is found in T10.39, while the other two show no similarity to T10.39 or each other. This could suggest the peptides bind different sites on the surface of human α-thrombin, with the DPGR containing peptides targeting the same epitope as T10.39 and the two unique peptides binding elsewhere on the surface; however, further experiments are needed to test this hypothesis.

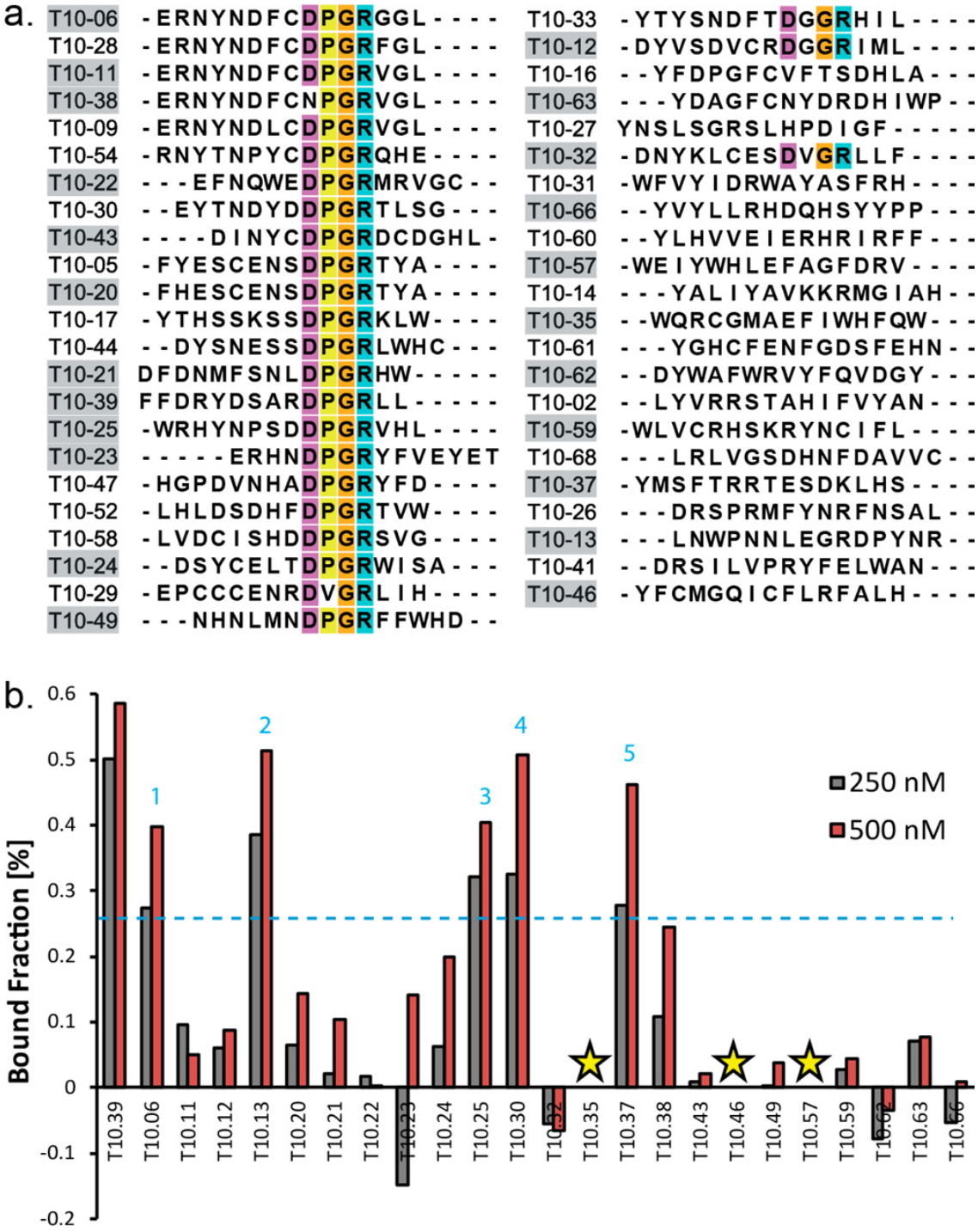


Figure 1.3. Screen of *in vitro* selected thrombin-binding peptides. (a) A list of 45 thrombin-binding peptides. Peptides selected for dot blot analysis (gray) (18). (b) Membrane-based screen of 24 thrombin-binding peptides for affinity to human α-thrombin. Stars indicate peptides with expression levels below the detection limit for dot blot analysis. High affinity peptides are numbered in blue. Arbitrary threshold (dashed blue line).

Given the importance of human α-thrombin as a potential therapeutic target (19), we sought to determine the binding affinity of the five novel thrombin-binding peptides. We began by validating our filter-binding assay using peptide T10.39, which was previously characterized and found to bind thrombin with a Kd of 166 nM (18). Peptide T10.39 was expressed and purified as described above for the SBP peptide. The peptide was then incubated with a range of thrombin concentrations, and peptide–protein complexes were separated from the free peptide by passing the solutions through the cellulose–nylon membranes. Analysis of the binding isotherm revealed Kd of 170 ± 40 nM, which closely approximates the known literature value (Figure 1.4a) (18). Moreover, no difference was observed when the T10.39 was measured as a free peptide or as an SBP fusion peptide (data not shown).

Next, we measured the equilibrium binding affinity of the five uncharacterized peptides using the same methodology described above. In each case, the peptides were expressed and purified as ^{35}S-labeled free peptide by eluting the peptides from the beads with TEV protease. The five peptides were incubated with a range of thrombin concentrations and their Kd values were determined by quantifying the amount of ^{35}S-label on the cellulose and nylon membranes. Peptides T10.25, T10.30, and T10.6, which share the DPGR motif with T10.39, have Kd values of 200, 360, and 460 nM, respectively (Figure 1.4a). By contrast, peptides T10.13 and T10.37, which are unique with respect to the T10.39 sequence, have Kd values of 160 and 310 nM, respectively (Figure 1.4b).

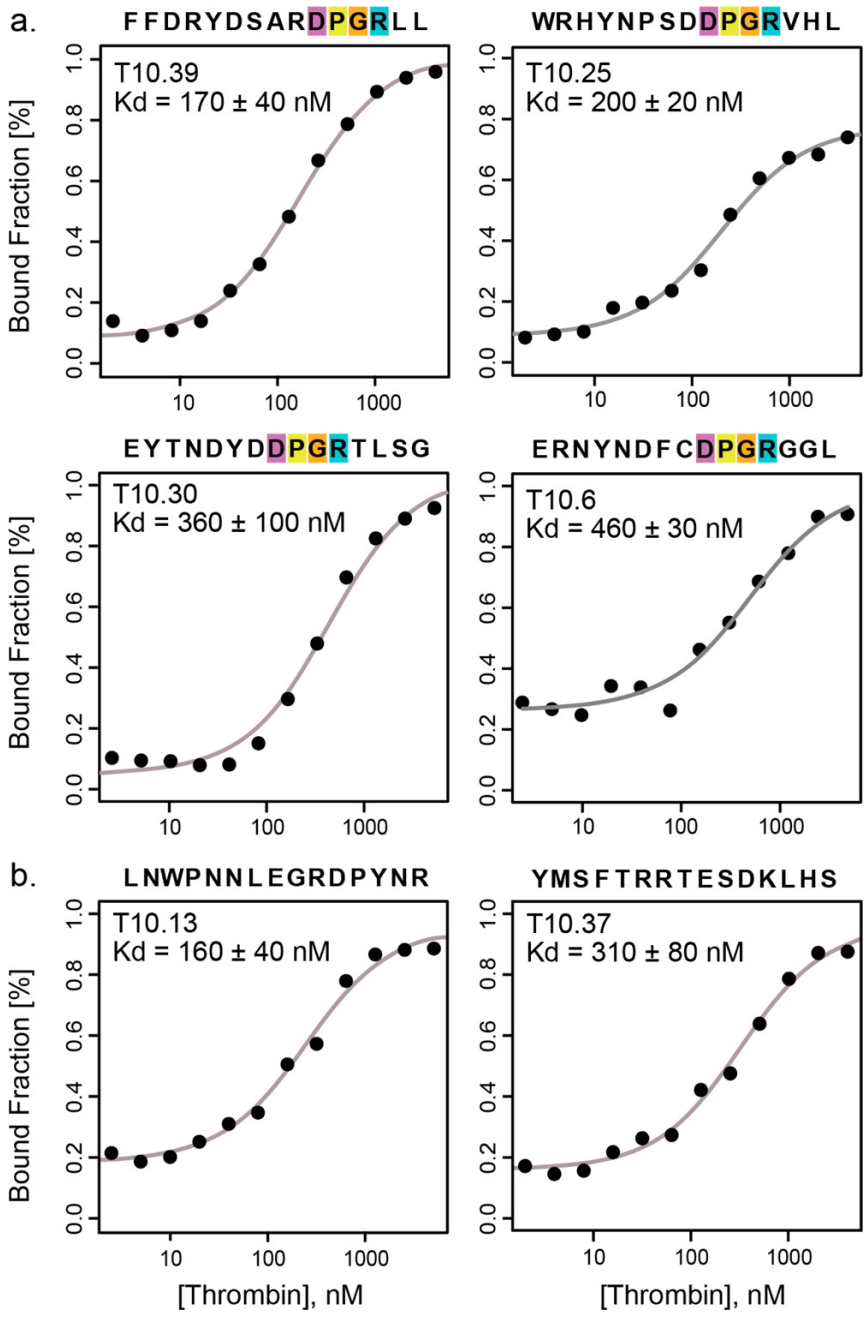


Figure 1.4. Equilibrium dissociation plots measuring the binding affinity for six high affinity thrombin-binding peptides. Binding isotherms were compared for peptides containing (a) and lacking (b) a conserved DPGR motif.

To further validate our results, we produced peptides T10.13 and T10.37 by solid-phase synthesis, purified both sequences by HPLC chromatography, and confirmed their binding affinity by microscale thermophoresis (Figure 1.5). This

technique measures changes in the hydration shell (due to conformational changes) along a temperature gradient, which makes it possible to determine Kd values using minimal amounts of sample (20). We found that peptides T10.13 and T10.37 bind thrombin with Kd values of 180 and 290 nM, which closely approximates the Kd values obtained using our double-filter binding assay.

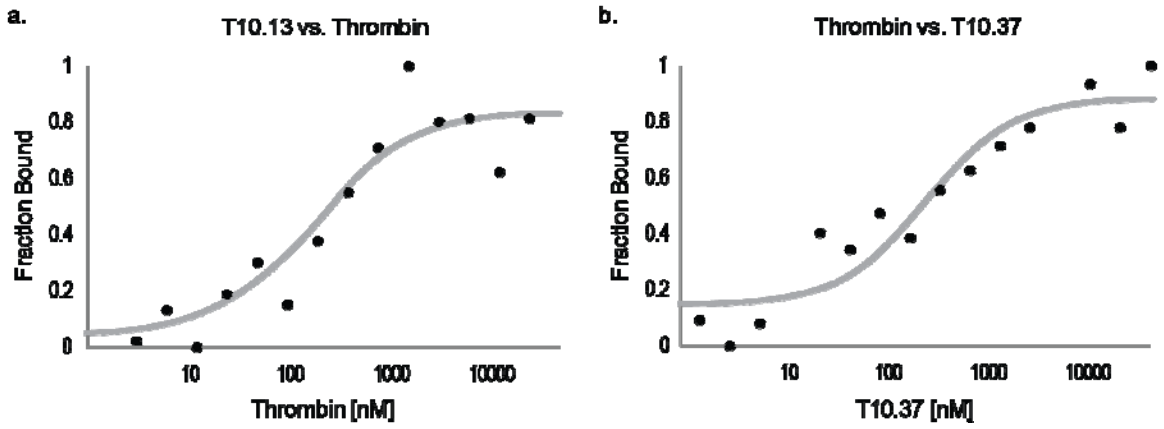


Figure 1.5. Microscale thermophoresis (MST) analysis of two thrombin-binding peptides. (a) A constant concentration of Cy5-labeled T10.13 peptide (50 nM) was equilibrated with varying concentrations of non-labeled thrombin protein before loading into glass capillaries for MST analysis. A Kd of 178 nM ± 22.5 nM was determined for this interaction. (b) A constant concentration of Cy5-labeled thrombin protein (100 nM) was equilibrated with varying concentrations of non-labeled T10.37 peptide before loading into glass capillaries for MST analysis. A Kd of 291 nM ± 35.2 nM was determined for this interaction.

In addition, we also demonstrated that the high affinity peptides function in a complex biological medium. In this case, peptides T10.13 and T10.37 were conjugated to streptavidin-coated magnetic beads and used to recover recombinant human α-thrombin that had been doped into HeLa cell lysate. After an incubation of 1 h at 25 °C, the beads were precipitated, the supernatant was removed, and the beads were washed with TBST buffer. The supernatant and bead samples were analyzed by SDS acrylamide gel electrophoresis. Both peptides pulled-down human α-thrombin from the cell lysate with efficiencies similar to peptide T10.39 and no

contamination above the bead-only control was observed (Figure 1.), demonstrating high affinity and high specificity binding.

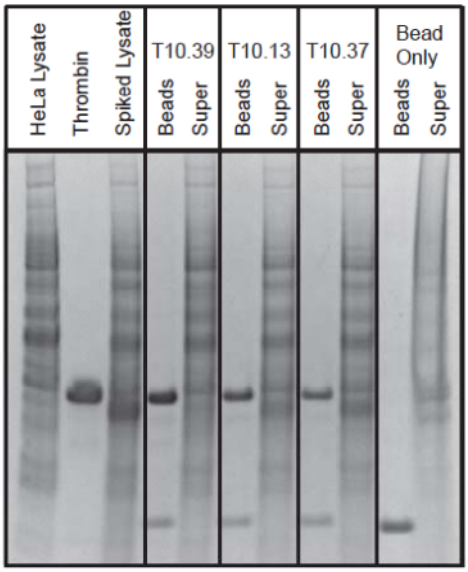


Figure 1.6. Pull-down assay using high affinity thrombin binding peptides. Biotinylated peptides were immobilized on streptavidin coated magnetic beads and incubated with thrombin spiked HeLa cell lysate (1 ug thrombin in 12.5 ug HeLa cell lysate). After an hour of incubation magnetic beads were pulled down and the bead and supernatant samples were run on an SDS-page gel stained with Coomassie blue.

Discussion

The past few years have witnessed an explosion in the demand for high-quality peptides that can be used to support a growing industry of peptide-based therapeutic and diagnostic applications. Unlike antibodies, peptides are amenable to chemical synthesis, generally nonimmunogenic, and their small size allows them to penetrate further into soft tissue (21, 22). These properties, along with improved strategies for increasing serum stability, warrant new methods to streamline the peptide discovery process (23). In line with these efforts, we present a general approach for characterizing the binding properties of *in vitro* selected peptides. This approach provides an inexpensive method to synthesize, purify, screen, and characterize peptides for high-affinity binding to their cognate protein target. We

validated the method using peptides with known protein-binding interactions and applied the strategy to identify five new peptides that bind to human α-thrombin with nanomolar affinity and high specificity.

In summary, we provide a new analytic technique to rapidly screen and characterize *in vitro* selected peptides with high protein binding affinity. We have successfully evaluated peptides that range in size from 22 to 74 amino acids and exhibit binding affinity constants of 1–500 nM. While it is likely that subnanomolar binding affinities could be measured using this approach, we suspect that weaker interactions may not be possible due to long transit times through the membrane. During the course of our study, we noticed that peptides that do not express well by *in vitro* translation tend to have high hydrophobic values, suggesting that peptide recovery after expression and purification could be an indicator of peptide solubility. This observation could provide a simple way to determine whether a peptide will be soluble in an aqueous solution. Relative to more conventional analytical techniques, like SPR or isothermal titration calorimetry (ITC), the method presented here allows for rapid screening of multiple peptide candidates in small sample volumes using cell-free translated peptides that can be obtained in a cost-effective manner. By contrast, SPR and ITC generally require large amounts of purified peptide and/or protein that can be cost prohibitive when screening large numbers of peptides. While our approach is ideal for peptide screening and characterization, high affinity ligands discovered using this method may require further characterization in order to obtain a complete kinetic and thermodynamic profile of the peptide–protein interaction. Recognizing the advantages of small sample volumes, low cost, and high throughput, we suggest that this strategy could be used to accelerate the pace of peptide characterization and help advance the development of peptide-based affinity reagents (24).

Experimental

Materials

Synthetic DNA oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA). Klenow Fragment (exo-) DNA polymerase, HindIII and SbfI restriction enzymes and T4 DNA ligase were purchased from New England BioLabs (Ipswich, MA). The *in vitro* transcription/ translation (TnT) rabbit reticulocyte lysate and ProTEV protease were purchase from Promega (Fitchburg, WI). Streptavidin agarose and the cellulose membrane (10 Kda nominal cut-off, PI88245) were obtained from Thermo Fischer (Waltham, MA). The vacuum Minifold-I Dot Blot apparatus was purchased from GE (Fairfield, Ct). All peptide-protein equilibration reactions were performed in phosphate buffered saline (PBS, 137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, 2 mM KH2PO4, pH 7.4) supplemented with 0.3% BSA and 0.025% Tween-20.

Synthesis of Peptide Expression Vectors

Partially complimentary DNA oligonucleotides were annealed by heating to 90°C followed by five minutes of incubation at 4°C and then extended with Klenow DNA polymerase for 30 minutes at 37°C (Figure 1.7). This produced a synthetic double-stranded DNA with the peptide-coding region flanked by fixed-sequences. Restriction sites compatible with our custom protein expression vector were added by overlapping PCR. Double-stranded DNA was digested and ligated into our expression vector using HindIII and SbfI restriction sites. Plasmids were transformed into E. coli TOP10 competent cells, grown at 37°C on solid agar plates containing ampicillin, and individual colonies were selected for peptide expression. PCR was performed to confirm that the colony contained the plasmid with the correctly sized insert and a subset of plasmids were verified by DNA sequencing (ASU Core Facility).

In Vitro Peptide Expression and Purification

Linear DNA encoding the peptide of interest and all genetic information required for cell-free transcription, translation, and purification was obtained by PCR amplifying miniprep DNA. Peptides were expressed *in vitro* with a C-terminal SBP affinity tag using a coupled *in vitro* transcription/translation (TnT) system (12). One microgram of PCR-generated dsDNA was used as template in a 100 µL reaction containing ^{35}S-methionine for 90 minutes at 30°C. The lysate samples were combined with three volumes of PBS and applied to an affinity matrix of streptavidin-coated agarose beads. After an incubation of 30 minutes at 4°C with rotation, the matrix was then washed with 60 column volumes of PBS, and thrombin peptides were eluted by separating the peptide of interest from the SBP tag with 10 units of TEV protease (overnight incubation at 24°C). The free thrombin binding peptides were isolated by eluting the beads with four column volumes of PBS and quantified by scintillation counting. Eluting the column with eight column volumes of deionized water isolated uncleaved thrombin-SBP fusion peptides.

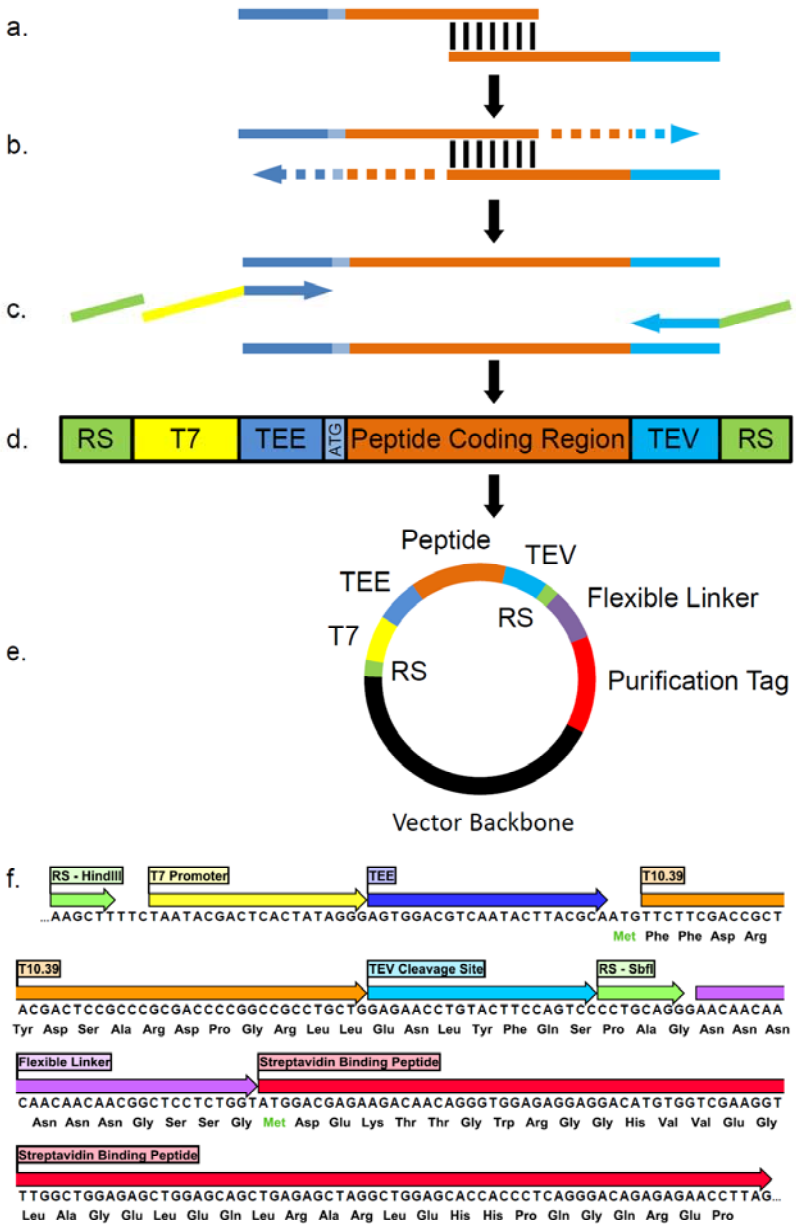


Figure 1.7. Peptide expression vectors construction. a) Two partially complimentary synthetic oligonucleotides encoding the peptide of interest as well as a translation enhancing element (TEE) and protease cleavage site (PCS) were denatured and then annealed. b) Partially complimentary oligonucleotides were extended with Klenow DNA polymerase. c) The T7 promoter site and two restriction sites (RS) were added by overlap extension PCR. d) The dsDNA PCR product was digested and ligated into an expression vector. e) Individual clones were selected that contain all of the information necessary for expression and purification in cell-free lysate. f) DNA sequence of the expression vector in the vicinity of the peptide coding region for the T10.39 peptide.

Dot Blot Binding Assay

Purified ^{35}S-labled peptides were equilibrated with their cognate protein target in PBS containing 0.3% BSA and 0.025% Tween-20 for 1 hour at room temperature and loaded onto a vacuum Minifold-I Dot Blot apparatus containing a cellulose membrane (top layer) and nylon membrane (bottom layer). The time required to bring the system to equilibrium may vary depending on the affinity of the peptide and the concentration of peptide isolated from the *in vitro* translation system. A vacuum pressure of 400 torr was applied overnight to pull the solutions (typically 20 µL sample volumes) through the membrane. While stronger vacuum pressure will decrease the transit time, we found that a low and constant vacuum pressure provided the most reproducible results. The amount of ^{35}S-labeled peptide on each membrane was quantified by phosphorimaging. The fraction of bound peptide for each well was determined by dividing the signal intensity from the cellulose membrane by the combined signal intensity from the cellulose and nylon membranes. For the peptide screen, the thrombin protein was poised at fixed concentrations of 250 and 500 nM. Additional peptide samples were also incubated in the absence of thrombin to determine the background binding of each peptide to the cellulose membrane. For affinity measurements, a dilution series of the target protein spanning at least 10-fold above and below the estimated Kd was used. The protein-bound and free peptide fractions were used to determine the Kd using the following equation:

$$Ib / (Ib + If) = C1 + C2 ([Protein]/([Protein] + Kd))$$

Ib and If are the signal from the bound and free peptide respectively and C1 and C2 are both constants. Affinity dissociation constants were obtained using a non-linear least-squares regression analysis performed using the R software environment (25). The above calculation is not valid if the concentration of peptide is similar to

the concentration of target protein. While most *in vitro* expression systems generate only small amounts of protein, caution should be taken as significant variability is observed between different expression systems and individual peptide sequences.

Microscale Thermophoresis (MST)

Peptides T10.13 and T10.37 were purchased from Biomatik (Cambridge, CAN) in >90% purity. Both peptides were synthesized with a GSKN3 extension at their C-terminus, where the KN3 residue carries an ε-azido lysine modification. The T10.13 peptide was labeled with Cy5 fluorophore using Cu-free click chemistry by adding equimolar concentrations of DBCO modified Cy5 fluorophore (Click Chemistry Tools, Scottsdale, AZ) and the ε-azido lysine peptide in water. The reaction was followed to completion by monitoring the decreasing DBCO absorbance at 310 nm. For Kd measurements, the Cy5-labeled T10.13 peptide (50 nM) was equilibrated with thrombin across a range of protein concentrations (1.5 – 50000 nM) in PBST for 1 hour and then loaded into glass capillaries for MST analysis. For analysis of the T10.37 peptide, human α-thrombin was fluorescently labeled according to the manufacturer's protocol using the Monolith™ Protein Labeling Kit RED-NHS from Nanotemper Technologyies (München, Germany) (20). Labeled thrombin protein (100 nM) was equilibrated with a series of non-labeled T10.37 peptide concentrations (1.5 – 50000 nM) in PBST buffer for 1 hour before loading into glass capillaries for MST analysis. Measurements were made using the Monolith NT.115 from Nanotemper Technologies and Data analysis was performed using Nanotemper Analysis software.

Pull-down Assays

Pull-down assays were performed using chemically synthesized peptides with a C-terminal KN3 (ε-azido lysine) were conjugated to DBCO-PEG12-biotin using Cu-free click chemistry as described above. The conjugated peptides were immobilized

on M-270 streptavidin coated magnetic beads (Pierce Biotechnology) by incubating a 10-fold excess of the magnetic bead nominal binding capacity for 1 hour at 25°C. The beads were then washed with PBST to remove free peptide-biotin conjugates. Peptide immobilized beads were then incubated with HeLa cell lysate spiked with human α-thrombin (1 µg thrombin in 12.5 µg HeLa cell lysate). After 1 hour of incubation at 25°C the magnetic beads were pulled down and the supernatant was collected into a fresh tube. The beads were washed with PBST buffer. The supernatant fractions were evaporated to dryness. The washed beads and dried supernatant samples were denatured in 1x SDS loading buffer by heating at 90°C for 10 minutes and run on a 4-12% gradient SDS-page gel followed by staining with Coomassie blue.

References

1. Baines IC, Colas P (2006) Peptide aptamers as guides for small-molecule drug discovery. *Drug Discov Today* 11(7-8):334–341.

2. Takahashi TT, Austin RJ, Roberts RW (2003) mRNA display: Ligand discovery, interaction analysis and beyond. *Trends Biochem Sci* 28(3):159–165.

3. Molek P, Strukelj B, Bratkovic T (2011) Peptide phage display as a tool for drug discovery: Targeting membrane receptors. *Molecules* 16(1):857–887.

4. Olson CA, et al. (2012) Single-round, multiplexed antibody mimetic design through mRNA display. *Angew Chemie - Int Ed* 51(50):12449–12453.

5. Matochko WL, et al. (2012) Deep sequencing analysis of phage libraries using Illumina platform. *Methods* 58(1):47–55.

6. Turunen L, Takkinen K, Söderlund H, Pulli T (2009) Automated panning and screening procedure on microplates for antibody generation from phage display libraries. *J Biomol Screen* 14(3):282–293.

7. Cung K, et al. (2012) Rapid, multiplexed microfluidic phage display. *Lab Chip* 12(3):562–565.

8. Baggio R, et al. (2002) Identification of epitope-like consensus motifs using mRNA display. *J Mol Recognit* 15(3):126–134.

9. Colwill K, Gräslund S (2011) A roadmap to generate renewable protein binders to the human proteome. *Nat Methods* 8(7):551–558.

10. Katzen F, Chang G, Kudlicki W (2005) The past, present and future of cell-free protein synthesis. *Trends Biotechnol* 23(3):150–156.

11. Wong I, Lohman TM (1993) A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions. *Proc Natl Acad Sci U S A* 90(12):5428–5432.

12. Wilson DS, Keefe AD, Szostak JW (2001) The use of mRNA display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci U S A* 98(7):3750–3755.

13. Kapust RB, Waugh DS (2000) Controlled intracellular processing of fusion proteins by TEV protease. *Protein Expr Purif* 19(2):312–318.

14. Keefe AD, Wilson DS, Seelig B, Szostak JW (2001) One-step purification of recombinant proteins using a nanomolar-affinity streptavidin-binding peptide, the SBP-Tag. *Protein Expr Purif* 23(3):440–446.

15. Yu H, Jiang B, Chaput JC (2011) Aptamers can discriminate alkaline proteins with high specificity. *ChemBioChem* 12(17):2659–2666.

16. Xu D, Shi H (2009) Composite RNA aptamers as functional mimics of proteins. *Nucleic Acids Res* 37(9):1–9.

17. Yu H, Zhang S, Chaput JC (2012) Darwinian evolution of an alternative genetic system provides support for TNA as an RNA progenitor. *Nat Chem* 4(3):183–187.

18. Raffler N a., Schneider-Mergener J, Famulok M (2003) A novel class of small functional peptides that bind and inhibit human alpha-thrombin isolated by mRNA display. *Chem Biol* 10(1):69–79.

19. Crawley JTB, Zanardelli S, Chion CKNK, Lane D a. (2007) The central role of thrombin in hemostasis. *J Thromb Haemost* 5:95–101.

20. Jerabek-Willemsen M, Wienken CJ, Braun D, Baaske P, Duhr S (2011) Molecular interaction studies using microscale thermophoresis. *Assay Drug Dev Technol* 9(4):342–353.

21. Sato AK, Viswanathan M, Kent RB, Wood CR (2006) Therapeutic peptides: technological advances driving peptides into development. *Curr Opin Biotechnol* 17(6):638–642.

22. Goodwin D, Simerska P, Toth I (2012) Peptides As Therapeutics with Enhanced Bioactivity. *Curr Med Chem* 19(26):4451–4461.

23. McGregor DP (2008) Discovering and improving novel peptide therapeutics. *Curr Opin Pharmacol* 8(5):616–619.

24. Williams BA, et al. (2009) Creating protein affinity reagents by combining peptide ligands on synthetic DNA scaffolds. *J Am Chem Soc* 131(15):17233–17241.

25. R Core Team R: A Language and Environment for Statistical Computing. Available at: http://www.r-project.org.

CHAPTER 2

Genome-Wide Profiling of Cap-Independent Translation Enhancing Elements

Publication Note

This research was originally published in Nature Methods. Wellensiek, B.P., Larsen, A.C., Stephens, B., Kukurba, K., Waern, K., Briones, N., Liu, L., Snyder, M., Jacobs, B.L., Kumar, S. and Chaput, J.C. (2013) Genome-Wide Profiling of Cap-Independent Translation Enhancing Elements. Nat. Methods 10:747-750 © Macmillan Publishers Limited.

Introduction

In eukaryotes, initiation of translation usually follows a cap-dependent mechanism, in which the 43S ribosomal preinitiation complex is recruited to a 7-methylguanosine cap located at the 5′ end of the mRNA strand via recognition of the cap-binding complex eIF4F (26, 27). Although we now have a detailed structural and mechanistic understanding of each step in the cap-dependent process (26, 27), very little is known about the molecular basis of cap-independent initiation of translation (28). Cap-independent translation occurs during normal cellular processes (for example, mitosis and apoptosis) or when the cap-dependent translation machinery is compromised by viral infection or disease (29, 30). To address this critical gap in our understanding of protein translation, we developed an in vitro selection strategy to identify sequences in the human genome that mediate cap-independent initiation of translation.

Our selection strategy relies on mRNA display, which is a cell-free method for covalently linking newly translated proteins to their encoding RNA message (31). In this approach (Figure 2.1a), a genomic library is inserted into the 5′ untranslated region (UTR) of a DNA construct containing the genetic information necessary for mRNA display. The library is in vitro–transcribed to yield a pool of uncapped single-

stranded mRNA that is photo-ligated at the 3′ end to a DNA linker containing a 3′ puromycin residue. When translated in vitro, RNA sequences that mediate cap-independent initiation of translation become covalently linked to a peptide affinity tag encoded in the open reading frame. Formation of a chemical bond between newly translated peptides and their encoding mRNA occurs via the natural peptidyl transferase activity of the ribosome, which recognizes puromycin as a tyrosyl-tRNA analog (Figure 2.1b). Functional RNAs are then isolated, reverse-transcribed and amplified by PCR to regenerate the pool of DNA for another selection cycle.

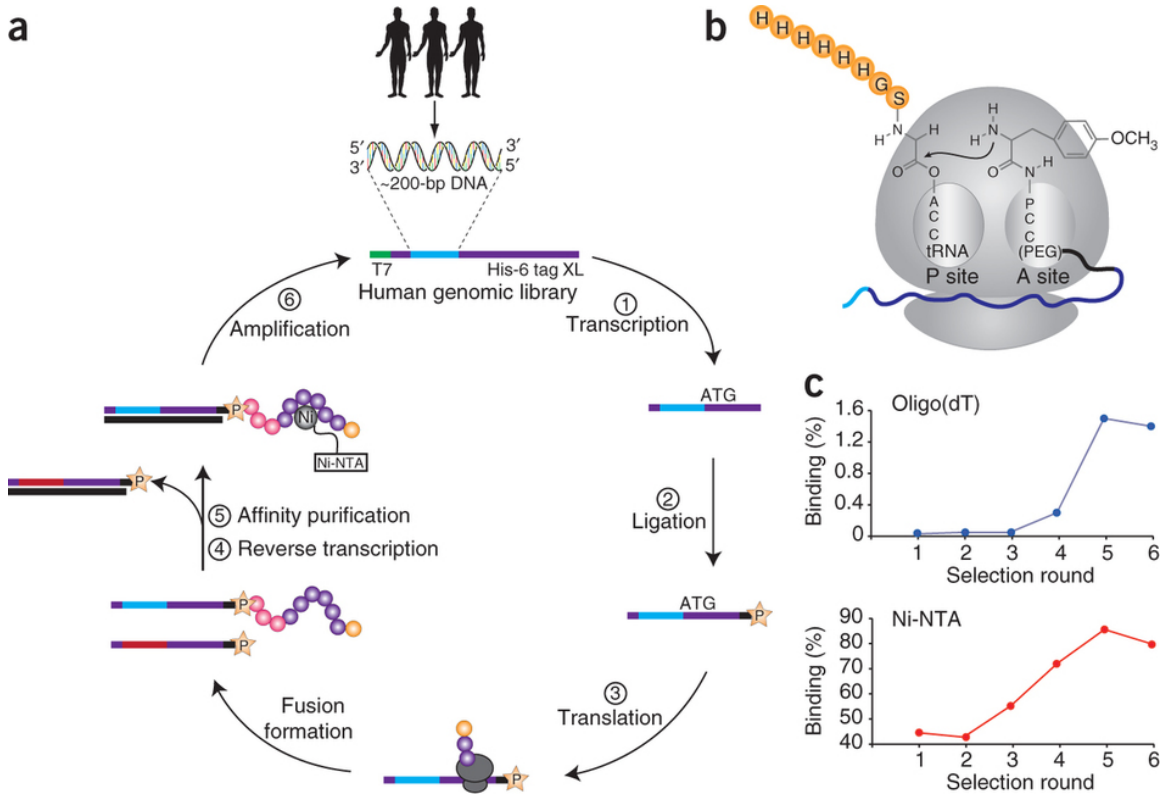


Figure 2.1. In vitro selection of RNA elements that mediate cap-independent translation. (a) A library of human genomic DNA fragments was inserted into a DNA cassette for mRNA display. For each selection round, the dsDNA pool was in vitro–transcribed into single-stranded RNA, conjugated to a DNA-puromycin linker and translated in vitro. Uncapped mRNA sequences that initiate translation of an intact open reading frame become covalently linked to a His-6 protein affinity tag encoded in the RNA message. Functional molecules are recovered, reverse transcribed and amplified by PCR to generate the DNA for the next selection cycle. T7, T7 RNA

polymerase promoter; XL, photo–cross-linking site. (b) Schematic of RNA-protein fusion molecule generated via the natural peptidyl transferase activity of the ribosome. (c) Percentage of ^{35}S-labeled fusion molecules recovered from the oligo(dT) and Ni-NTA affinity columns.

Results

We began the selection with a library of ~10^{13} RNA-DNA-puromycin molecules containing a random region of genomic fragments (~150 nucleotides) derived from total human DNA (32). We translated the library for 1 h at 30 °C and then incubated the translation mixture overnight at −20 °C under high-salt conditions to promote formation of mRNA-peptide fusions. We isolated the fusions from the crude lysate by oligo(dT) affinity purification, reverse-transcribed the mRNA portion into cDNA to form chimeric cDNA-RNA heteroduplexes and immobilized sequences displaying a His-6 affinity tag on Ni-NTA agarose beads. After washing the beads to remove RNA molecules that did not form mRNA-peptide fusions or did not translate in the correct reading frame, we eluted the remaining mRNA-peptide fusions with imidazole, exchanged the eluate into buffer and performed PCR amplification to reinitiate another selection cycle.

The abundance of mRNA-peptide fusions plateaued after six rounds of mRNA display, indicating that the library had become dominated by sequences that could enhance cap-independent initiation of translation (Figure 2.1c). To assess the level of sequence diversity that remained in the pool, we cloned and sequenced individual members from the selection output. We identified 636 unique sequences, 225 of which exhibited 100% identity to the human reference genome (hg18). The remaining 411 sequences had high homology (85–99% identity) but contained sequence variation that included single nucleotide polymorphisms in addition to small insertions and deletions. Such variation is expected for individuals in a population,

and it is known that functionally relevant sequences can differ between individual genomes (33, 34).

To test our selected sequences for functional activity in human cells, we modified two luciferase reporter vectors used previously to evaluate translation initiation by adding a promoter sequence specific to our cell-based system (35) (Figure 2.2a). The first vector contained an unstructured 5′ UTR designed to quantify the activity of TEEs. The second vector contained a stable stem-loop structure (Gibbs free energy (ΔG) = −58 kcal mol^{-1}) upstream of the insert, which blocks translation in the absence of an internal ribosomal entry site (IRES). Translation of both mRNA templates containing a no-insert 13-nucleotide control sequence confirmed that the stem-loop structure inhibited translation (~99% inhibition) in vitro and in cells (Figure 2.2b). Quantitative real-time PCR (qRT-PCR) confirmed that the differences in translation were not caused by differences in RNA expression.

Because cryptic splicing activity is a common cause of IRES misinterpretation (36), we used a cytoplasmic expression system that bypasses nuclear expression (37). In this system, mammalian cells transfected with an expression vector carrying a vaccinia virus (VACV)-specific promoter are immediately infected with VACV. The virus produces its own RNA polymerase that recognizes the viral promoter and mediates RNA expression in the cytoplasm. We confirmed that nuclear expression did not contribute to translation by measuring the luciferase activity of transfected cells that were not infected with VACV. These cells yielded luciferase values equivalent to those for untreated control cells (data not shown).

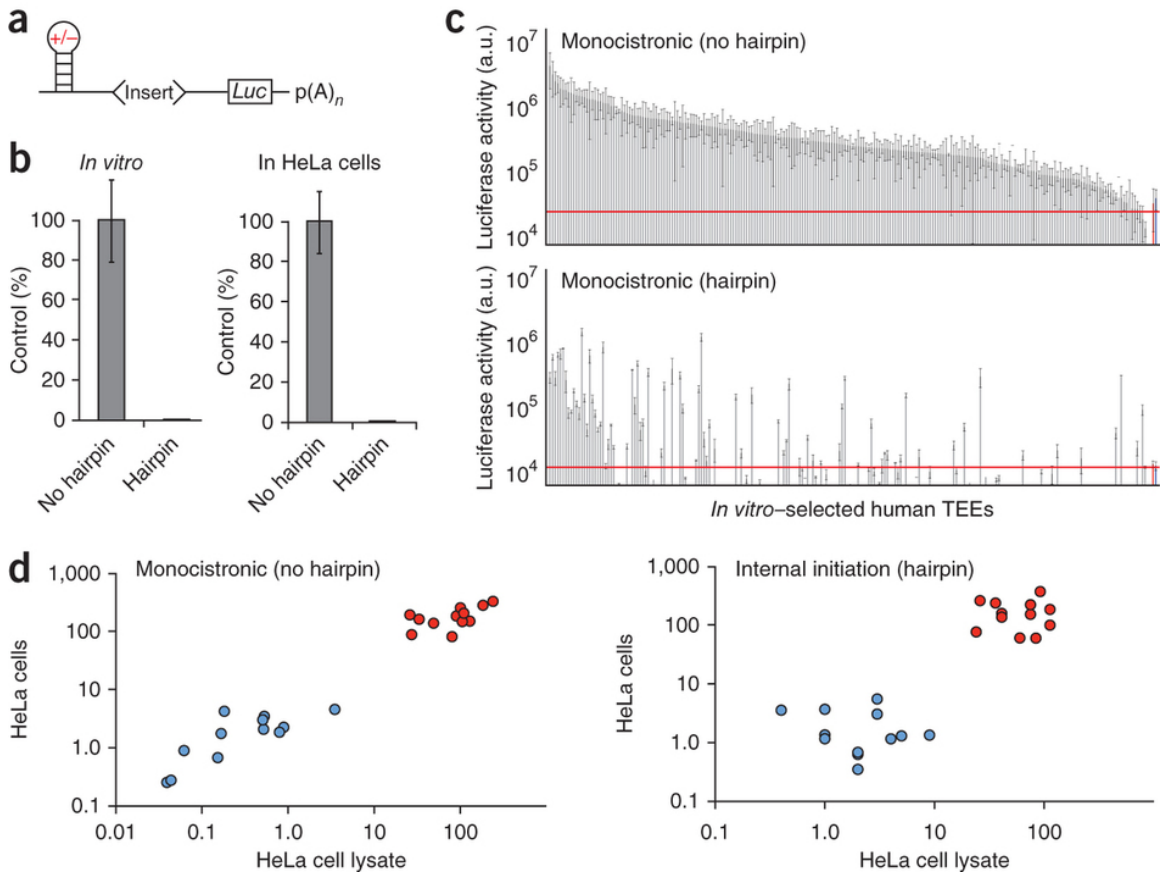


Figure 2.2. Functional analysis of selected TEEs in human cells and in vitro. (a) Firefly luciferase reporter (Luc) with or without (+/−) a stable stem-loop structure in the 5′ UTR. p(A)n, polyadenylation signal. (b) Translation efficiency, as measured by luciferase activity, of a no-insert control in the absence and presence of the stem-loop structure, assayed in HeLa cell lysate (in vitro) and in HeLa cells. Error bars, s.d.; n = 3. (c) Translation-enhancing activity of 225 representative sequences after six rounds of in vitro selection, assayed using a luciferase reporter construct in the absence and presence of the stem-loop structure (hairpin) in HeLa cells. Results were compared to data for an unstructured 13-nucleotide insert (red), which defined the basal level of bioluminescence activity for the reporter plasmid. Error bars, s.d.; n = 2. (d) Comparison of 12 high-activity sequences (red) to an equal number of unselected sequences from the starting library (blue) in the absence and presence of the stem-loop structure in HeLa cells and in HeLa cell lysate. Fold enhancement of translation was measured relative to a no insert reporter containing a 13-nucleotide unstructured sequence in place of the TEE. Data shown represents an average of 2 experiments. Raw data are provided in Table 2.1. Luciferase values were normalized to luciferase mRNA data for cell-based experiments in b and d but not in c.

Next, we tested perfectly matched sequences for TEE and IRES function in human cells. Using the unstructured vector, we found that the selected sequences produced up to 100-fold more luciferase than the no-insert control (Figure 2.2c), demonstrating that our in vitro selection strategy enriched for sequences that enhance translation. Approximately 20% of our TEEs remained functional when tested in the stable stem-loop structure (Figure 2.2c), suggesting that a subset of our in vitro–selected TEEs function as IRESs. To ensure that the observed IRES activity was not due to a cryptic promoter (38), we screened 20 high-activity sequences in HeLa cells using a vector lacking the VACV promoter. This assay identified 8 sequences with modest to high luciferase activity, indicating that these sequences harbored a cryptic promoter (Figure 2.3). We considered the remaining 12 sequences to be human IRESs, as their function was not an artifact of RNA splicing or cryptic promoter activity.

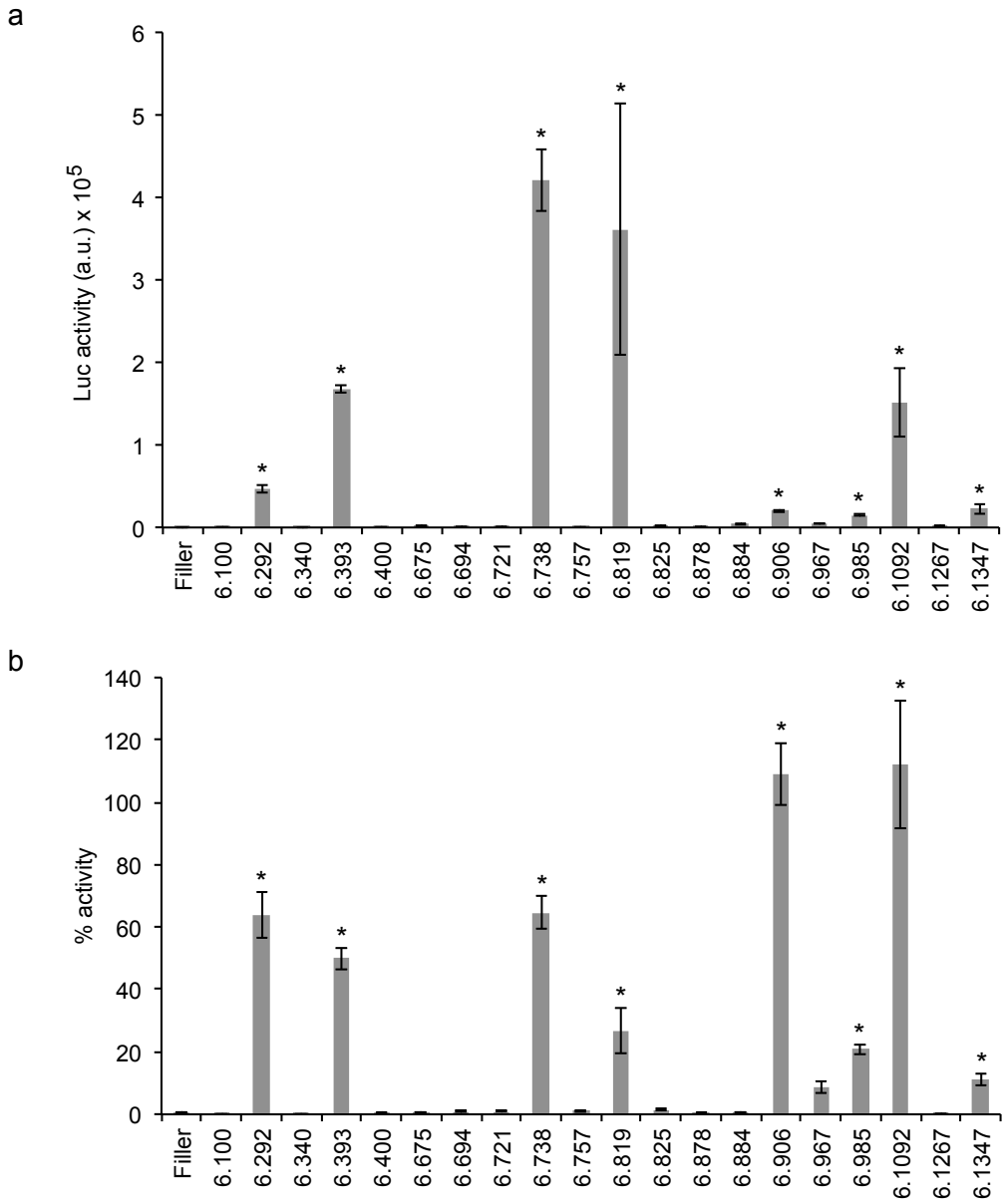


Figure 2.3. Cryptic promoter activity of TEEs. (a) Twenty of the top translation enhancing sequences identified from round 6 of the selection were evaluated for cryptic promoter activity in the transfect/infect assay using a plasmid that does not contain any known promoters. Results for plasmids containing sequences from round 6 were compared to those of a plasmid which contained a simple filler sequence in place of the TEE. (b) The luciferase activity from the promoter-less plasmids was compared to the activity of the twenty TEEs in a plasmid containing a stable stemloop upstream of the TEE in the 5' UTR. Any TEEs which generated promoter-less activity greater than 10% of the stemloop activity, indicated with a *, were not included in further studies.

We then compared the 12 human IRESs to 12 randomly chosen sequences from the starting library in the unstructured and stem-loop luciferase reporter vectors, both in HeLa cells and in HeLa cell lysates. In the unstructured luciferase reporter system, we observed strong concordance between luciferase assays performed in HeLa cells and in HeLa cell lysates, which resulted in ~100-fold greater translation-enhancing activity for the 12 human IRESs relative to the randomly chosen sequences from the starting library (Figure 2.2d and Table 2.1). We observed a similar trend for the stem-loop luciferase reporter system, which showed that the selected sequences exhibit up to ~400-fold higher activity in cells and up to ~100-fold higher activity in vitro than the randomly chosen sequences from the starting library (Figure 2.2d and Table 2.1). Collectively, these results establish the ability of our in vitro selection strategy to identify RNA sequences from the human genome that function as efficient translation-enhancing elements, a subset of which function as IRESs.

Table 2.1
Translation enhancing activity of 12 selected TEEs following 6 rounds of mRNA display selection. Activity was calculated as fold enhancement over results obtained when using a 13-nt unstructured (no insert) sequence in place of the TEE.

Fold Enhancement (±SD)				
	No hairpin		**Hairpin**	
Clone ID	**In cells**	**Cell lysate**	**In cells**	**Cell lysate**
6.100	182±4	90±1	235±9	36±3
6.340	160±14	33±2	76±3	24±4
6.400	87±2	27±1	156±22	41±4
6.675	252±17	100±4	258±10	26±3
6.694	278±11	184±8	151±10	75±13
6.721	80±8	80±2	60±12	60±2
6.757	149±10	129±3	59±2	84±7
6.825	191±11	26±2	135±6	41±11
6.878	137±5	49±3	220±18	75±10
6.884	146±13	106±2	183±17	113±21
6.967	325±55	240±4	99±9	113±5
6.1267	203±16	110±2	368±37	92±26

One caveat of our HeLa cell assay is that the mRNA transcripts likely contain a 5′ cap because of the strong capping enzymes encoded in the VACV genome (37). This is not a concern for the hairpin construct as the stem-loop structure blocked cap-dependent initiation of translation (Figure 2.2b). However, in the case of the unstructured templates, where a 5′ cap could aid initiation of translation, additional experiments are needed to define the activity of the TEE. We therefore selected 26 sequences that exhibited a range of TEE activities but had no observable IRES activity (Figure 2.2c). We then measured their luciferase activity under cap-independent conditions relative to the no-insert control. Consistent with the functional constraints of our in vitro selection, the selected TEEs maintained their activity in the absence of a 5′ cap (Figure 2.4). In some cases, activity increased considerably when the 5′ cap was missing, suggesting that certain TEEs prefer cap-independent pathways for initiation of translation. This observation provides new insight into the mechanism of initiation of translation where the 5′ cap is thought to inhibit alternate pathways (39).

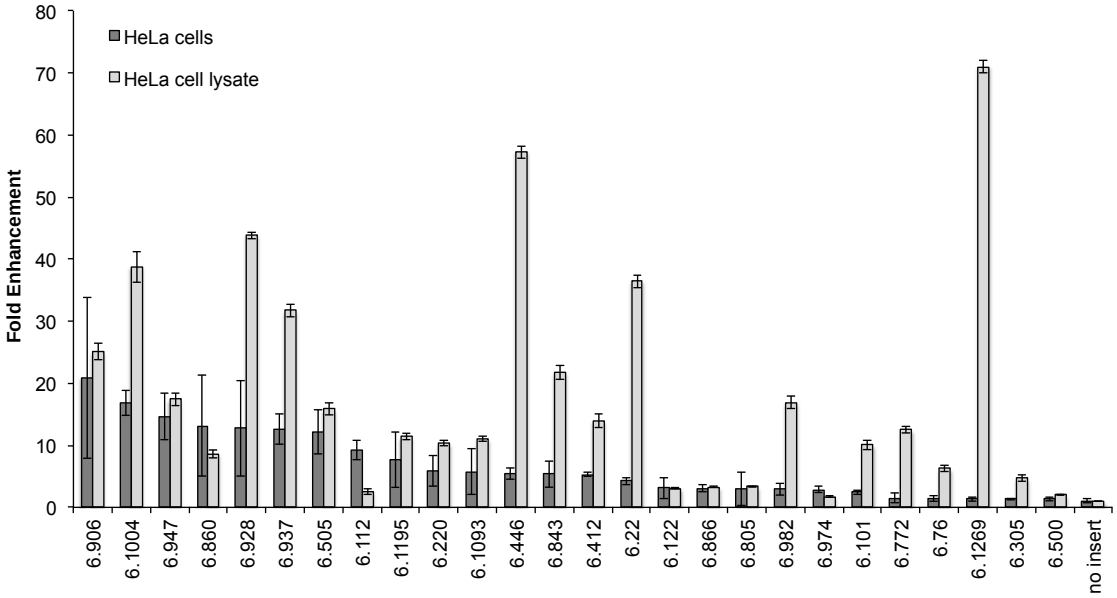


Figure 2.4. Cap-independent activity of TEEs. A set of 26 sequences were selected that exhibit a range of TEE activity but no observable IRES activity. Each sequence was assayed in the absence of a 5′ cap in HeLa cell lysate and values of fold enhancement over a no insert control were plotted (light grey). For comparison purposes, the fold enhancement observed under cap-dependent translation in HeLa cells was also plotted for each sequence (dark grey).

As only a few human TEEs are known (40), we performed Illumina deep sequencing on the starting library (round 0, *R0*) and the selection output (round 6, *R6*). Sequence analysis revealed that only 2% of the *R0* sequences remained in the pool after six rounds of selection. We aligned the *R0* and *R6* sequences to the reference human genome (hg19) and identified 12,278 unique regions that were enriched by at least tenfold (Figure 2.5 and Table 2.2). The in vitro–selected TBRs mapped to ~2 million base pairs. A vast majority of TBRs were shorter than 250 base pairs (99.5%) and were widely dispersed across all 24 chromosomes (Figure 2.6a and 3.7). Of these, 12% (1,532 TBRs) mapped to genomic regions containing known genes, even though genic regions (introns and exons) account for ~40% of the human genome (Figure 2.6b) (41). This underabundance in genic regions may be a result of negative selection against TEEs aimed at avoiding disruptive

translation in nature, which would be consistent with our results of TEE activity in vitro and in cells (Figure 2.2). Moreover, TBRs were preferentially located in 5′ UTRs of genes (threefold over-representation), which would suggest potential functional roles for these elements. We also observed a small but significant enrichment of TBRs in long noncoding RNA regions as compared to the entire human genome (12.2% versus 11.5%, binomial test, P = 0.003), which could lead to the production of novel proteins as these sites are located in intragenic regions of the genome.

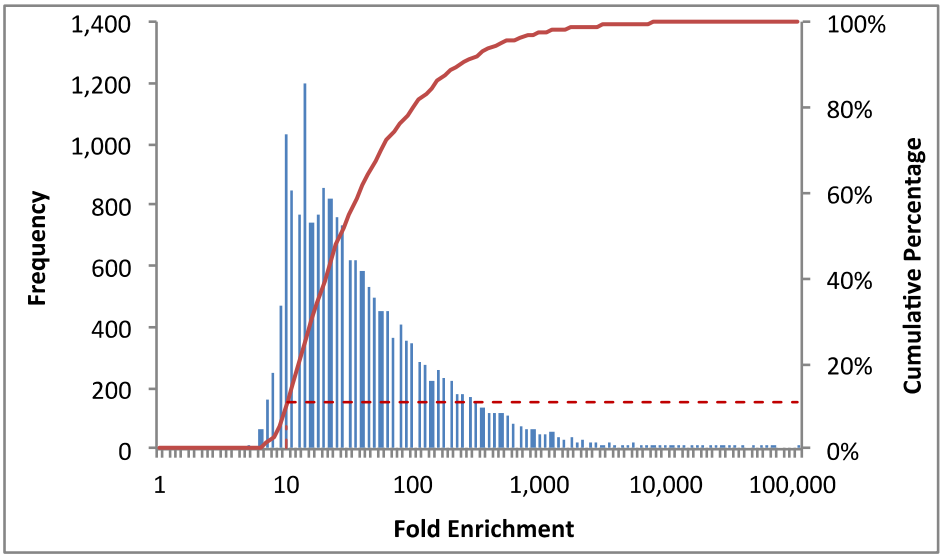


Figure 2.5. Frequency (blue bars) and cumulative distributions (solid red line) of fold enrichment for all single-copy peaks. Fold enrichment represents the ratio of normalized R6 reads over R0 reads for a specific peak4. The broken red line illustrates that 11% of the peaks are enriched less than 10-fold.

Table 2.2
Outputs from each step of the sequence processing pipeline: (**a**) raw Illumina reads, (**b**) after primer trimming, (**c**) after genome mapping, (**d**) after peak calling, (**e**) after enrichment estimation, and (**f**) after repeat masking.

(a)

	library	reads	bases
Raw Illumina Reads	R0	44,444,004	3,555,520,320
	R6	15,822,677	1,265,814,160

(b)

	library	reads	bases
Primer Trimmed	R0	37,844,023	2,214,081,562
	R6	15,487,289	880,257,653

(c)

		library	reads	bases
Aligned to Genome	***Single Copy***	R0	30,344,547	1,812,193,574
		R6	8,208,028	467,542,238
	Low Copy	R0	4,665,039	259,320,667
		R6	3,834,883	204,430,440
	High Copy	R0	1,815,116	111,163,504
		R6	1,444,416	86,354,599
	Unaligned	R0	1,019,321	31,403,817
		R6	1,999,962	121,930,376

(d)

		reads	peaks
Peak Calling	***Single Copy***	4,833,027	18,353
	Low Copy	3,041,808	4,544
	High Copy	799,386	8,267

(e)

		fold	peaks
Enrichment	***Single Copy***	≥10	17,349
		≥100	3,662
		≥1,000	495
	Low Copy	≥10	4,246
		≥100	1,020
		≥1,000	113
	High Copy	≥10	7,949
		≥100	745
		≥1,000	44

(f)

		fold	peaks
Repeats Masked	***Single Copy***	≥10	12,278
		≥100	2,291
		≥1,000	312

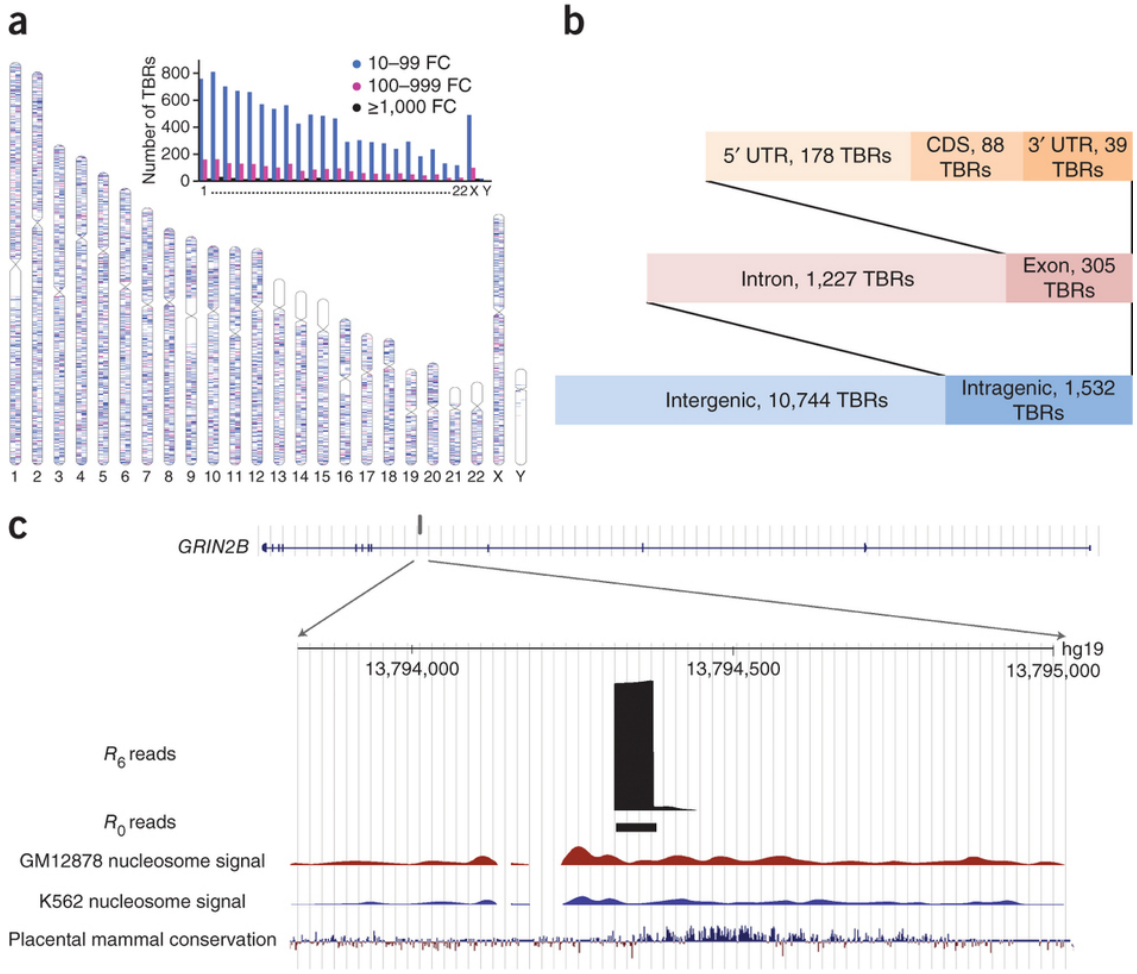


Figure 2.6. Genomic landscape of human TEEs. (a) Chromosomal ideogram of TBRs with different levels of sequence enrichment between the starting pool (*R0*) and the selected library (*R6*): low (10–99-fold), medium (100–999-fold) and high (≥1,000-fold). The blank regions in the chromosome correspond to the unsequenced regions in the reference genome (hg19). Inset, total number of TBRs per chromosome, sorted by enrichment level. FC, fold change. (b) Quantity of TBRs in various genomic regions. TBRs were underrepresented in intragenic and exonic regions (binomial test, both $P < 10^{-16}$) and overrepresented in 5′ UTRs (binomial test, $P < 10^{-16}$). CDS, coding sequence. (c) Genomic context of an example TBR residing in an intron of the GRIN2B gene.

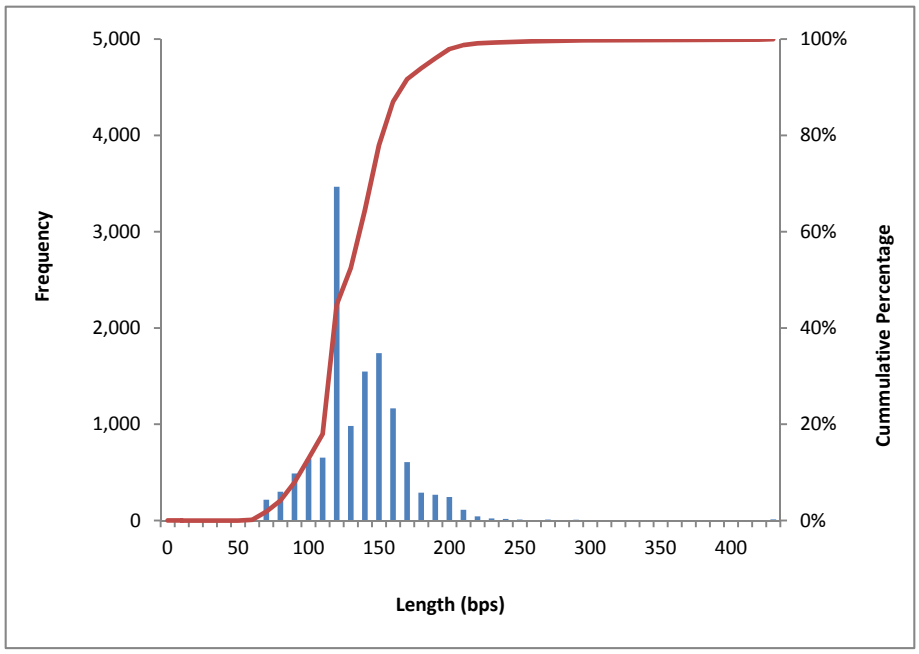


Figure 2.7. Frequency (blue bars) and cumulative distribution (red line) of TBR lengths. TBRs have an average length of 132 bps with a standard deviation of 34 bps.

Gene Ontology analysis revealed that many TBRs associate with genes involved in signal transduction, cell communication and neurological system development pathways (Figure 2.8). These functional categories are frequently reported for genes that have undergone adaptive evolution (42, 43). One example is genes encoding glutamate receptors, which are important for neural communication, memory formation, learning and regulation (44). Among the 21 human genes encoding glutamate receptors, eight harbor TBRs in their introns. Of these, two were enriched by more than 1,000-fold after in vitro selection using mRNA display. Some of these sequences are flanked by regions that are highly conserved among species and exhibit transcriptional activity in cells, indicating a possible role for TBRs in the translation of proteins involved in important developmental pathways. One example is a TBR located in an intron of the GRIN2B gene (Figure 2.6c). This sequence overlaps with active nucleosome binding sites in the Encyclopedia of DNA Elements

(ENCODE) cell lines GM12878 and K562, and is upstream of a highly conserved region among placental mammals. We identified population polymorphisms upstream of, but not within or downstream of, this TBR.

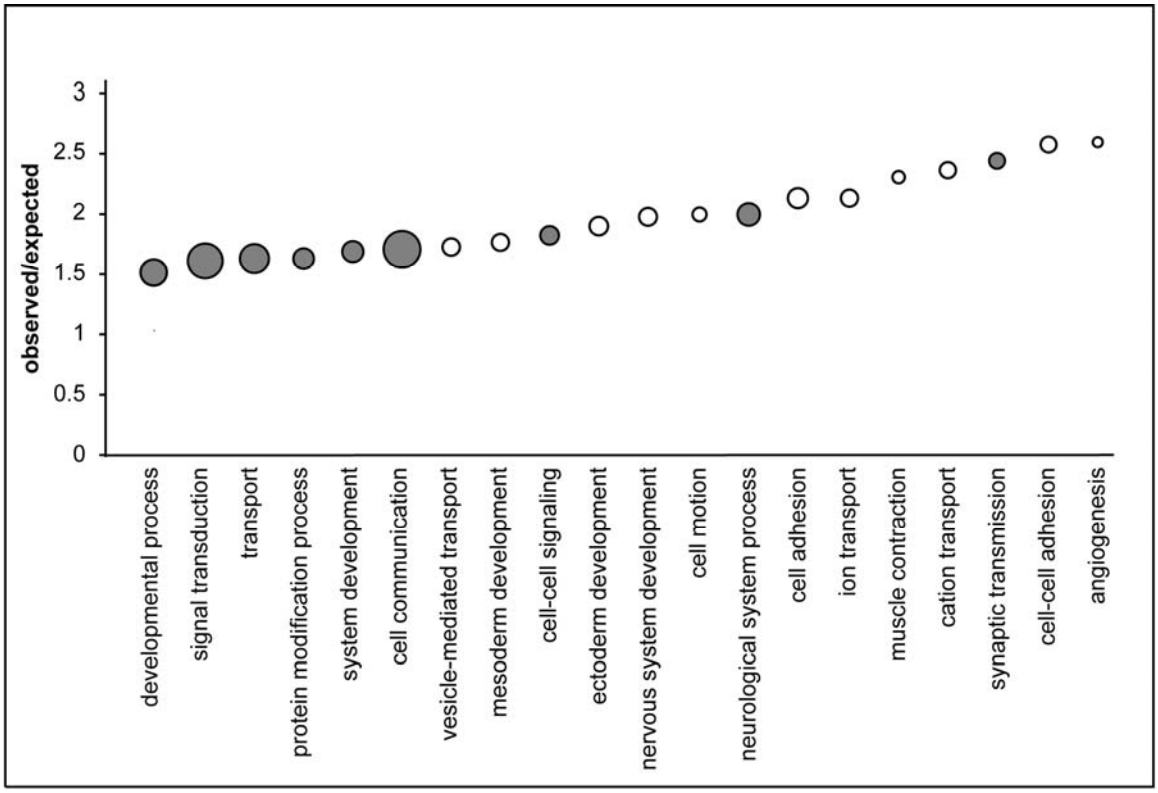


Figure 2.8. GeneOntology enrichment analysis of genes harboring TBRs. Among the 1,236 genes found to contain a TBR, 1,156 were mapped and annotated in the PANTHER Classification System6. Biological processes enriched at p-value < 0.01 after Bonferonni correction were displayed, as compared to all genes encoded in the human genome. The size of the bubbles represents the number of genes classified into a particular category, ranging from 25 to 348 genes. Closed bubbles correspond to biological processes that are significantly enriched after adjustment for gene length.

Discussion

In summary, we present an in vitro selection strategy for searching entire genomes for RNA sequences that enhance cap-independent initiation of translation. Using this technique, we identified >12,000 TEEs in the human genome, generated a high-resolution map of human TEE-bearing regions and validated the function of a subset of sequences in vitro and in cells. Our approach is time-effective, cost-effective, cell line–independent and scalable, making it an effective tool for studying translation mechanisms in other genomes.

Experimental

Library assembly and mRNA display selection

The pool of fragmented human genomic DNA was previously constructed with conserved sequences flanking the random region (32). The library was modified by overlap PCR to add all necessary sequence information required for mRNA display. This included a T7 RNA polymerase promoter site upstream of the random region and an open reading frame and photo–cross-linking site downstream of the random region. The open reading frame included a canonical AUG start site followed by a nucleotide sequence encoding a flexible linker and His-6 protein affinity tag. The library was amplified using the forward primer (5′-TTCTAATACGACTCACTATAGGGGGATCCAAGCTTCAGACGTGCCTCACTACG-3′) and reverse primer (5′-ATAGCCGGTGTCCACTTCCATGATGATGGTGATGGTGGGCCATGGCTGAGCTTGACGCTTTGC-3′). For each round of selection, 120 pmol of the dsDNA library was transcribed with T7 RNA polymerase into single-stranded RNA and purified after separation by 10% denaturing urea-PAGE. Purified RNA was photo-ligated to a psoralen-DNA-puromycin linker (5′-psoralen-TAGCCGGTG-(PEG9)2-A15-ACC-puromycin) by irradiating at 366 nm for 15 min. The RNA-DNA-puromycin product was ethanol-

precipitated, and the cross-linked RNA (400 pmol) was translated in vitro by incubating the library with micrococcal nuclease–treated rabbit reticulocyte lysate and [35S]methionine for 1 h at 30 °C. The mixture was then incubated overnight at −20 °C in the presence of KCl (600 mM) and MgCl2 (75 mM) to promote formation of fusions. The mRNA-peptide fusion molecules were purified from the crude lysate using oligo (dT)-cellulose beads (NEB) and reverse-transcribed with SuperScript II (Invitrogen) by extending the DNA primer (5′-TTTTTTTTTTTTTTATCCACTTCCATGATGATGGT-3′) with dNTPs. Fusion molecules containing the correctly translated His-6 tag were isolated on Ni-NTA agarose beads (Qiagen). Functional sequences were recovered by eluting the column with 500 mM imidazole, dialyzing the sample into water and amplifying the cDNA by PCR using previously described overlap PCR primers to add back the necessary sequences for mRNA display. The selection progress was monitored by measuring the fraction of 35S-labeled mRNA-peptide fusions that bound to and eluted from the oligo(dT) and Ni-NTA affinity columns. After six rounds of selection and amplification, the dsDNA library was cloned into a pJET plasmid (Fermentas), and individual isolates were sequenced at the Arizona State University core DNA sequencing facility.

Luciferase reporter plasmids

A monocistronic luciferase reporter vector with an unstructured 5′ UTR, that contains both a T7 RNA polymerase promoter and a vaccinia virus synthetic late promoter (slp), was constructed from a pT3_R-luc<IRES>F-luc(pA)62 luciferase reporter plasmid (35). The vector was first modified using PCR to exchange the T3 promoter with a T7 promoter (forward primer 5′-GATCCCGGGATTAATAACGACTCACTATAGGGAACAAAAGCTGGGTACCGG-3′ and reverse primer 5′-GATCCCGGGTGCGCGCTTGGCGTAATCATGG-3′). The resulting PCR product was cut with SmaI restriction endonuclease and recircularized using T4 DNA

ligase. A synthetic dsDNA molecule containing the slp promoter was inserted immediately downstream of the T7 promoter using KpnI and XhoI restriction sites. Finally, the Renilla luciferase gene was removed by PCR using forward primer 5′-ACTAGGATCCGCTTCTGTTGGGAAATGC-3′ and reverse primer 5′-CGCGGATCCAAGCTTATCGATACCGTCGAC-3′. The PCR product was cut with BamHI restriction endonuclease and recircularized using T4 DNA ligase. To assay for IRES activity, two additional luciferase reporter vectors were used, both of which contain a stable stem-loop structure in the 5′ UTR. The first vector was the pT7-stem_F-luc(pA)62 luciferase reporter plasmid described previously (26). This plasmid contains a T7 RNA polymerase promoter upstream of the stem-loop. The second vector was constructed by removing the stem-loop structure from pT7-stem_F-luc(pA)62 using StuI and XhoI restriction sites and reciprocally inserting it into the unstructured vector, immediately downstream of the slp promoter. Plasmids to assay for cryptic promoter activity were generated by removing the T7 and slp promoters from the unstructured vector using SmaI and BamHI restriction sites. T4 DNA ligase was then used to insert a 22-nucleotide spacer (5′-ATAGCGCCACCGAGATATCTGG-3′) in place of the promoters. To insert the human genomic sequences into the luciferase reporter vectors, the genomic fragments were amplified by PCR (forward primer 5′-TAGGGGGATCCCAGACGTGCCTCACTACGT-3′ and reverse primer 5′-TGGGCCATGGCTGAGCTTGACGCTTTGCT-3′) to add BamHI and NcoI restriction sites to the 5′ and 3′ ends, respectively. The PCR products were then reciprocally inserted into the vectors immediately upstream of the luciferase coding region by restriction endonuclease digestion.

Cell culture

HeLa cells, obtained from American Type Culture Collection, were maintained in DMEM (Invitrogen) supplemented with 5% (v/v) FBS (HyClone) and 5 μg/ml

gentamicin (Invitrogen). Cells were kept at 37 °C in a humidified atmosphere containing 5% CO2. The cells were free of mycoplasma contamination, as determined by PCR during routine monitoring of cell lysates.

Luciferase reporter assay

HeLa cells were seeded at a density of 15,000 cells per well in white 96-well plates 18 h before transfection. Cells were transfected with a complex of the luciferase reporter plasmid (200 ng) and Lipofectamine 2000 (0.5 µl) in Opti-MEM (Invitrogen) and immediately infected with the Copenhagen strain (VC-2) of wild-type vaccinia virus at a multiplicity of infection of 5 plaque-forming units per cell. Cells were lysed (6 h after infection) in the 96-well plates, and luciferase activity was measured using the Promega Luciferase Assay System with a Glomax microplate luminometer (Promega). Cell-free characterization of the top translation-enhancing sequences was performed using a Human In Vitro Protein Expression Kit (Pierce). Luciferase expression was achieved following the manufacturer's protocols using 300 ng of linear template for a 2-h transcription at 32 °C followed by a 90-min translation at 30 °C.

RNA characterization

A portion of the cells used in the luciferase reporter transfection studies were separately lysed to evaluate the quality of the cellular RNA. RNA isolation was performed using the PerfectPure RNA cultured cell kit (5 Prime) according to the manufacturer's protocol. Isolated RNA was reverse-transcribed with an oligo(dT) primer and SuperScript II (Invitrogen). Real-time PCR (iQ SYBR Green Supermix, Bio-Rad) was used to determine the mRNA levels of luciferase (forward primer 5′-GCTGGGCGTTAATCAGAGAG-3′ and reverse primer 5′-GTGTTCGTCTTCGTCCCAGT-3′) as well as the housekeeping gene hypoxanthine-guanine phospho-ribosyltransferase (HPRT, forward primer 5′-TGCTGAGGATTTGGAAAGGGTG-3′ and reverse primer 5′-

CCTTGAGCACACAGAGGGCTAC-3′). Using the ΔΔCt method, the amount of luciferase mRNA was normalized to HPRT mRNA levels. Luminescence values were then adjusted according to the normalized luciferase mRNA levels.

Sequence analysis

An in-house pipeline was used to process Illumina HiSeq sequences. First, base-calling and quality control were performed using the Illumina HiSeq2000 according to the manufacturer's instructions (Table 2.2). The average length of reads was 80 base pairs (bp). To detect and trim the PCR primers at both ends of each Illumina read, we used the 'cutadapt' program (http://code.google.com.ezproxy1.lib.asu.edu/p/cutadapt/) allowing a maximum of two mismatches. Both primers were detected in a vast majority of the reads (85% in *R0* and 98% in *R6*). However, multiple primers were found to be concatenated in some reads, which is common for HiSeq data. For these reads, we used 'cutadapt' iteratively until all primer sequences were trimmed. Finally, reads shorter than 35 bp or longer than 75 bp were discarded because they contained too many or no copies of the primers (Table 2.2). To ensure correct orientation for all reads, sequences were reverse-complemented if the 5′ primer was present at the 3′ end or the 3′ primer was present at the 5′ end.

All trimmed reads were aligned to the human reference genome build 19 (hg19) using iterative execution of 'bowtie' alignment and end trimmings (45). Sequentially, with one base at a time, 16 bp from the 3′ end, 5 bp from the 5′ end and another 15 bp from the 3′ end were trimmed from unaligned reads, which is done to ensure low-quality base calls do not interfere with sequence alignment. In all iterations, 'bowtie' was executed in "-n" mode with "-n 2 -e 70" setting. Reads uniquely mapped to exactly one location, 2–10 locations and more than 10 locations

in the hg19 genome were denoted as 'single-copy', 'low-copy' and 'high-copy' reads, respectively (Table 2.2c).

Based on reads mapped to the human genome, we used the command-line version of the CisGenome (46) to call peaks where *R6* served as the positive sample and *R0* served as the negative control sample; parameters were set as "-c 1 -m 10 -w 60 -s 20 -p 0.009948 -br 0 -ssf 0." Because TEEs are directional, we applied single-strand filtering and labeled a peak as 'forward' or 'reverse' depending on which strand of the genome it resided on. To further reduce spurious peaks, we required a peak to have a strand-specific global false discovery rate less than 10%, total number of reads greater than ten and at least one read present in the R0 library (Table 2.2d). The CisGenome program compared the normalized number of R6 reads with the normalized number of *R0* reads in a peak, which represented the fold enrichment level (Table 2.2e). Because repetitive elements can complicate downstream analysis, we focused on peaks derived from single-copy reads. Furthermore, single-copy peaks containing low-complexity sequences were detected using RepeatMasker with parameters "-noint -species human -q." Peaks with no repeat masked and with more than tenfold enrichment were called putative TBRs (Table 2.2f). Chromosomal distributions of TBRs were converted into ideograms using the Idiographica website (47).

We performed bionomical tests for evaluating the null hypothesis that TBRs are randomly distributed in the human genome. In this case, the random probability of a base to belong to a genomic category was first estimated using the RefSeq database to be 0.43, 0.005, 0.005 and 0.57, for genes (all exons and introns), 5' UTRs, 3' UTRs and intergenic regions, respectively. We also conducted Gene Ontology enrichment analyses to identify functional categories that were over-represented in the collection of genes found to harbor TBRs (Figure 2.8). We used

Gene Ontology classifications from the PANTHER (48) website and applied Bonferroni correction for multiple testing, using a cutoff P value of 10^{-3}. Enriched biological processes were reported (Figure 2.8). Because the naive library was generated by randomly sampling the genome, longer genes were sampled more often than shorter genes. To account for this gene-length effect, we constructed a background sample from the human genome that matched the length distribution of genes bearing TBRs and redid the Gene Ontology enrichment analysis. This process was repeated ten times. The Bonferroni-corrected P values from each analysis were combined using Fisher's method. Biological processes with $P < 0.01$ in at least one of these ten gene length–adjusted analyses or with combined $P < 0.05$ ($\chi 2$ test) were highlighted.

Construction and generation of Illumina library

The Illumina sequencing libraries were generated according to Illumina DNA Sample Kit Instructions (Illumina part 0801-0303). The protocol was modified such that enzymes were obtained from other suppliers, as previously described (49). Briefly, DNA from *R0* and *R6* was end-repaired and phosphorylated using the 'End-It' kit (Epicentre). The blunt, phosphorylated ends were treated with Klenow fragment (3’ to 5’ exo minus; NEB) and dATP to yield a 3’ A overhang for ligation of Illumina's adaptors. After adaptor ligation (LigaFast, Promega) DNA was PCR-amplified with Illumina genomic DNA primers 1.1 and 2.1. The final libraries were isolated (150–300 bp) from an agarose gel to remove residual primers and adaptors. Purified library DNA was captured on an Illumina flow cell for cluster generation and sequenced on an Illumina HiSeq 2000 following the manufacturer's protocols.

References

26. Jackson RJ, Hellen CUT, Pestova T V (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* 11(2):113–127.

27. Sonenberg N, Hinnebusch AG (2009) Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell* 136(4):731–745.

28. Shatsky IN, Dmitriev SE, Terenin IM, Andreev DE (2010) Cap- and IRES-independent scanning mechanism of translation initiation as an alternative to the concept of cellular IRESs. *Mol Cells* 30(4):285–293.

29. Spriggs KA, Stoneley M, Bushell M, Willis AE (2008) Re-programming of translation following cell stress allows IRES-mediated translation to predominate. *Biol Cell* 100(1):27–38.

30. Johannes G, Carter MS, Eisen MB, Brown PO, Sarnow P (1999) Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc Natl Acad Sci U S A* 96(23):13118–13123.

31. Roberts RW, Szostak JW (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc Natl Acad Sci U S A* 94(23):12297–12302.

32. Salehi-Ashtiani K, Lupta A, Litovchick A, Szostak JW (2006) A Genomewide Search for Ribozymes. *Science (80-)* 313(5794):1788–1792.

33. Kasowski M, et al. (2010) Variation in transcription factor binding among humans. *Science (80-)* 328(5975):232–235.

34. Korbel J, Urban A, Affourtit J, Godwin B (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science (80-)* 13(5849):350–358.

35. Gilbert W V, Zhou K, Butler TK, Doudna JA (2007) Cap-Independent Translation Is Required for Starvation-Induced Differentiation in Yeast. *Science (80-)* 217(5842):1224–1227.

36. Baranick BT, et al. (2008) Splicing mediates the activity of four putative cellular internal ribosome entry sites. *Proc Natl Acad Sci U S A* 105(12):4733–4738.

37. Moss B (2013) Vaccinia Virus : Vaccine for Tool Development. *Science (80-)* 252(5013):1662–1667.

38. Van Eden ME, Byrd MP, Sherrill KW, Lloyd RE (2004) Demonstrating internal ribosome entry sites in eukaryotic mRNAs using stringent RNA test procedures. *RNA* 10(4):720–730.

39. Mitchell SF, et al. (2010) The 5’-7-methylguanosine cap on eukaryotic mRNAs serves both to stimulate canonical translation initiation and to block an alternative pathway. *Mol Cell* 39(6):950–962.

40. Mokrejš M, et al. (2009) IRESite A tool for the examination of viral and cellular internal ribosome entry sites. *Nucleic Acids Res* 38:131–136.

41. Sakharkar MK, Chow VTK, Kangueane P (2004) Distributions of exons and introns in the human genome. *In Silico Biol* 4(4):387–393.

42. Akey JM, et al. (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* 19(5):711–722.

43. Sabeti PC, et al. (2006) Positive natural selection in the human lineage. *Science (80-)* 312(5780):1614–1620.

44. Traynelis SF, et al. (2010) Glutamate Receptor Ion Channels: Structure, Regulation, and Function. *Pharmacol Rev* 62(3):405–496.

45. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.

46. Ji H, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26(11):1293–1300.

47. Kin T, Ono Y (2007) Idiographica: A general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics* 23(21):2945–2946.

48. Thomas PD, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* 13(9):2129–2141.

49. Auerbach RK, et al. (2009) Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* 106(35):14926–14931.

CHAPTER 3

A leader sequence capable of enhancing RNA expression and protein synthesis in mammalian cells

Publication Note

This research was originally published in Protein Science. Wellensiek, B. P., Larsen, A. C., Flores, J., Jacobs, B. L., & Chaput, J. C. (2013). A leader sequence capable of enhancing RNA expression and protein synthesis in mammalian cells. Protein Science: 22, 1392–1398 © The Protein Society.

Introduction

Many applications in biotechnology require human proteins generated from human cells. Stable cell lines commonly used for this purpose are difficult to develop, and scaling to large numbers of proteins can be problematic. Transient expression can circumvent this problem, but protein yields are generally too low for most applications. Here we report a novel 37-nucleotide leader sequence that promotes rapid and high transgene expression in mammalian cells. This sequence was identified by in vitro selection and functions in a transient vaccinia-based cytoplasmic expression system. Vectors containing this sequence produce microgram levels of protein in just 6 hours from a small-scale expression in 10^6 cells. This level of protein synthesis is ideal for high throughput production of human proteins, and could be scaled to generate milligram quantities of protein. The technology is compatible with a broad range of cell lines, accepts plasmid and linear DNA, and functions with viruses that are approved for use under BSL1 conditions. We suggest that these advantages provide a powerful method for generating human protein in mammalian cells.

The synthesis of human proteins in human cells is necessary when properly modified protein is needed for biomedical assays (50–53). This requires developing

stable cell lines or engineered viruses (53), which is technically challenging, because it requires integrating a foreign gene of interest into the genome of the host cell or virus (54, 55). Even when properly constructed, stable cell lines are prone to contamination by viruses and microorganisms present in the laboratory environment. Consequently, human proteins are often synthesized in prokaryotic systems, even though these systems lack the capacity to produce post-translational modifications (56).

Here, we describe a novel 37-nucleotide RNA sequence that promotes strong protein synthesis in a vaccinia virus (VACV)-based cytoplasmic expression system. This system is ideal because of its activity in a broad range of mammalian cell lines, high expression capacity, and rapid timeframe (57). Biochemical analysis of our novel leader sequence reveals an unusual dual activity that leads to enhanced expression and translation. As a proof-of-concept, we show that 12 arbitrarily chosen human proteins express without the need for optimization, suggesting a straightforward method for generating human proteins in human cells.

Results

In a previous in vitro selection experiment, we isolated translation enhancing elements (TEEs) from the human genome (58). The selected TEEs were evaluated in a VACV cytoplasmic expression system (Figure 3.1), and found to enhance translation by up to 100-fold when compared with unselected sequences from the naïve library or a traditional VACV synthetic late promoter (SLP) alone. Subsequent screening led us to identify one sequence, hTEE-658, with unusually high activity in our VACV system (Figure 3.2). Comparative studies showed that hTEE-658 enhances translation more than 5,000-fold over a standard SLP VACV promoter. This observation suggested a possible strategy for increasing protein synthesis levels in mammalian cells.

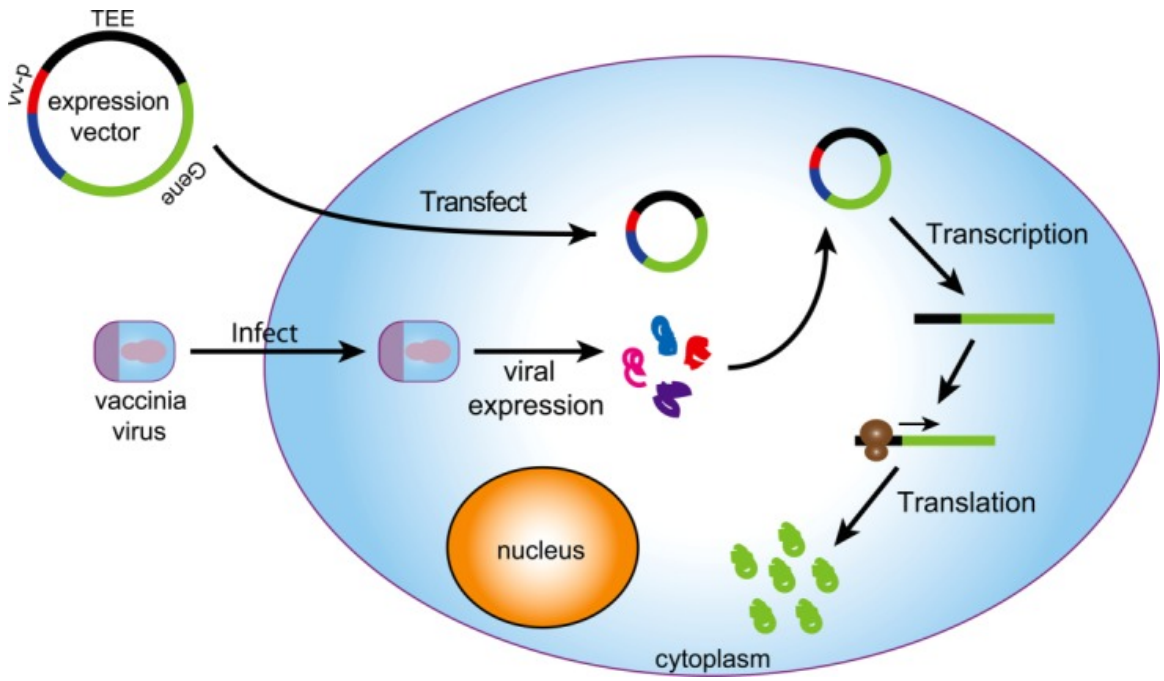


Figure 3.1. Vaccinia-based cytoplasmic expression of recombinant genes in mammalian cells. Cells transfected with a viral protein expression vector are infected with the vaccinia virus. Infected cells produce a viral RNA polymerase that recognizes a viral promoter in the protein expression vector and mediates the cytoplasmic transcription of gene-encoded RNA messages. Expressed mRNAs are translated using the translational machinery present inside the cell.

To understand the function of hTEE-658, we used quantitative real-time PCR (qRT-PCR) to measure RNA levels from cells transfected with a luciferase reporter plasmid containing sequences from the naïve library, selection output, and hTEE-658. After normalization, the hTEE-658 plasmid produces ~10-fold more RNA and leads to ~5-fold more luciferase than the most active sequence previously identified from our selection (Figure 3.2b,c). We confirmed by qRT-PCR that plasmid copy number was not altered (Figure 3.3), demonstrating that stronger mRNA expression and translation was not due to differences in plasmid replication by the virus. These results indicate that hTEE-658 enhances transcription and translation levels in the cell. The observation that a single sequence can affect both steps of protein

synthesis is unusual, but not unprecedented. We are aware of at least one other RNA motif, the TISU element, which functions in this capacity (59).

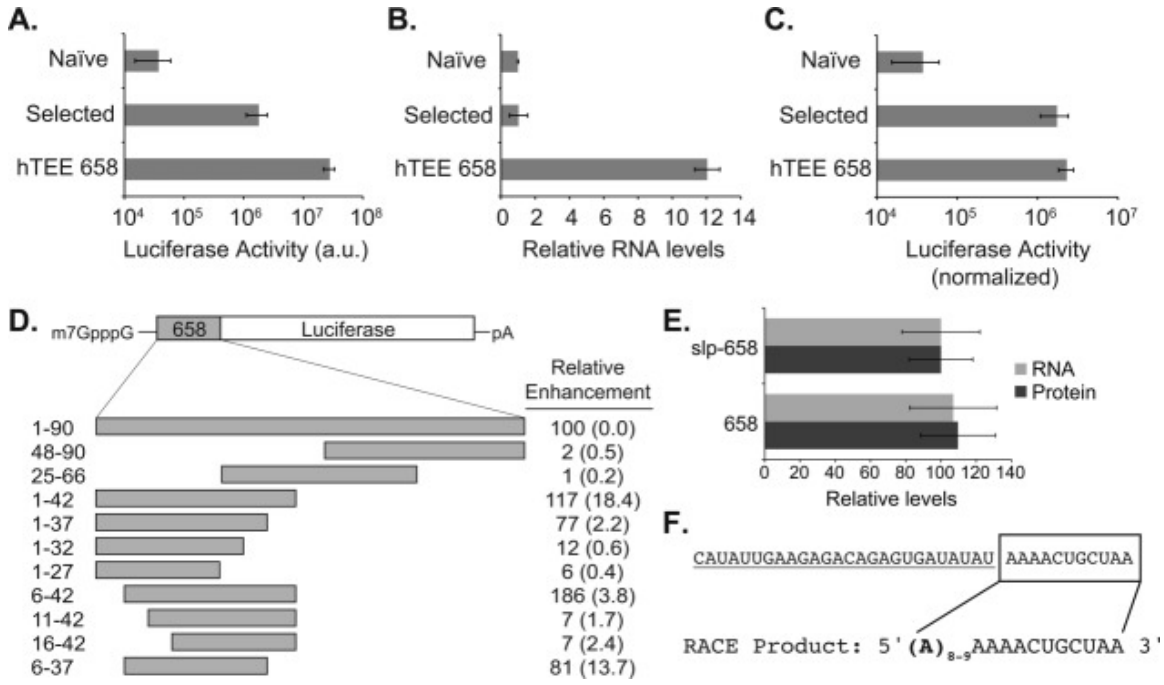


Figure 3.2. Functional characterization of hTEE-658. (A) Luciferase production driven by hTEE-658 compared to the average of 9 in vitro selected human TEEs and four randomly chosen human sequences from a naïve library. Results from the naïve library are equivalent to the SLP promoter alone. (B) Luciferase mRNA levels determined by qRT-PCR after normalization to HPRT. (C) Luciferase activity normalized to cellular mRNA. (D) Reporter constructs containing 5′ and 3′ deletions were used to identify the core functional domain of hTEE-658. Labels indicate the precise nucleotide fragment analyzed in vaccinia-infected cells. Relative enhancement is given as a percentage of full-length hTEE-658 with normalized percent error shown in parenthesis. (E) Luciferase mRNA and protein levels observed for vectors carrying and lacking the vaccinia SLP promoter upstream of hTEE-658. (F) 5′ RACE analysis was used to identify the viral promoter region (underlined) and ribosomal TEE (boxed) within the core functional region of hTEE-658.

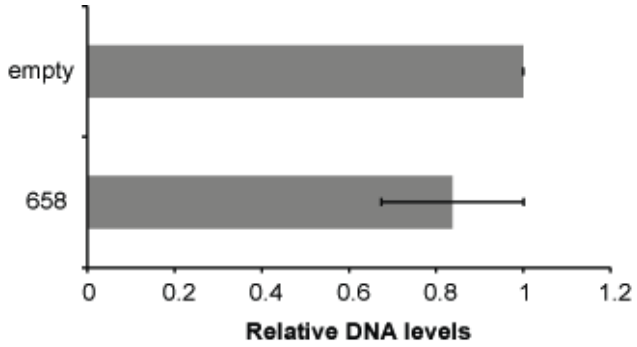


Figure 3.3. Plasmid levels in vaccinia infected cells. HeLa cells were transfected with luciferase reporter plasmids containing either the hTEE-658 sequence in the 5′ UTR or a 13-nt unstructured sequence (empty) and then immediately infected with wildtype (VC2) vaccinia virus. Plasmid DNA levels were measured six hours post-infection by realtime-PCR.

Next, we determined the minimal region required to achieve strong gene expression. A set of hTEE-658 variants were generated by first separating the parent sequence into the 5′ half, 3′ half, and central portion, which revealed that the functional region resided in the 5′ portion of the parent sequence (Figure 3.2d, Table 3.1). We then performed an incremental deletion analysis on the 5′ half to identify the minimal sequence necessary for function. Sequential deletions from the 5′ and 3′ ends allowed us to identify a core functional region of 37-nts spanning a boundary from residues 6–42. This region is ~2-fold more active than the full-length sequence and additional deletions that extend into either end led to significant drops in luciferase activity (Figure 3.2d). The remainder of our study focuses on the activity of the 37-nt core region of hTEE-658.

Table 3.1. Relative RNA levels (±SD) for reporter constructs containing the nucleotide fragments of hTEE-658 used for deletion analysis in vaccinia infected cells. Values are the average of 3 replicates determined by the ΔΔCt realtime PCR method.

Nucleotide fragment	Relative RNA levels
1-90	1.00 (0.12)
48-90	0.10 (0.01)
25-66	0.09 (0.01)
1-42	1.41 (0.05)
1-37	1.00 (0.30)
1-32	1.27 (0.35)
1-27	0.99 (0.01)
6-42	0.63 (0.09)
11-42	0.99 (0.16)
16-42	1.23 (0.17)
6-37	0.92 (0.13)

To verify that hTEE-658 functioned as a VACV promoter, we removed the vaccinia SLP promoter from the luciferase plasmid. Analysis of cellular RNA and luciferase activity values from vectors containing and lacking the SLP promoter showed no detectible difference in mRNA and protein levels (Figure 3.2e), confirming that hTEE-658 functions as a VACV promoter. To discern which region of the sequence is responsible for promoter activity and which region is responsible for TEE activity, we sequenced the 5′ end of the luciferase mRNA by rapid amplification of cDNA ends (RACE). cDNA sequencing indicated that transcription initiated within the AAAACUGCUAA portion of the sequence, which was preceded by a stretch of 8 or 9 non-templated adenosine residues (Figure 3.2f). We anticipated the presence of short polyA ends since VACV encodes strong poly-adenylation enzymes that modify the 5′ and 3′ ends of primary transcripts (60). This analysis suggests that the first 26 nucleotides of hTEE-658 function as a VACV promoter, while the last 11 nucleotides function as a TEE.

We established the activity of hTEE-658 relative to known VACV promoters using viral vectors that contain the SLP and I1L promoters alone and in combination

with hTEE-658 (Table 3.2). Vectors designed to express the luciferase and HIV-1 gag genes were tested in our cytoplasmic expression assay. After 6 h of expression, protein abundance was detected by western blot analysis using antigen specific antibodies. Analysis of the resulting gel indicates that vectors carrying hTEE-658, either alone or in tandem with SLP and I1L, produce substantial amounts of luciferase or HIV gag when compared with vectors containing only the SLP and I1L promoters alone (Figure 3.4). This result is consistent with our quantitative luciferase measurements.

Table 3.2. Common sequence elements used for protein expression.

Element	**Sequence**
I1L vaccinia promoter	CTATTGATATATTTGTATTTAAAAGTTGTTTGGTGAACTAA
SLP vaccinia promoter	AGCTTTTTTTTTTTTTTTTTGGCATATAAATGGA
T7 promoter	TAATACGACTCACTATA
EMCV IRES	GGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCC GGAAACCTGGCCCTGTCTTCTTGACGAGCATTCCTAGGGGTCTTT CCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTCGTGAAG GAAGCAGTTCCTCTGGAGGCTTCTTGAAGACAAACAACGTCTGTA GCGACCCTTTGCAGGCAGCGGAACCCCCCACCTGGCGACAGGT GCCTCTGCGGCCAAAAGCCACGTGTATAAGATACACCCGCAAAG GCGGCACAACCCCAGTGCCACGTTGTGAGTTGGATAGTTGTGGA AAGAGTCAAATGGCTCACCTCAAGCGTATTCAACAAGGGGCGGA AGGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGG CCTCGGTGCACATGCTTTACATGTGTTTAGTCGAGGTTAAAAAAC GTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAA ACACGATGATAAT
c-Myc Tag	GAACAGAAACTGATCAGCGAAGAGGATCTGTAATGA

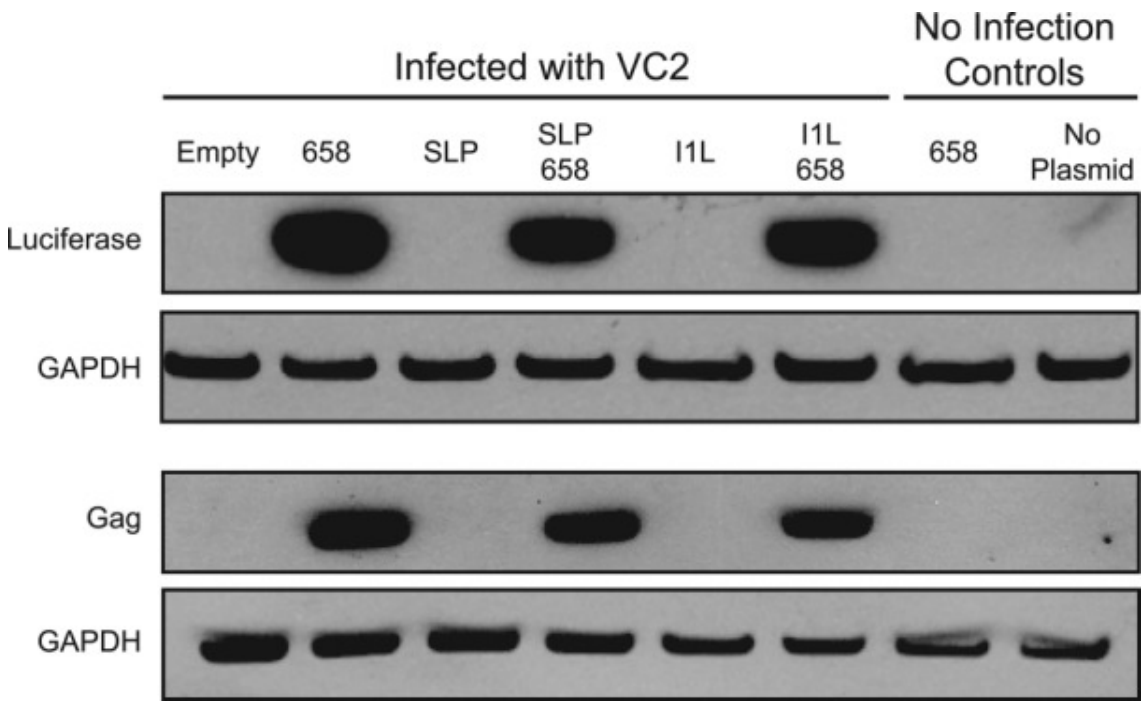


Figure 3.4. Western blot analysis confirms that hTEE-658 is a strong VACV promoter. Luciferase and HIV Gag proteins were produced in HeLa cells from vectors carrying hTEE-658, SLP, I1L or a combination of hTEE-658 in tandem with SLP or I1L. Western blot analysis was performed using antibodies directed against luciferase and HIV Gag proteins. GAPDH was used as a loading control. Empty refers to cells that were infected, but not transfected. No infection controls confirm that protein synthesis is VACV-driven. SLP and I1L protein is visible after prolonged exposure (data not shown).

Next, we evaluated cell line and viral strain compatibility by measuring luciferase production in three different cell types using three different viral strains. In this case, HeLa, HEK, and BHK cells were chosen for analysis with the VACV strains VC2, vTF7-3, and MVA. VC2 is a wild-type Copenhagen strain, while vTF7-3 is an engineered VACV designed to express the T7 RNA polymerase (61). MVA is a highly attenuated VACV that is non-pathogenic to humans and compatible with biosafety level 1 (BSL1) conditions (62). Plasmids carrying an internal ribosomal entry site (IRES) from the encephalomyocarditis virus (EMCV), in combination with a T7 or SLP promoter, were used as controls. The EMCV IRES is a ~500-nt noncoding RNA motif that is commonly used for protein synthesis in mammalian cells (63).

Time-dependent measurements were collected over the course of 24 h. In nearly all cases, hTEE-658 proved superior to the EMCV IRES with luciferase expression following a general trend of early rapid expression that plateaued after 6–9 h (Figure 3.5). While expression from the EMCV plasmid followed a similar trend, this plasmid generally required longer expression times and produced less overall protein (~10-fold). In only two cases were the hTEE-658 and EMCV plasmids similar; however, this required the engineered VACV strain vTF7-3, an efficient virus optimized for EMCV. Among the three cell lines, BHK cells consistently produced the highest levels of luciferase, consistent with previous VACV expression results (64). These findings indicate that hTEE-658 vectors produce significant quantities of protein in a time frame competitive with most prokaryotic expression systems.

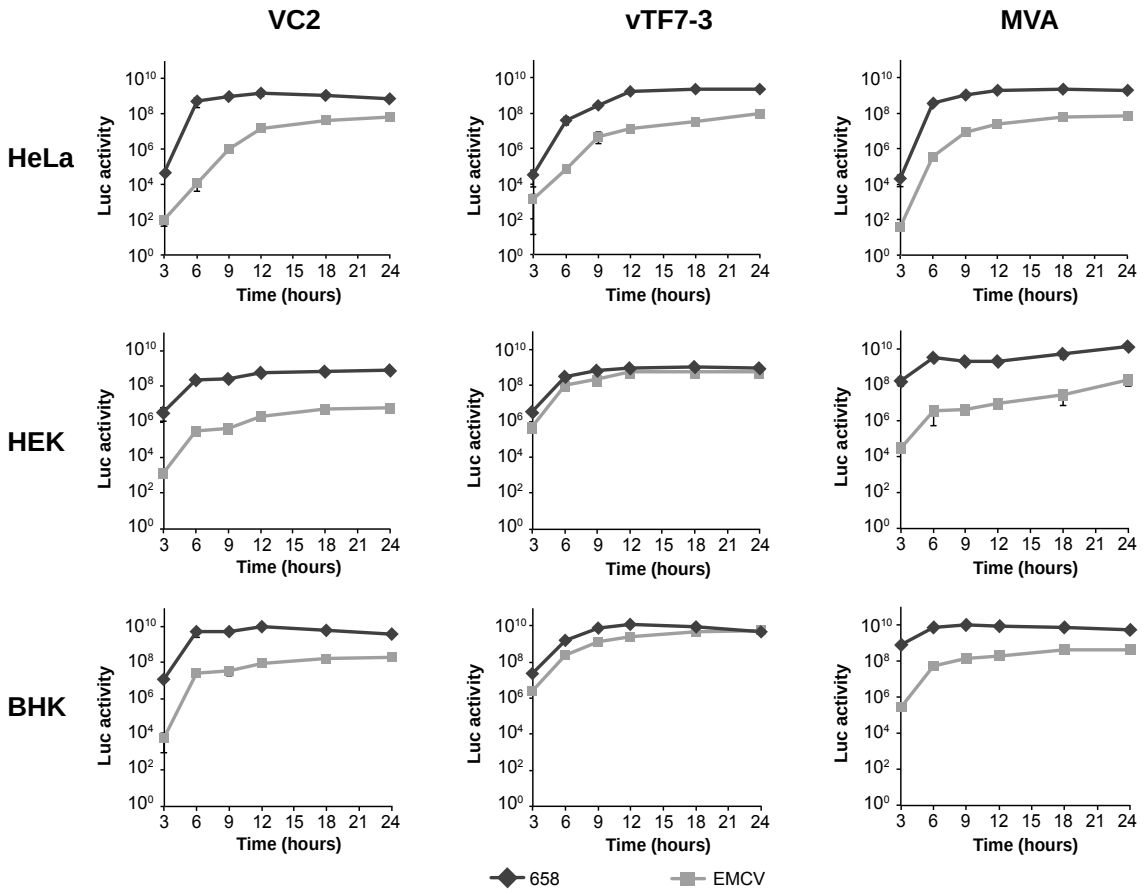


Figure 3.5. Time course analysis of luciferase production in multiple cell lines. Luciferase synthesis was measured in three mammalian cell lines (HeLa, HEK 293T, and BHK), each infected with three different vaccinia virus strains (VC2, vTF7-3, and MVA). Protein synthesis was monitored in triplicate over a 24-hour period using viral expression vectors engineered with hTEE-658 (diamonds) or the EMCV IRES (squares).

To demonstrate the potential for broad protein synthesis, 12 human proteins of different sizes and functional categories were arbitrarily chosen for analysis (Table 3.3). In all cases, the gene encoding sequence was inserted into an expression vector containing hTEE-658 upstream of the coding region, and protein production levels were monitored after expression in HeLa cells using a common c-Myc epitope tag. Western blot analysis of cell lysates indicated that full-length proteins were obtained in all cases (Figure 3.6). This result is important given the approximate 10-fold range in protein sizes. The ability of hTEE-658 to mediate the production of such

a variety of proteins from a plasmid expression system conveys a significant advantage over prokaryotic and cell-free expression systems, where success rates for human proteins are highly variable and typically less than 50% (65). For example, we have found that six of the 12 human proteins analyzed above (PI3K, SRC, P53, MYOT, HADH, and HRAS) are undetectable or barely detectable in a coomassie stained gel after expression in E. coli (data not shown).

Table 3.3. Full name and reference ID for all genes.

Gene	**Reference**
Firefly Luciferase	AB644228.1
HIV-1 Gag	See methods
v-akt murine thymoma viral oncogene homolog 1 (AKT1)	NM_005163.2
BCL2-related protein A1 (BCL2-A1)	NM_004049.3
hydroxyacyl-CoA dehydrogenase (HADH)	NM_005327.4
v-Ha-ras Harvey rat sarcoma viral oncogene homolog (HRAS)	NM_005343.2
mitogen-activated protein kinase 1 (MAPK1)	NM_002745.4
myotilin (MYOT)	NM_006790.2
nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha (NFKB-IA)	NM_020529.2
phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit gamma (PI3K)	NM_002649.2
tumor protein p53 (P53)	NM_000546.5
v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian) (SRC)	NM_005417.3
tumor necrosis factor receptor superfamily, member 21 (TNFRSF21)	NM_014452.3
tubulin polymerization-promoting protein family member 3 (TPPP3)	NM_016140.2

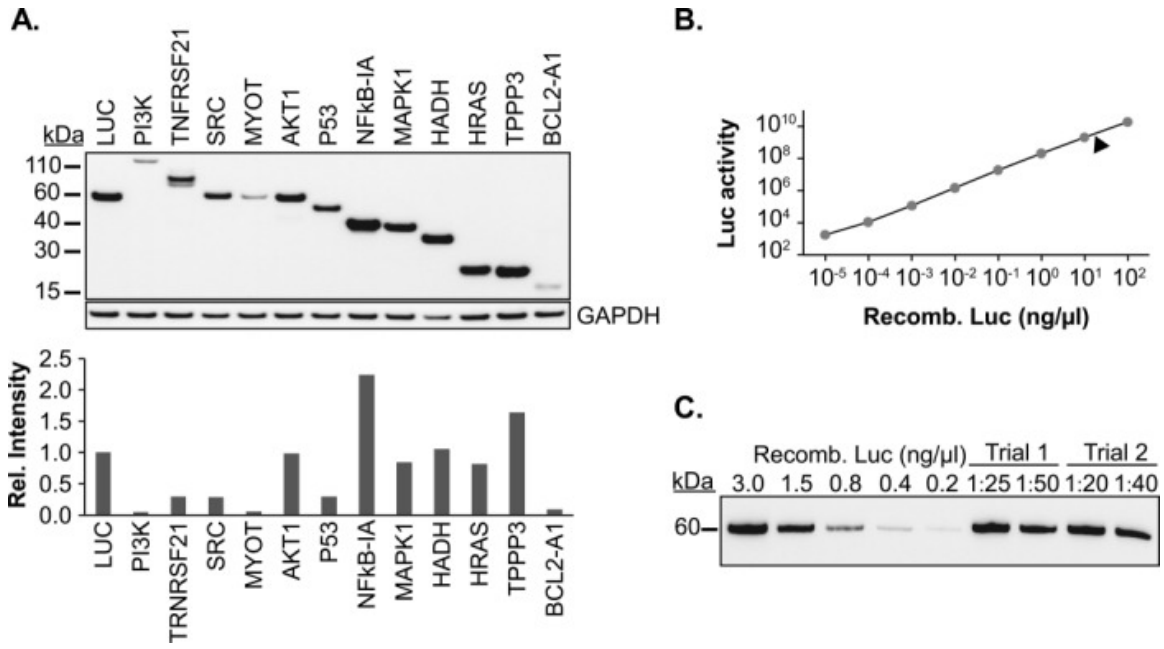
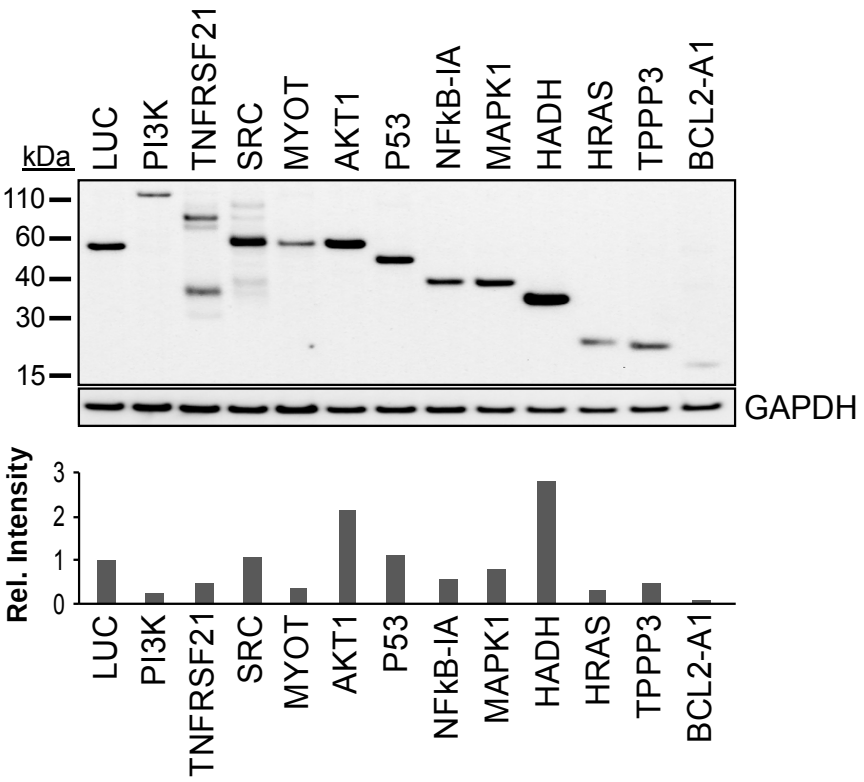


Figure 3.6. Synthesis of 12 human proteins in HeLa cells. (A) Twelve recombinant human proteins were generated from protein expression vectors engineered with hTEE-658. C-terminal myc-epitope tags were used to compare protein levels by Western blot analysis. Relative protein synthesis levels were determined by densitometry. (B) Quantification of luciferase production using a luciferase activity curve. The arrow indicates the average amount (20–50 ng/µL) of luciferase generated from 106 HeLa cells. This corresponds to 2–5 µg of total protein. (C) Western blot analysis showed strong concordance between the luciferase activity assay and protein synthesis levels from two independent trials. Protein samples were diluted to fit to the linear range of the Western blot.

Two different assays were used to quantify protein production in our expression system. First, luciferase enzyme generated from hTEE-658-mediated expression in HeLa cells was quantified by linear calibration using known amounts of commercial recombinant luciferase to measure enzymatic activity (Figure 3.6b). Second, Western blot analysis was performed using the same protein standards and anti-luciferase antibody to measure protein production (Figure 3.6c). Both methods gave similar results, yielding 2–5 µg of luciferase protein from 10^6 HeLa cells. This result indicates that all or nearly all of the luciferase protein was properly folded and enzymatically active. Comparison of the luciferase levels to the 12 human proteins observed in the western blot indicates that protein expression levels ranged from 0.1

to 2-fold, with NFkB-IA showing the highest levels of expression. These results suggest that this transient cytoplasmic expression protocol could produce milligram quantities of protein by scaling the expression to 10^9 cells.

To further simplify our expression system, linear DNA was assayed for activity in the cytoplasmic expression assay. Overlap PCR was used to add the hTEE-658 sequence and the c-Myc tag to our set of 12 human proteins. The linear DNA was transfected into HeLa cells and protein levels were analyzed by western blot after overnight expression. Analysis of the cell lysates revealed the presence of all 12 full-length human proteins (Figure 3.7). Only one protein, TNFRSF21, showed a truncated product that was presumably due to incomplete translation. Quantification of protein levels using the luciferase activity assay indicates that linear DNA produces ~10-fold less protein than plasmid DNA. Nevertheless, the simplicity of this approach makes it an attractive method for generating smaller amounts of protein for a large number of targets.



Figre 4.7. Synthesis of twelve human proteins from linear DNA. Twelve recombinant human proteins were synthesized in HeLa cells by transfecting PCR amplified DNA carrying hTEE-658 and a C-terminal myc-epitope tag. Western blot analysis was performed using an anti-myc antibody with GAPDH as a loading control. Relative protein levels were determined by densitometry.

Discussion

The ability to produce significant quantities of human protein in mammalian cells without the need for stable cell lines or recombinant viruses is a major advantage of our translation enhancing technology. This advance is based on the discovery of hTEE-658 as a short genetic sequence capable of rapid and high transgene expression in a VACV cytoplasmic expression system. Relative to common IRESs, like EMCV, hTEE-658 is substantially shorter (37 vs. >500 nts), making it easy to engineer into vectors. hTEE-658 is also more effective than EMCV at engaging the ribosomal machinery, and functions with viruses that are non-pathogenic to humans. We suggest that this new technology provides a versatile

platform for protein synthesis in mammalian cells. This could be especially useful in cases where prokaryotic and cell-free systems fail to produce protein or when post-translationally modified protein is needed for biological analysis. While further optimization, could lead to higher yields, the system is already ideal for routine protein synthesis.

Experimental

Cell culture

All cells used in this study were obtained from the American Type Culture Collection (ATCC). HeLa and HEK293 cells were maintained in DMEM (Invitrogen), while BHK cells were maintained in MEM (Invitrogen). Media was supplemented with 5% fetal bovine serum (FBS, HyClone) and 5 mg/mL gentamicin (Invitrogen). Cells were kept at 37°C in a humidified atmosphere containing 5% CO2.

Vaccinia virus strains

The vaccinia virus *Copenhagen* (VC2) and vTF7-3 viral strains were obtained from Virogenetics and ATCC, respectively. The modified vaccinia virus Ankara (MVA) was obtained from Dr. Bernard Moss at the National Institute of Allergy and Infectious Diseases. VC2 is considered a wild-type vaccinia virus, MVA is an attenuated vaccinia virus strain that is non-pathogenic in humans, and vTF7-3 is a recombinant vaccinia virus strain derived from the *Western Reserve* (WR) strain that has been engineered to express T7 RNA polymerase. Viral stocks were stored in MEM with 2% FBS.

Cytoplasmic expression system

Cells were seeded 18 h before transfection according to Table 3.4. Transfections were carried out using Lipofectamine 2000 (Invitrogen). In brief, complexes containing either plasmid or linear DNA and Lipofectamine 2000 were formed in Opti-MEM (Invitrogen). During complex formation, culture media was

removed from the cells and replaced with fresh Opti-MEM. Complexes were then carefully overlaid onto the cells. Plasmid DNA was obtained by standard mini or maxiprep (Qiagen), while linear DNA templates were generated by high fidelity PCR (accuprime taq, Invitrogen) using expression vectors as templates. Primers were designed so that the product included a T7 promoter, hTEE-658 core, gene of interest, c-Myc tag, and poly-adenosine track. Immediately following DNA transfections, cells were infected with VC2, MVA, or vTF7-3 at a multiplicity of infection (moi) of five plaque forming units (PFU)/cell for all 6 or 18 h assays and 30 PFU/cell for 24-h time course assays.

Table 3.4. Description of conditions used for various sized transfect-infect assays.

Size of Well	Cells Plated	Plasmid Template	Linear Template	Volume of Lipofectamine	Volume of Lysis Buffer
96-well	15,000	200 ng	---	0.5 µL	20 µL
24-well	200,000	800 ng	---	2.0 µL	50 µL
6-well	1,400,000	4000 ng	800 ng	10.0 µL	100 µL

Luciferase activity assay

Post-transfect-infect cells were lysed using passive lysis buffer (Promega). Luciferase activity was determined by mixing a portion of the lysate with the Promega Luciferase Assay System and measuring light production with a Glomax microplate luminometer (Promega). Luciferase concentration was quantified by comparison to a standard curve of QuantiLum Recombinant luciferase (Promega) generated using the manufacturer's recommended protocol.

RNA characterization

RNA was isolated from HeLa cells 6-h post-infection. Lysate from 2-wells of a 96-well plate was pooled and RNA isolation was performed using the PerfectPure RNA cultured cell kit (5’) following the manufacturer's protocol. Isolated RNA was reverse transcribed with Superscript II (Invitrogen) using an oligo (dT) primer. Quantitative

real-time PCR (iQ™ SYBR® Green Supermix, Bio-Rad) was used to measure luciferase mRNA levels, which were normalized to the housekeeping gene hypoxanthine-guanine phospho-ribosyltransferase (HPRT) using the ΔΔCt method.

End-mapping deletion analysis

To determine the core functional region of the 658 sequence, constructs were designed where various amounts of either the 5′ or 3′ end were removed. Each construct was built by Klenow extension of synthetic DNA oligos containing the desired fragment of hTEE-658 along with BamHI and NcoI restriction sites. The double-stranded DNA was restriction digested and ligated into a monocistronic firefly luciferase reporter plasmid carrying a vaccinia virus SLP upstream of the insert. Reporter plasmids containing truncated variants were assayed for activity by transfect-infect assay.

Expression vectors

Expression plasmids were obtained by engineering a monocistronic reporter vector with a leader sequence of interest inserted into the 5′ UTR. This vector contains a T7 RNA polymerase promoter site, a 5′ UTR which directly precedes an ORF containing the firefly luciferase gene followed by a poly-adenosine track. In order to test the expression of additional proteins, the luciferase was replaced with either HIV-1 Gag (a kind gift of Dr. Ralf Wagner of the University of Regensburg) or one of 12 human genes obtained from the DNASU Plasmid Repository (DNASU.asu.edu). A c-Myc tag was also inserted at the 3′ end of the human gene constructs to be used as an epitope tag for Western blotting. The full list of human genes is located in Supporting Information Table S3.

Western blotting

Proteins were expressed using the transfect-infect assay described above. After expression, HeLa cells were lysed with Passive Lysis Buffer (Promega) and

cellular debris was removed by centrifugation. For protein analysis, samples were diluted with NuPage 4× LDS sample buffer (Invitrogen) and proteins were denatured by heating for 10 min at 95°C before being run on a NuPage 4–12% Bis-Tris gel (Invitrogen). Proteins were transferred to a nitrocellulose membrane using the iBlot Gel Transfer system (Invitrogen). After blocking for 1 h at 24°C in TBS-T (20 mM Tris, 125 mM NaCl, pH 7.5, and 0.05% Tween20) supplemented with 3% milk, the membrane was incubated with the appropriate primary antibody concentrations overnight at 4°C. Membranes were then incubated with appropriate concentrations of goat anti-mouse or goat anti-rabbit HRP conjugated secondary antibodies (Cell Signaling) for 1 h at room temperature. Chemiluminescent signal was visualized after reaction with SuperSignal West Pico or Dura Chemiluminescent Substrate (Pierce Biotechnology). Anti-luciferase antibody was obtained from AbDSerotec, anti-GAPDH from Abcam, anti-Myc Tag (clone 4A6) from Millipore and the HIV-1 Gag antibody was generously provided by Dr. Hohne at the Charite Institute for Biochemie in Berlin, Germany. Where possible, membranes were cut to immunoblot for glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and proteins of interest separately. Alternatively, after the proteins of interest were detected the blots were stripped by incubating three times for 10 min with 0.2M glycine, 0.1% SDS, 2% Tween20, pH 2.2. After stripping, blots were washed twice for 10 min with phosphate-buffered saline (PBS), twice for 5 min with TBS-T and then placed back into block solution for 1 h before immunoblotting for GAPDH. Western blot signals were quantified using ImageJ to determine the relative intensity for bands of interest. Known quantities of QuantiLum Recombinant luciferase (Promega) were run as a standard curve to enable quantification of luciferase protein produced by transfect-infect assay.

RACE

RNA was isolated using the PerfectPure RNA cultured cell kit (5 Prime). RACE was performed with the 5′ RLM-RACE kit (Invitrogen) using total RNA following the small reaction protocol provided by the manufacturer with primers specific to the luciferase gene. RACE sequences were ligated into the pJET 1.2 vector (Fermentas), cloned, and sequenced at the ASU DNA Sequencing Facility.

DNA isolation and real-time PCR

Cellular and plasmid DNA was isolated from transfected HeLa cells 6-h post-infection with VC2 using the Trizol Reagent (Invitrogen) according to the manufacturer's protocol. Following isolation, DNA was ethanol precipitated and re-suspended in water. Quantitative real-time PCR (iQ™ SYBR® Green Supermix, Bio-Rad) was used to determine the levels of plasmid DNA as well as the housekeeping gene Ribonuclease P (RNase P) and normalized using the ΔΔCt method.

References

50. Zhu J (2012) Mammalian cell protein expression for biopharmaceutical production. *Biotechnol Adv* 30(5):1158–1170.

51. Swiech K, Picanço-Castro V, Covas DT (2012) Human cells: New platform for recombinant therapeutic protein production. *Protein Expr Purif* 84(1):147–153.

52. Schmidt FR (2004) Recombinant expression systems in the pharmaceutical industry. *Appl Microbiol Biotechnol* 65(4):363–372.

53. Wurm FM (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol* 22(11):1393–1398.

54. Colosimo A, et al. (2000) Review Transfer and Expression of Foreign Genes in Mammalian Cells. *Biotechniques* 29(2):314–331.

55. Yin J, Li G, Ren X, Herrler G (2007) Select what you need: A comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. *J Biotechnol* 127(3):335–347.

56. Demain AL, Vaishnav P (2009) Production of recombinant proteins by microbes and higher organisms. *Biotechnol Adv* 27(3):297–306.

57. Jacobs BL, et al. (2009) Vaccinia virus vaccines: Past, present and future. *Antiviral Res* 84(1):1–13.

58. Wellensiek BP, et al. (2013) Genome-wide profiling of human cap-independent translation-enhancing elements. *Nat Methods* 10(8):747–50.

59. Elfakess R, Dikstein R (2008) A translation initiation element specific to mRNAs with very short 5'UTR that also regulates transcription. *PLoS One* 3(8):e3094.

60. Schwer B, Visca P, Vos JC, Stunnenberg HG (1987) Discontinuous transcription or RNA processing of vaccinia virus late messengers results in a 5' poly(A) leader. *Cell* 50(2):163–169.

61. Fuerst TR, Niles EG, Studier FW, Moss B (1986) Eukaryotic transient-expression system based on recombinant vaccinia virus that synthesizes bacteriophage T7 RNA polymerase. *Proc Natl Acad Sci U S A* 83(21):8122–8126.

62. Hebben M, et al. (2007) High level protein expression in mammalian cells using a safe viral vector: modified vaccinia virus Ankara. *Protein Expr Purif* 56(2):269–278.

63. Elroy-Stein O, Fuerst TR, Moss B (1989) Cap-independent translation of mRNA conferred by encephalomyocarditis virus 5' sequence improves the performance of the vaccinia virus/bacteriophage T7 hybrid expression system. *Proc Natl Acad Sci U S A* 86(16):6126–6130.

64. Drexler I, Heller K, Wahren B, Erfle V, Sutter G (1998) Highly attenuated modified vaccinia virus Ankara replicates in baby hamster kidney cells, a potential host for virus propagation, but not in various human transformed and primary cells. *J Gen Virol* 79:347–352.

65. Büssow K, et al. (2005) Structural genomics of human proteins - target selection and generation of a public catalogue of expression clones. *Microb Cell Fact* 4:21.

CHAPTER 4

Evolving Engineered Polymerases for the Production of Synthetic Nucleic Acid Polymers

Introduction

Synthetic biology and synthetic genetics

Synthetic biology is often viewed as the deliberate redesign of biological systems to perform new functions that benefit mankind (66, 67). Over the years, significant effort has gone into engineering cellular systems with modular circuitry and establishing minimal genomes as frameworks for building organisms with designed functions. Existing SB strategies rely on engineered DNA, which is recognized by all forms of life, opening the possibility for alteration and assimilation. An alternative strategy is to explore the possibility of generating SB organisms using synthetic genetic systems. Synthetic genetics is an emerging field that merges chemistry and biology to design, generate, and explore the properties of non-natural genetic polymers. Significant achievements in this field include but are not limited to examples of modified base pairs, extension of the genetic alphabet beyond the four natural bases, and expanded forms of base pairing (68–72). Work has even begun to merge these genetic polymers with biological systems by providing the first evidence for the replication and translation of unnatural codons in bacterial cells (73–77). The extent of chemical dissimilarity between particular synthetic genetic polymers and natural DNA or RNA varies significantly, and greatly impacts how they can be used. Some synthetic genetic systems have the unique advantage of traveling unrecognized by natural enzymes. This trait provides a firewall that separates the natural and synthetic systems, offering many potential benefits, but also creating challenges when trying to merge the chemistry into a biological system.

In the Chaput laboratory, we study a synthetic genetic system comprised of α-L-threofuranosyl nucleic acids (TNA) (78, 79). TNA is one member of a general class of nucleic acid molecules termed xeno-nucleic acids (XNAs), where the natural ribose (or deoxyribose) sugar is replaced by a moiety (X) not found in natural genetic systems, see Figure 4.1. The substitution for threose leads to a nucleic acid polymer with a backbone repeat unit that is one-atom shorter than DNA or RNA, but is still able to cross-pair with natural polymers. Its ability to hybridize with complimentary DNA or RNA coupled with its chemical simplicity has generated considerable interest for TNA as a candidate RNA progenitor in origins of life research (78, 80, 81). TNA also has the advantage that it is generally unrecognized by natural enzymes. Significant effort has been made not only for the chemical synthesis of TNA building blocks that contain the natural A, C, G, and T bases, but also to develop enzyme based systems for the transcription of TNA polymers from a DNA template and the reverse-transcription of TNA polymers back into DNA. A long-term vision is to create a new generation of SB organisms that have genetic information encoded using TNA polymers. The work presented here helps progress towards this long-term goal by developing engineered polymerases that will improve the ability to generate and express information encoded in synthetic genetic matieral composed of threose nucleic acid (TNA).

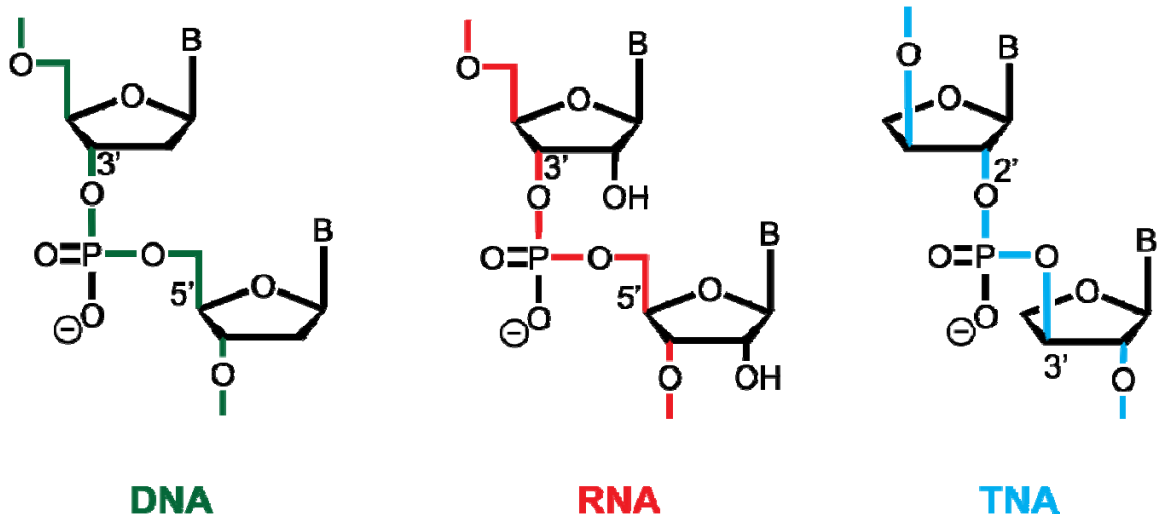


Figure 4.1. Backbone structures for DNA, RNA, and TNA. Constitutional structures for the linearized backbones of DNA (left), RNA (center), and α-l-threofuranosylnucleic acid, TNA (right). The phosphodiester linkage for TNA occurs between the 3′ and 2′ carbons on the threose sugar moieties, unlike DNA and RNA that share a common 5′-3′ sugar linkage. The backbone repeat unit for TNA is also one atom shorter than the backbone repeat unit found in RNA and DNA.

Polymerases

Polymerases are highly specialized enzymes that catalyze the synthesis of nucleic acid polymers by directing the addition of nucleotide 5′-triphosphates onto the 3′ end of a growing strand. Although all polymerase enzymes share a common purpose, nucleotide incorporation, and accomplish this task uses the same two-metal ion catalysis, a great deal of structural variation exists (82). Most generally, polymerases can be classified into three categories based on their substrates and products: DNA polymerases (DNAPs), RNA polymerases (RNAPs), and reverse transcriptases (RTs). DNA polymerases are further classified into families that share evolutionary and structural relationships, as well as similar biological functions. The most studied polymerases belong to Family A (found in prokaryotes, eukaryotes and bacteriophages) and family B (found in prokaryotes, eukaryotes, Archaea, and viruses) (83). Family C polymerases are involved in duplication of bacterial chromosomes and have no eukaryotic equivalents (84). The family D polymerases

are from Archaea (85), with families X and Y having function in DNA repair mechanisms (86, 87). The reverse transcription (RT) family consists of polymerases generally found in viruses, but this family also includes eukaryotic telomerases.

Polymerases serve the core biological function of maintaining both the storage and flow of genetic information. Polymerases have evolved the ability to perform these tasks with the catalytic efficiency, processivity, and fidelity required both to support current life and to foster adaptive evolution. The ability to copy long genes, and in some cases, entire genomes, with high speed and accuracy make polymerases valuable tools in biotechnology and molecular medicine. Since the initial discovery of this class of enzyme, research aimed at characterizing and improving their function as progressed with great fervor. While natural polymerases offer a variety of capabilities to choose from, researchers are continually searching for new variants with improved function (i.e., longer processivity, increased fidelity, and resistance to changes in salt, pH, temperature, chemical inhibition, and UV exposure). The very strengths that enable polymerases to faithfully propagate genetic information often limit their deployment in a laboratory setting. A particularly challenging need arises from the desire to engineer polymerases that incorporate unnatural nucleic acids, such as TNA.

Because polymerases represent an ancient class of enzymes that diverged long ago, each family, and even each individual polymerase, has a unique set of abilities. These subtle differences relate to their catalytic efficiency, template recognition, substrate specificity, processivity, and fidelity and help to determine which polymerase is most suitable as a starting point for engineering a desired function. Family A polymerases are the most studied and widely used in molecular biology and biotechnology. DNA pol I from *E. coli* was the first DNA polymerase to be characterized enzymatically (88) and provided the first crystal structure of a

polymerase (89). The structure resembles the shape of a right hand, with domains that are referred to as fingers, thumb and palm. While their three-dimensional structures can vary significantly, most polymerases share the same three core domains (90). The catalytic amino acids of the active site located in the palm, a thumb that binds double-stranded DNA and fingers where the incoming nucleotide binds and interacts with the template. Differences in the topology of the palm domain have led some polymerases to be referred to as left-handed (91). A separate structural category exists where the catalytic domain resides in a double-psi β-barrel structures (92). Family A polymerases also provided some of the first efforts relating to polymerase engineering, with the deletion of full domains. Two important examples of such engineering are the Klenow (93) and Stoffel (94) fragments. In both cases the 5'-3' exonuclease domain of DNA polymerase I, from *E. coli* and *T. aquaticus* respectively, was removed to greatly increase the usefulness of these enzymes for *in vitro* experiments. Improvements in genetic engineering led to more systematic investigations of individual point mutations, leading to great advances in the understanding of polymerase function.

Family B polymerases, isolated mainly from hyperthermophilic Achaean, are also widely used in molecular biology and biotechnology. Members of family B share the right-handed structural design with finger, palm, and thumb core domains. While family A members are the most frequent source for polymerase engineering, strong thermal stability, high fidelity, and strong affinity for primer-template complexes have driven much interest in family B polymerases (95, 96). Interestingly, family A and B polymerases were shown to have different biases for the incorporation of unnatural substrates (97). Early engineering efforts identified key mutations in Vent DNA polymerase, a family B member, that could alter substrate specificity (98). Polymerase screens later identified that family B polymerases, especially those with

mutations that alter substrate specificity could recognize TNA triphosphates and extend TNA polymers using a DNA template (99, 100). Significant efforts to improve these polymerases for TNA function have been reported and are discussed below.

TNA Replication

Nucleic acids in which the canonical ribose or deoxy ribose sugar is replaced by an alternative ring or other structure are typically poor polymerase substrates. This is due in part to the non-canonicai helical conformations observed even for compounds in the direct chemical neighbourhood of ribofuranose (79). One notable exception is TNA. The ability for TNA to form stable helical structures with complementary strands of itself, RNA and DNA, coupled with its chemical simplicty relative to ribose, has led to considerable attention for TNA as a possible RNA progenitor (78, 101, 102). The ability for TNA to cross-pair with RNA is remarkable considering that TNA polymers have a sugar-phosphate backbone that is one atom shorter. This ability also provides a plausible mechanism for the transfer of genetic information between these systems. Considerable efforts have been reported to develop a polymerase-mediated replication system that makes it possible to copy and store genetic information as TNA (17, 99, 103–106). Using an engineered form of the Archaean replicative DNA polymerase 9°N, known commercially as Therminator DNA polymerase (New England BioLabs, Inc.) with optimized conditions, TNA polymers can be generated from DNA templates sequential extension of a primer with TNA nucleotide triphosphates (107). We term this process TNA transcription. Single-stranded TNA can then be purified by denaturing polyacrylamide gel electrophoresis (PAGE) and reverse-transcribed back into DNA using Superscript II (SSII) (108). The overall fidelity of the combined transcription and reverse transcription process has been determined under a variety of conditions using DNA templates of a known sequence. DNA libraries composed of a three-letter genetic alphabet (A,T,C) can be converted to and from TNA with high

efficiency, and individual sequences replicate with >99% fidelity (104). Considering the structural deifferences between DNA and TNA it is surprising that with only minor modification natural polymerases can obtain this level of processivity and fidelity. However, this level of fidelity is significantly lower than natural polymerases and poses a potential problem for many applications of TNA, including the *in vitro* selection of functional TNA molecules. One likely source of poor fidelity is the incorporation of manganese in the TNA transcription reaction. Manganese is common in many XNA reactions as it helps relax the enzyme active site, increasing the processivity for TNA and other XNA polymers, but also decreasing the fidelity (109, 110).

Additional biases are observed that limit the ability to generate TNA polymers. For instance, synthesis of TNA libraries using unbiased DNA templates that harbor a four-letter genetic alphabet (A,C,T,G) leads to termination events that inhibit TNA synthesis. Investigation of this phenomenon discovered that TNA transcription could proceed to completion using DNA templates containing low numbers of isolated dG residues. Analysis of the fidelity revealed a G to C transversion mutation rate between 3-25% during TNA transcription, depending on the identity of the preceeding base in the template (104). DNA templates of known sequence that contain successive dG residues yield TNA transcription products indicative of transcription inhibition at the site of the G-repeats. This inhibition was found to be caused by tG:dG mispairing in the enzyme active site (103). This mispairing is likely a result of tG:dG Hoogsteen base pair formation, which adopts a similar structure to the canonical dG:dC base pair (111). This structural similarity is one reason why many polymerases (natural and engineered) struggle to read through templates that are rich in G nucleotides (112). Addition of the unnatural base analogue 7-deazaguanine (7dG) in the DNA template suppresses the tGTP misincorporation by

inhibiting the formation of Hoogsteen tG:dG base pairs (103). TNA transcription using DNA templates, even libraries of sequences, where all dG residues are replaced with 7dG proceeds with high efficiency and >99% overall fidelity.

A TNA replication system is highly valuable for the ability to select for functional TNA aptamers or catalysts. Although the system works remarkably well given the use of a DNA polymerase that harbors only a single point mutation, improvements to this system are still needed. Strategies such as the use of 7dG in the DNA template provide useful workarounds to address some of the issues, but this strategy is costly both for the added time and the cost of the 7dG analog. In order to generate large, unbiased pools of TNA for in vitro selection polymerases with enhanced activity, specifically in the absence of manganese ions, are required.

The ability of polymerases to efficiently copy genetic information has long made them important engineering targets to improve their functions for *in vitro* experiments. For this same reasons, engineered polymerases have emerged as powerful tools in the synthesis of unnatural genetic polymers (113). However, creating such enzymes is a challenging task because it is difficult to identify the genetic changes needed to elicit new functional activities (114). Poor recognition of the modified nucleotides by natural polymerases currently limits the development of expanded genetic systems and thus is a major obstacle toward achieving these ambitious goals. The large size of polymerases limits the ability to systematically examine all mutants in the surrounding sequence space. In addition, polymerase-engineering efforts are sensitive to enzyme concentrations, catalytic activity, and background noise that can interfere with assay detection. The following sections highlight some of the major advances and potential future hurdles in polymerase engineering.

Polymerase screening

The simplest approach to identify polymerases with new functions is to screen variants *in vitro* using a polymerase activity assay (PAA) (115). Many styles of PAAs have been developed to study polymerase function, ranging from laborious gel-based separation of extended products, to monitoring radioactive incorporation (116, 117), to a diverse assortment of fluorescence-based approaches (118–122). While screens are relatively straightforward to apply, even high-throughput screens are limited in their ability to search sequence space surrounding natural polymerases.

Because the numbers of mutations that improve activity are rare relative to those that diminish activity, a large number of variants must be screened. The number of unique single mutants of any particular polymerase is equal to 19x(N), where (N) represents the amino acid length of the polymerase, discounting the start codon. For polymerases that can be in excess of 1000 amino acids, a significant undertaking would be required to screen every single point mutant. Typical academic screens are carried out in eppendorf tubes or microtiter plates with library sizes ranging from tens to a few thousand mutations. Automated workstations can be used to increase throughput, but these systems come with significant cost due to the volume and quantity of reagents consumed in each assay. The theoretical complexity of a screen grows exponentially when one considers that multiple mutations are possibly required to impart a desired function. Screening through iterative rounds of single point mutants is one method to tackle this problem, and maximize the search of local sequence space, but this approach is inadequate for the identification of epistatic mutations (123). Screens often prove most successful when significant information regarding the sequence and structure of the polymerase is known *a priori*. Information about key residues located in the polymerase active site, regions that help form contacts with the template, or residues that show strong evolutionary

conservation can help researchers to limit the scope of a screen and increase the likelihood of identifying improved mutants. For instance, using the crystal structure of Taq DNA polymerase a small number of variants were screened to yield improved mutants that incorporate all four dideoxynucelotides at a more consistent rate during sequencing reactions (124). Although screening approaches are inherently low-throughput, they have been successfully applied to a diverse set of polymerases, altering characteristics such as thermostability, substrate specificity and fidelity.

A major ingredient in the ability to increase throughput of screening approaches has been the development of fluorescence-based PAAs. These assays make use of fluorescent nucleotides (121), fluorescent intercolating agents (115) and labeled nucleic acids that carry fluorophore-quencher pairs. Not all PAAs are equally effective, and are often too narrow in the types of polymerases or polymerase functions they can analyze (120). Marx et al. have reported the greatest strides at adapting high-throughput liquid-handling with fluorescence-based detection to identify novel polymerase variants through screening efforts. By monitoring the fluorescence increase of SYBR Green upon intercalation into duplex DNA during primer extension reactions, or qPCR, the Marx group has identified Taq polymerase mutants with improved mismatch discrimination (125), reverse transcriptase activity (115), and the ability to amplify damaged DNA with lesions that are often difficult for polymerases to bypass (126). Recently the Marx group has also reported the identification of T7 RNA polymerase variants with greater flexibility for the incorporation of 2'-modified nucleotides (127).

While screening approaches have proven fruitful for many engineering efforts, they are limited in scale both by the number of mutants that can be screened and by the reaction volume required to query each variant. While *in vitro* PAAs can be performed in what might appear to be small volumes, microliters quickly add up to

milliliters and possibly liters when large libraries are queried. For many types of polymerase function this is a costly, but not altogether unrealistic scale. But when commercially unavailable, non-natural substrates are being investigated there are often not enough triphosphates in the world to satisfy the demand. Several selection approaches have been developed to engineer polymerases for improved or altered function; either by letting bacterial cells take over the leg work, or by harnessing the power of bacteriophage display, or *in vitro* compartmentalization methods. While these selection approaches overcome some of the limitations of screens for the ability to search larger regions of sequence space, screening methods will always remain crucial for proper characterization of selection outputs.

In Vivo Selections

Selections have been performed *in vivo* using bacterial host cells transformed with synthetic polymerase sequences and screened against external pressure to identify functional mutants. Loeb et al. pioneered the use of an *in vivo* selection approach that relies on the genetic complementation of *E. coli* (128). The key to the technique is a temperature sensitive mutant of *E. coli* DNA pol I that is permissive to growth at lower temperatures, but has impaired function at 37°C making it lethal to the cell. Functional mutants of polymerases that can rescue the DNA pol I activity are identified following transformation by their ability to form colonies at the non-permissive temperature. This approach has been used to study various polymerase scaffolds and identified mutants of HIV type 1 reverse transcriptase with increased fidelity of DNA synthesis (129) and mutants of Taq pol I with lower fidelity of dNTP incorporation (130, 131). This method has also been used to help study the mechanism of polymerase activity selecting for all variants from a randomized library that retained function (132, 133). This enabled the authors to identify which amino acid residues of Taq DNA pol I were mutable and which mutations were permitted.

Genetic complementation was also extended in yeast to identify variants of human DNA polymerase η with an enhanced ability to bypass site-specific DNA lesions (134). Genetic complementation has proven to be a straightforward and convenient selection strategy to identify active variants from large pools of polymerase mutants. However, only polymerases that can rescue cellular polymerase activity can be identified and in its current form it is not amenable for non-natural templates or substrates.

A second *in vivo* selection approach involves a self-amplification strategy for the evolution of T7 RNA polymerases (135). Libraries of T7 RNA pol variants are cloned downstream from mutated versions of their own promoter sequence. Polymerase mutants capable of recognizing the altered promoter generate increasing copies of their own mRNA and by extension more protein. Extraction of the mRNA followed by reverse transcription and cloning leads to a new library that is enriched for functional variants (both polymerase and their new promoter). This approach was capable of selection from libraries with as many as 10^6 variants and led to the identification of one T7 RNA pol that recognizes a T3-like promoter and one that has an expanded promoter range, identifying multiple mutations of the T7 promoter sequence. While current in vivo selection approaches are limited to the types of function that can be achieved, they highlight excellent examples of ingenuity for selection design and the ease of their implementation is highly attractive. It is conceivable that these approaches could be re-imagined or extended to broaden their capacity for alternative polymerase function.

Phage Display

Phage display is a protein selection technology that uses bacteriophage to maintain a physical link between polymerase genotype and phenotype (136). While most phage display selections aim to identify binding interactions for the displayed

protein, phage display has also been adapted for the selection of functional polymerases by proximal display of the polymerase and a primer-template complex on the surface of the phage particle (137–139). Phage display is a widely used selection technique that enables large libraries to be searched to identify functional variants. The application to polymerase selection introduces new challenges as the enzyme and template nucleic acid strand must be linked. This approach tends to constrain the processivity of the selected enzymes because each enzyme acts on a single, well-defined template that is often very short. However, this does enable tight control over the primer and template sequence used in any round of selection and opens the possibility for templates that contain unnatural nucleic acids. Recovery of functional variants relies on the incorporation of a biotinylated nucleotide, which introduces background through chromatographic separation, but enables selection based on single nucleotide incorporation. Phage display also suffers from cross-reactivity of variants since they are not physically separated during the selection step. Cross-reactivity can lead to decreased enrichment per round and increase the number of variants that must be screened post-selection to identify optimal variants.

Phage selection has enabled the discovery of mutants with increased substrate recognition, including Stoffel fragments of Taq DNA polymerase with strikingly improved incorporation of NTPs (139), or the ability to incorporate 2′-OCH_3 substituted NTPs (140). By coupling an RNA template to the phage library, a selection was performed to identify variants of the Stoffel fragment with reverse transcription activity (141). Phage display has even been used to identify a polymerase variant of the Stoffel fragment capable of working through an unnatural base pair (142). While the Klenow fragment and Taq DNA polymerase were known to synthesize a propinyl isocarbostyril (PICS) self-pair, the Stoffel fragment could not. Using phage display a variant of the Stoffel fragment was identified that could extend

a primer template pair where the terminal 3' base of the primer was a PICS residue that self-paired to an identical residue in the template strand. The identified variant not only extended this self-pair, but was also shown to be capable of synthesizing a PICS residue into a growing strand. Although the efficiency of this reaction was limited, the results demonstrated the potential to use phage display to select for unnatural substrates. The ability to incorporate unnatural primers, templates or triphosphate substrates coupled with the potential for large library diversities makes phage display an attractive candidate for XNA work.

Miniaturization using artificial reaction compartments

Screening reactions in cells is a useful way to interrogate libraries for new enzymatic activity under conditions that promote multiple turnover kinetics (143). However, these systems are limited to substrates that can diffuse into a cell and remain in the cell after the reaction is complete. To overcome this problem, water-in-oil (w/o) droplet emulsions have been developed as artificial compartments with cell-like dimensions (144). Artificial w/o droplets have volumes in the picoliter to femtoliter range, which reduces the reaction volume (and cost) by > 10^5-fold compared to robotic or manual screens. Trapping the enzyme and substrate in a compartment allows one to select for multiple turnover kinetics, as opposed to phage-based selections, which use single turnover kinetics. Polymerase selections using w/o emulsion droplets that rely on bulk mixing techniques suffer from variations in compartment size (68, Figure 4.2). These differences complicate the selection by negatively effecting the distribution of functional molecules (i.e., active and inactive variants in the same compartment) and reducing the signal-to-noise ratio of the fluorescent readout. Using microscale emulsification techniques, monodisperse w/o droplet emulsions (146, 147) and water/oil/water double

emulsions (148) can be formed with to statistically reduce assay variation and favor predictable distributions of functional molecules.

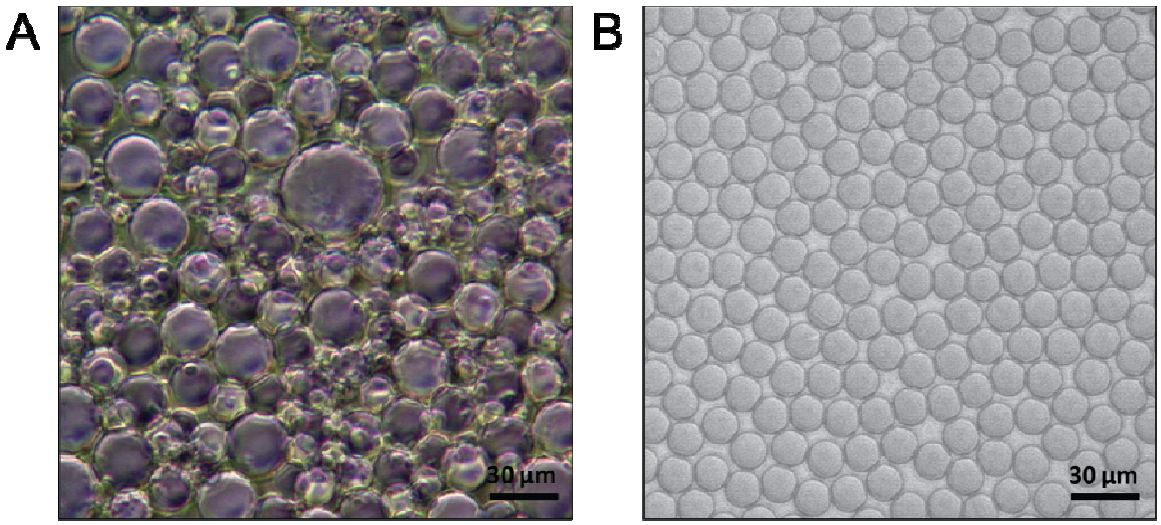


Figure 4.2. Water-in-oil emulsions. A) Emulsion formed using a bulk emulsification approach where aqueous and oil phases were mixed by repeated passage through a 12 µm filter in an extruder device following a literature approach (149). B) Emulsion formed by microfluidic flow focusing junction as described here in the experimental section.

Compartmentalized self-replication (CSR)

CSR is based on a feedback loop, where the ability of a polymerase to amplify its own genotype leads to its increased prevalence in the population (145). Since individual variants are encapsulated in emulsion droplets, functional variants are only capable of self-replication. In this system, polymerase variants expressed in *E. coli* are encapsulated in bulk w/o droplet emulsions with dNTPs, primers, and buffer. Heat is used to release the polymerase and plasmid from the *E. coli* without damaging the w/o compartment. Upon thermocycling, variants are challenged to amplify their own gene from the plasmid. The enzymes that are successful become enriched over the inactive population by making linear copies of their DNA sequence. The amplified DNA is then recovered, inserted into a new expression vector, and transformed back into *E. coli* to repeat the selection cycle. By controlling the

contents within the aqueous compartments or altering the extension conditions, the selective pressure of each round can be easily modulated. In this system the adaptive gains of useful variants translated into an increased genetic copy number, thus the prevalence of a variant is proportional to its catalytic activity. Despite its simplicity, CSR is limited to template-copying reactions that can undergo PCR; and therefore, may not be suitable for unnatural XNA backbones.

CSR was developed as a general strategy to improve the stability of DNA polymerases used in the polymerase chain reactions (PCR) and the original publication yielded variants of Taq DNA polymerase with increased thermostability and increased resistance heparin. Selections performed under increased heat denaturation have produced DNA polymerases with increased thermal stability, while selections performed in the presence of heparin have yielded variants that can amplify DNA directly from blood (145). Adapting this strategy with the addition of primer mismatch at the a 3' end yielded variants of Taq DNA polymerase with increased DNA lesion bypass, as well as the ability to incorporate unnatural substrates such as fluorescent dye-labeled nucleotide triphosphates (150). A significant limitation of the original CSR approach is the requirement that successful variants are capable of copying their entire genotype. This could prove overly challenging when trying to impart specificity for a new substrate of interest and precludes the ability to select for the use of unnatural templates.

An extension of CSR, short-patch compartmentalized self-replication (spCSR) requires only a small region (or patch) of the genotype to be copied by the polymerase variant. While this approach still relies on the DNA genotype to serve as the template, it reduces the burden on processivity. This approach was used to identify mutations in Taq DNA polymerase that increased its substrate spectrum to enable NTP incorporation while maintaining dNTP incorporation (151). Identified

mutants also displayed activity for additional 2'-substituted nucleotides. In a second example of spCSR standard dCTP was replaced with nucleotide triphosphates modified with cyanine fluorescent dyes inside the selection droplets. A mutant library of family B DNA polymerase from *Pyrococcus furiousus* (Pfu) was searched for the ability to replicate a defined segment of its encoding genotype with the modified substrates. With only two rounds of selection a variant was isolated with the ability to perform PCR amplification of long DNA polymers where all dC bases were substituted by Cy3- or Cy5-labeled dC equivalents (152). These demonstrations highlight the potential for the use of CSR-based selection approaches to identify variants with increased substrate specificity, including for unnatural substrates. The key drawback to CSR is the requirement for the encoding DNA to function as a template and the need for PCR-like amplification, a task that may be too challenging as a starting point when working with XNAs.

Table 4.1. Comparison of strategies for the identification of enhanced polymerase variants.

Strategy	Advantages	Disadvantages
Screening	- Each reaction is simple to setup - Direct control over template, substrate and conditions - Flexible	- Large volume of reagents (µLs) required to screen each polymerase variant - Labor intensive - Only "small" screen sizes are feasible
In Vivo Selections	- No cross reaction between mutants - Sufficient activity to replicate genome ensured - Screening process is simple	- Only natural substrates possible - Selection pressure difficult to control - Only a subset of polymerase scaffolds possible
Phage Display	- Straightforward to select for modifications in primer, template and/or nucleotide triphsophate, including modifications in both strands simultaneously; - Control of selection pressure; - Challenging selection conditions easy to apply - Selection possible without amplification and even with the incorporation of a single nucleotide	- Cross reactivity with sufficient enrichment, mandating the use of post-selection screen; - Complicated - Requires biotin-streptavidin purification
CSR	- No cross reaction between mutants - spCSR overcomes some of the limitations for processivity - Selection for unnatural substrates possible	- Requires amplification of double stranded DNA - Full length CSR has stringent demands on enzyme processivity - Selection pressure is difficult to modulate
CST	- No cross reaction between mutants - Primer extension only, no requirement for double stranded amplification - Selection for unnatural substrates possible	- Chromatographic separation based on biotin-streptavidin purification - Limited ability to modulate template
CFA (compartmentalized fluorescence amplification - outlined in this work)	- No cross reaction between mutants - Any primer/template complex - Any nucleotide triphosphate substrate - Control over processivity requirements by altering template length	- Complex selection strategy involving microfluidics double emulsification and FACS

Compartmentalized self-tagging (CST)

CST was developed to address some of the complications of CSR when evolving DNA polymerases that can copy DNA templates into XNA (113). Like CSR, polymerase variants expressed in *E. coli* are encapsulated in bulk w/o droplet emulsions to create a library of artificial cells in which the DNA and polymerase of each enzyme variant are compartmentalized in a single microreactor. Heat is used to release the plasmid and polymerase from the *E. coli*. Polymerases that extend a

biotin-modified DNA primer annealed to a region of the plasmid create a stable DNA-XNA hybrid duplex that can be separated by affinity purification on streptavidin-coated beads. Plasmids encoding active variants are recovered, their genes are PCR amplified, inserted into an expression vector, and cloned back into *E. coli* to repeat the process.

CST has been used to evolve DNA polymerases that display the ability for primer extension using multiple different XNA substrates with modest fidelity (113). While CST overcomes the requirement for double stranded amplification using CSR, use of the genotype as a template strand limits the ability to modulate the template used for selection. This prevents the ability to evolve polymerases with other types of XNA activity, such as the ability to copy XNA into XNA or XNA back into DNA. In addition, the technique requires chromatographic separation of extended primers, which often introduces significant background through nonspecific DNA binding to streptavidin-coated beads.

Polymerase diversification

A central challenge to any effort attempting to identify engineered versions of a natural polymerase with new or improved function is the decision of which mutations to search. The use of screening versus selection technology will help influence this decision, as screens can only cover a vanishingly small fraction of neighboring sequence space. A key aspect of all *in vitro* selection technologies is the ability to generate a pool (or library) of variants at the genetic level that can be used for input into the selection. Random mutagenesis is common in many forms of *in vitro* selection and can be used to mimic mutations that might arise spontaneously in nature. This is often accomplished through error prone PCR (ePCR), which employs PCR conditions that increase the error rate of a polymerase. However, for the selection of large proteins like polymerases random mutagenesis is unattractive due

to vast size of sequence space and the low frequency of useful mutations. Since all of the selection approaches described above rely on bacterial transformation, initial library diversity is limited to 10^9 members or less. While this number is vast compared to what is capable with screening technology, full randomization of only seven amino acid positions with any of the 20 natural amino acids would yield $20^7=1.3x10^9$ possible variants. For this reason researchers often choose to randomize a section of the polymerase, such as the active site, or just a few key residues in an effort to search a larger fraction of the possible variants that can be generated.

To simplify library generation for our selection experiments, we have studied what is known about the structure and function of family B polymerases as well as any previous efforts for their directed evolution to choose key amino acid residues for randomization. By taking a targeted approach, we are attempting to limit our search through protein sequence space to those regions that are most likely to yield mutations that modulate substrate specificity. Using sequence and structure alignments with comparison to known structure-activity data we have identified genetic "hotspots" that exhibit strong sequence and structural conservation among related polymerases from different clades including: archaea, prokaryotes, eukaryotes, and viruses. Our analysis indicate that the amino acid positions 408, 409, 485, 521, 664, and 730 represent ideal starting points for identifying polymerases with altered substrate specificity. Using cassette mutagenesis we can saturate each of these sites with all possible amino acid mutations and select for improved TNA activity. While structural and evolutionary comparison along with mechanistic analysis are important tools in polymerase engineering, to date our understanding of chemical theory does not provide the tools to predict the exact outcome of any amino acid mutation on the performance of any protein, including

polymerases. This means that a significant amount of trial and error is generally inherent in protein engineering endeavors.

Microfluidic-based droplet generation

Emulsion droplet technologies are an excellent way to isolate and manipulate nano-liter and sub-nanoliter scale reactions from each other and from their surroundings. Given a minimum presence of biochemical reactants within each volume, the tiny nano-liter reactions will contain a high signal over background by virtue of their scale. This makes them ideally suited for dilution free measurement of their contents and also makes each reaction more robust against inhibitory effects. For example, an individual cell trapped inside a small droplet containing reagents can be lysed and thermocycled to perform PCR without being inhibited by ancillary cellular contents. The technology naturally lends itself to high-throughput biological assays because the droplet size allows microliter reactions to be broken up into thousands to millions of reaction compartments that can be individually monitored in each experiment. Droplet formation is possible at frequencies of 5-10 kilohertz. Both CSR and CST style selections are performed using w/o emulsion droplets to physically separate each polymerase variant. To date these approaches have employed bulk-mixing techniques for emulsion droplets (113, 153). This strategy leads to significant variation in the size of each droplet, greatly affecting the concentration of polymerase and template (plasmid DNA) from droplet to droplet. Although this variation has not hindered the selection of functional polymerases, it introduces a significant variable. Microfluidics technology offers an ideal solution by tightly controlling droplet formation within a microfluidics device. Monodispersed droplets can be reproducibly generated with variation in the droplet diameter maintained within 3% (154).

Emulsion droplet generation requires two distinct solution phases, an aqueous phase which contains the biological reagents and an oil phase to surround the reactions, along with biocompatible surfactants to stabilize the droplet emulsion (155). Droplets are formed at droplet generating junctions within a microfluidic chip where both fluid phases are forced through a small channel intersection. Manipulation of phase viscosity, surface tension and the velocity of the dispersed and continuous phases allows precise control over the size of droplets being formed (156). The side-walls of the containing microfluidic channel or tubing need to be of the correct chemical structure to preferentially wet the oil phase and repel the aqueous phase (157). Proper surface wetting is crucial for both droplet formation and stability as well as to prevent sample carryover from reaction to reaction. Different from hydrophobic-hydrophilic interactions, the oils used in many biological emulsion droplet microfluidic systems are fluorinated to exhibit a strong fluorophillic binding to Teflon/PTFE tubing or other materials such as glass, PDMS and metal which have been stably modified with a fluoropolymer surface coating (158). The fluorinated oil has the added benefits of being very heavy to improve phase separation, exhibit a low-viscosity to reduce back-pressure, and have superior inert qualities with respect to anything suspended in the aqueous phase. These droplets can be fluorescently interrogated and sorted using custom fluorescence activated droplet sorting (FADS) approaches (159), or formed into double emulsions to more easily manipulate them with commercial technologies such as fluorescence-activated cell sorting (FACS) equipment (160).

Double emulsions are generated using multi-phase droplet microfluidic technology. By running the formed w/o droplets through another droplet generating junction, with the opposite hydrophilic and fluorophobic surface properties, it is possible to generate droplets within droplets with the critical aqueous phase inside

an oil phase which is inside an aqueous carrier phase. So called water-in-oil-in-water (w/o/w) double emulsions present a powerful means of conducting complex biological assays with the ability to maintain an outer aqueous phase that is compatible with technologies such as fluorescence-based cell sorting.

Recently, microfluidics-based droplet generation and FACS have been applied to select for enzymes where enzymatic conversion of substrate to product yields a fluorescent signal. However, the systems (microfluidics and FACS sorting) have not been applied to the evolution of polymerases. Not all fluorescence-based PAAs are amenable to miniturization in droplets, mainly due to poor signal-to-noise ratios that are confounded when populations of droplets with varying sizes are examined. Here we discuss a new selection approach that was designed not only to help generate polymerase variants for TNA transcription, but should also be adaptable to other XNA substrates and the ability to select for variants that accept XNA template molecules. We have validated the use of a donor-quencher fluorescence based PAA in emulsion droplets as a viable selection strategy to identify novel polymerases with XNA function, demonstrated enrichment values of approximately 1000-fold per selection round and used this technology to identify polymerase variants with improved activity for TNA. Using this system we were able to identify new variants of the 9°N polymerase that show enhanced activity in the absence of manganese ions, leading to high fidelity reactions that do not require 7dG in the DNA template.

Results

The simplest way to identify polymerases with new or improved functions is to screen large numbers of variants using a primer-extension assay. However, because mutations that improve activity are rare relative to those that diminish activity, a large number of variants must be screened to identify the novel, complex functions needed to further the field of synthetic genetics. To accelerate the pace of

polymerase discovery we set out to develop a novel microfluidics-based approach that could be used as a general strategy for evolving DNA polymerases with TNA activity. Although emulsion-based approaches have been applied to the evolution of polymerase enzymes, our work represents the first use of microfluidics for this purpose. This was made possible using a fluorescence-based PAA to monitor polymerase activity inside of emulsion droplets. A conceptual overview of our selection scheme is outlined in Figure 4.3. Specifically, a population of individual *E. coli* cells are encapsulated in their own w/o droplets generated using a microfluidic chip. Prior to droplet encapsulation the bacterial cells were trasnfromed with a library of polymerase variants. Encapsulation of an *E. coli* cell enables delivery of the expressed polymerase variant along with its encoding genotype. Each droplet also contains all of the reagents required to achieve a fluorescence signal from a PAA. Artificial w/o compartments that contain active polymerase variants extend a primer-template complex and elicit a fluorescence signal by disrupting a donor-quencher pair at the end of the template. Droplets that exhibit strong fluorescence are recovered using fluorescence-activated cell sorting (FACS). In order to enable droplet sorting on a commercial flow cytometer the surface characteristics of the w/o droplet must be changed from hydrophobic to hydrophilic. This is achieved using a second microfluidic chip to encapsulate the w/o population into monodisperse double emulsions known as water-in-oil-in-water (w/o/w) emulsions. The selected (FACS sorted) w/o/w compartments are extracted to recover the encoding DNA plasmids, which are transformed back into *E. coli* to initiate another round of selection.

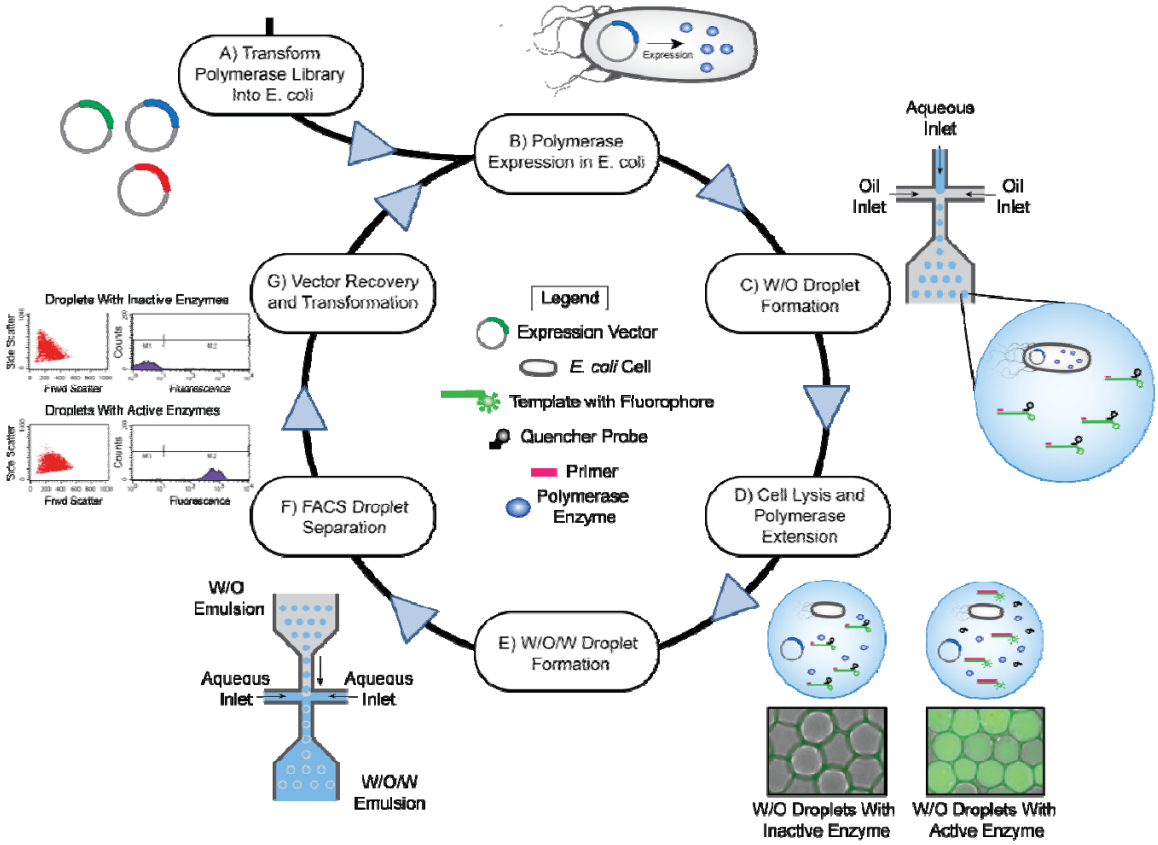


Figure 4.3. In vitro selection of XNA polymerases inside monodisperse compartments. A) A random library of polymerase variants is generated cloned into a protein expression vector and transformed into *E. coli*. B) The population of *E. coli* are grown to log phase in liquid media and induced with IPTG. C) Water-in-oil (w/o) emulsion droplets are generated in a microfluidic devise to produce artificial compartments that contain one or no *E. coli* cells and the substrates required for the polymerase activity assay. D) The w/o emulsions are heated to lyse the bacteria and release the polymerase into the droplet. The droplets are then incubated under conditions that allow for primer extension. Extension of the primer displaces the quencher probe from the template, which causes the droplet to fluoresce. E) After extension, w/o emulsions are passed through a second microfluidic device to generate water-in-oil-in-water (w/o/w) emulsions in a bulk aqueous phase. F) Fluorescence-activated cell sorting (FACS) is used to sort w/o/w droplets based on their fluorescence signal. G) DNA is recovered and transformed into *E. coli* to start a new round of selection or sent for sequencing to identify the polymerase variant.

Several *in vitro* selection strategies have been designed previously for use in identifying novel polymerase variants with improved function (114). Although these approaches have been successful and have made significant strides in the use of

synthetic genetic systems, we sought to develop a more flexible technology that removed any limitations on the desired substrate or template. For example, existing strategies are limited to the use of DNA or RNA templates delivered to the system as encoding plasmid or expressed mRNA (113, 145). With the future goal of evolving polymerases that recognize XNA templates we chose to validate universal fluorescence-based PAA that would function using substrate and template molecules that we deliver during the emulsification process. This also offers the opportunity to modify the template for each new round of selection, providing stringent control over the selective pressure. Our PAA is modeled on a standard primer-extension reaction using a template strand carrying a fluorescent label at the 5' end (Figure 4.4). Along with a primer, a third oligonucleotide that is complimentary to the 5' end of the template and carries a quencher dye at its 3' end is introduced into the system. Hybridization of the quencher probe to the template brings the quencher into close proximity with the template fluorophore and quenches the fluorescence signal. Polymerases that extend the primer to the end of the template disrupt binding of the quencher probe, causing the w/o compartment to generate fluorescent signal. The universal nature of this assay makes it amenable to just about any type of XNA function, which is quite different from past selection efforts where individual assays were developed for specific applications.

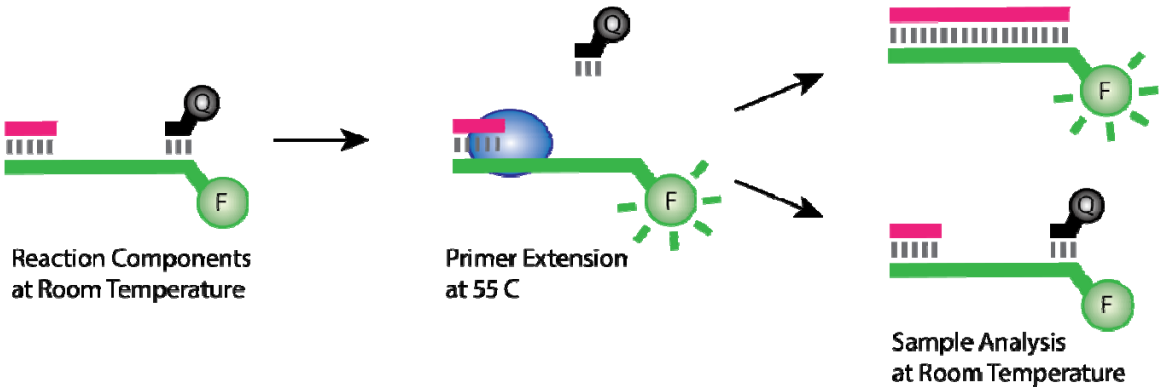


Figure 4.4. Fluorescence-based reporter assay for polymerase function. The reporter assays consists of a primer-template complex that contains a downstream fluorescent donor-quencher pair. At room temperature, the primer (pink) is annealed to a template (green), which is also annealed to a short probe bearing a 3' quencher (black). At elevated temperatures, the probe dissociates from the template and thermophilic polymerases have the opportunity to extend the primer to the end of the template. When the temperature is lowered for sample analysis, fully extended primers create a fluorescent signal by preventing the probe from re-annealing with the template. Primers that are not fully extended fail to generate an optical signal because of quenching by the probe.

Two key characteristics of our assay required experimental validation to optimize the system. The first parameter we explored was the length of the quencher probe. Many commercial fluorescence based PAAs, used for applications like qPCR, rely on polymerase functions such as exonuclease or strand displacement activity to help generate a signal. Taqman® probes are an example of such a PAA, relying on the 5'-3' exonuclease activity of *Taq* DNA polymerase to degrade a dual-labeled probe, eliciting a fluorescence signal by releasing the fluorophore from the quencher (161). Polymerases commonly used for XNA work are mutated to silence their exonuclease domains and often have weak or no strand displacement activity. To overcome this limitation we designed our quencher probe such that it dissociates from the template at the permissive temperature for polymerase extension (Figure 4.5 A and B). The second important parameter was the selection of an appropriate donor quencher pair. Although a simple binary fluorescence system could be achieved in bulk solution with a variety of different fluorophores and quenchers,

miniturization proved challenging. Even though microfluidics offers tight control and strong reproducibility over the size of emulsion droplets, some variation is inevitable when attempting to generate droplets at a rate that is feasible for selections. When droplets are later sorted by FACS, the distribution in droplet size leads to a distribution in the amount of fluorescence signal, limiting the ability to distinguish populations of quenched versus fluorescent droplets. A fluorophore-quencher pair with signal-to-noise ratio of ~10-fold for two solutions, one with an unextended primer and the other with a fully extended primer, yielded droplets with a continuum of fluorescence signal. By screening fluorophore and quencher pairs we were able to identify candidates with ~200-fold signal-to-noise ratio in bulk solution, which translated to two distinct populations of droplets that could be clearly distinguished by FACS (Figure 4.5 C and D).

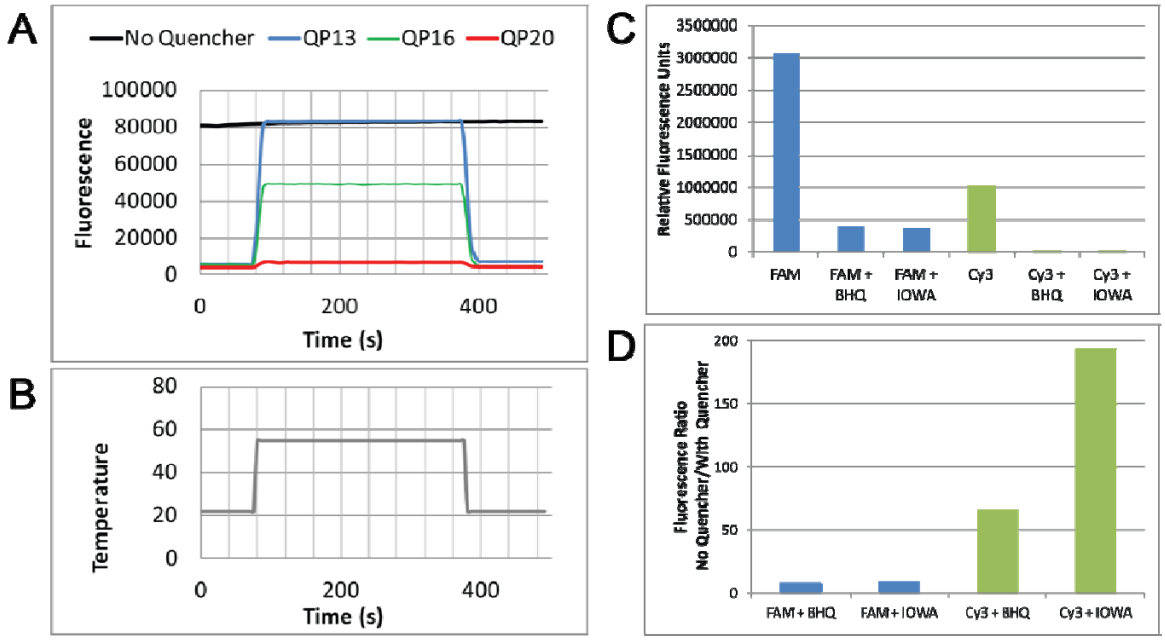


Figure 4.5. Optimizing thermal and fluorescence properties of the polymerase activity assay. A) Template fluorescence was monitored in a real time PCR instrument in absence of a quencher probe (QP) or in the presence of three QPs of varying length in 1x ThermoPol buffer. The DNA template was labeled at the 5′ end with a FAM residue and the quencher probes were labeled at the 3′ end with Black Hole Quencher Dye 1. Fluorescence was monitored as the temperature was raised to a permissive temperature for polymerase extension and then brought back down to room temperature (B). C) Solution fluorescence for template molecules carrying either a 5′ FAM or Cy3 label in the presence or absence of 14 nt quencher probes. Fluorescence was measured at 22°C. D) Calculated signal-to-noise ratio of template fluorescence without quencher probe divided by with quencher probe. See Table 4.3 for template and quencher probe sequences.

An additional key to this selection approach is the creation of w/o droplets that encapsulate one *E. coli* cell per compartment. Each *E. coli* cell will deliver the plasmid and polymerase (i.e., genotype and phenotype) for an individual library member. The compartments are formed using an emulsion droplet microfluidic device (Dolomite) that is attached to a series of pumps and mounted on a microscope equipped with a digital camera that allows us to monitor droplet formation in real-time. This process begins with a population of *E. coli* that is engineered to express a library of polymerase variants. The cells are grown to log phase and induced to express the protein of interest. After expression, the cells are

washed to remove lysed cellular debris and unwanted media, placed into the microfluidics device with the primer-template complex, buffer and substrates required for primer extension, and encapsulated into w/o droplets using a commercial fluorocarbon oil (3M) as the carrier phase (Figure 4.6). The w/o compartments are stable for weeks when kept at room temperature; however, when temperatures are elevated, the compartments are only stable for hours. This makes them an effective way to separate a population of enzymes into individual microreactors. Upon heating, the population of *E. coli* cells lyse without damaging the w/o compartments, maintaining the physical linkage between the polymerase genotype and phenotype. This step introduces the polymerase to the contents of the artificial compartment, which contains our PAA. Because the distribution of bacteria encapsulated in w/o droplets is dependent on cell density, single compartment occupancy is calculated using a Poisson distribution.

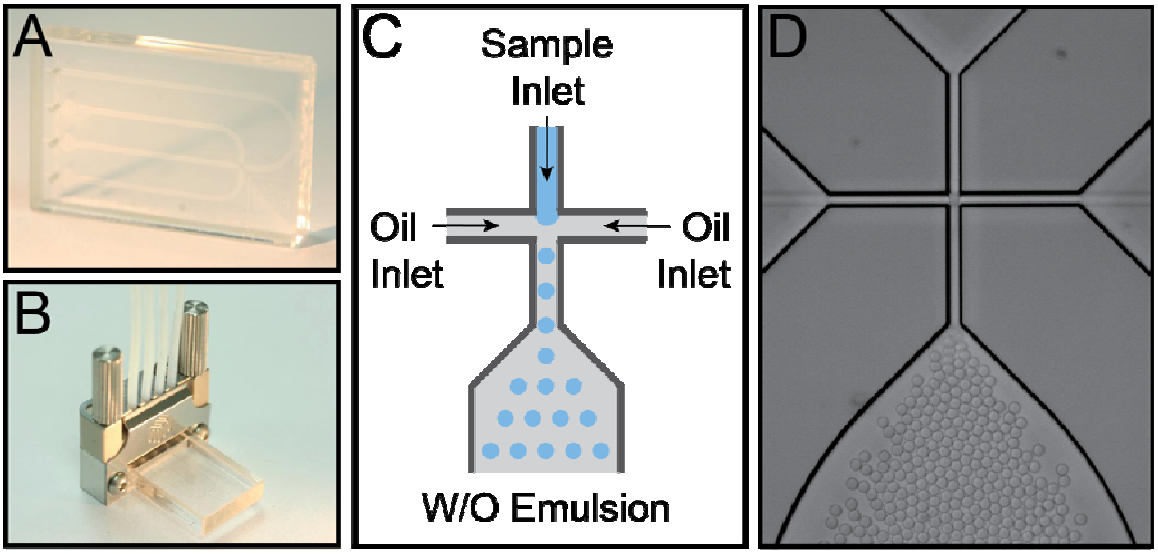


Figure 4.6. Microfluidics formation of water-in-oil (w/o) droplets. A) A commercial glass microfluidics device with a flow-focusing junction designed to form w/o droplets. B) Polytetrafluoroethylene tubing is connected to the microfluidic devise using connections that seal the tubing to the microchannels etched in the device. C) A cartoon depiction of the flow focusing junction inside the microfluidic device. The microchannels have a hydrophobic coating that enables the aqueous sample to form droplets inside of a bulk oil phase at the channel junction. The throughput and consistence of droplet formation is affected by the solution flow rate, surface tension, and surfactant composition. D) Bright field micrograph of the flow-focusing junction of the microfluidic device with stable w/o droplets (bottom).

To validate the ability to generate droplets containing single bacterial cells, we encapsulated *E. coli* engineered to express the green fluorescent protein (GFP) and measure cell occupancy and protein function using bright field and fluorescence microscopy. Overlaying images of w/o compartments enabled assessment of occupancy for various input cell densities (Figure 4.7). In doing so, we demonstrated that we are able to construct uniform w/o droplet emulsions that contain cell occupancy that follows a Poisson distribution, and that we are able to use *E. coli* as a vehicle to deliver expressed proteins. Next, we demonstrated that we could deliver functional polymerase that could be released from their *E. coli* host and function in our PAA.

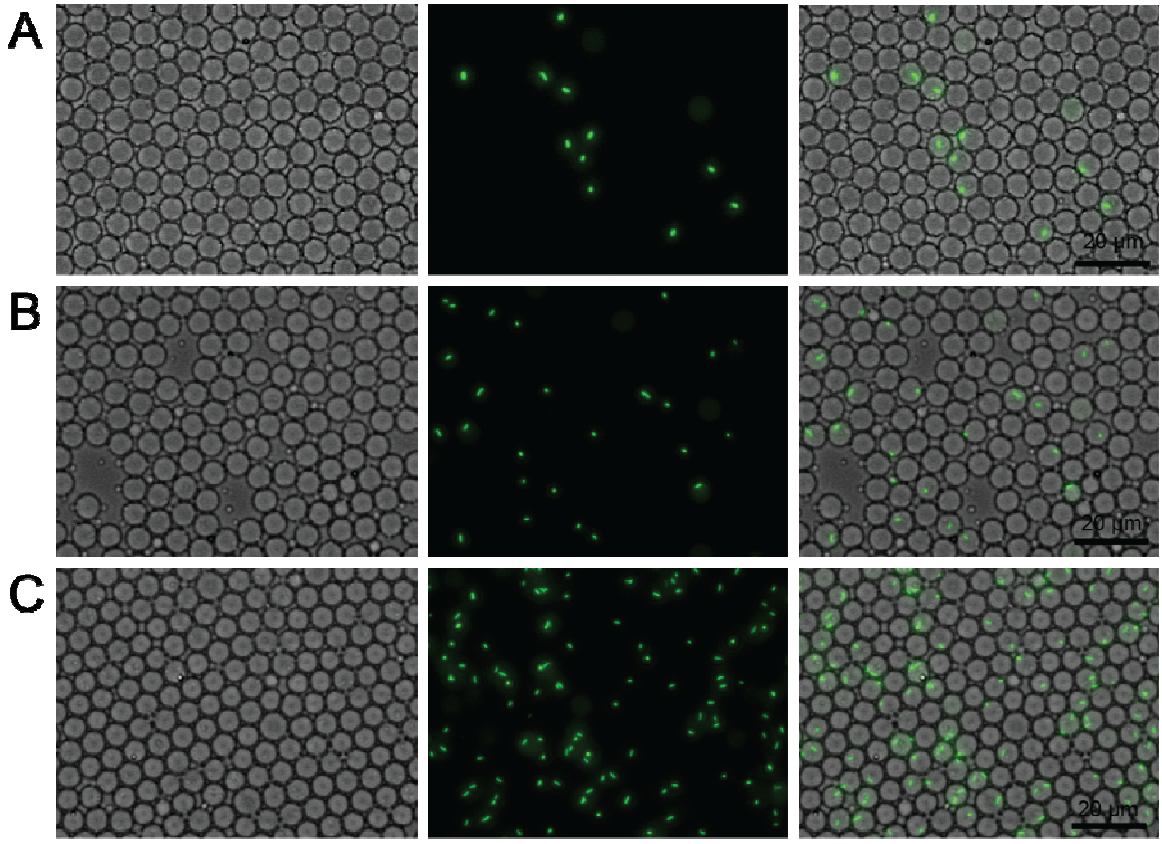


Figure 4.7. Determining distribution of *E. coli* cells in w/o emulsion droplets. *E. coli* cells were diluted to a final OD_{600} of A) 0.5, B) 1.0 and C) 2.0 prior to emulsification. Brightfield (left) and fluorescence microscopy (middle) were combined (right) to determine the fraction of droplets that contain 0, 1, 2 or multiple *E. coli* cells.

To demonstrate that bacteria can deliver functional polymerase to w/o droplets we encapsulated two populations of *E. coli* cells. The first population expressed the DNA polymerase 9°N bearing the mutations V93Q, Y409G, A485L, and E664K (termed 9°N-QGLK), which endows the enzymes with strong RNA synthesis activity. The second population expressed 9°N carrying only the A485L mutation (9°N-L), which is unable to extend a DNA primer-template complex with RNA and served as a negative control. We encapsulated the *E. coli* in separate populations of w/o droplets containing our PAA and NTP substrates. After droplet formation the emulsions were heated to 90°C for five minutes to release the expressed polymerase followed by incubation at 55°C for three hours. Next, the droplets were analyzed by

bright field and fluorescence microscopy (Figure 4.8). Droplets that contain no bacteria have very low fluorescence, indicating that in the absence of a functional polymerase the PAA remains quenched. With only NTP substrates added to the droplets, the 9°N-L polymerase is not expected to extend the primer-template complex. Low fluorescence from droplets containing cells expressing the 9°N-L polymerase demonstrates that a non-functional polymerase will not lead to a fluorescent signal. These droplets also highligh an important control, revealing that polymerases and nucleotide triphosphates endogenous to the bacterial are not sufficient to extend the primer-template complex. Finally, droplets containing cells expressing the 9°N-QGLK variant yield strong fluorescence, demonstrating that bacterial cells are able to deliver sufficient polymerase to function in our PAA.

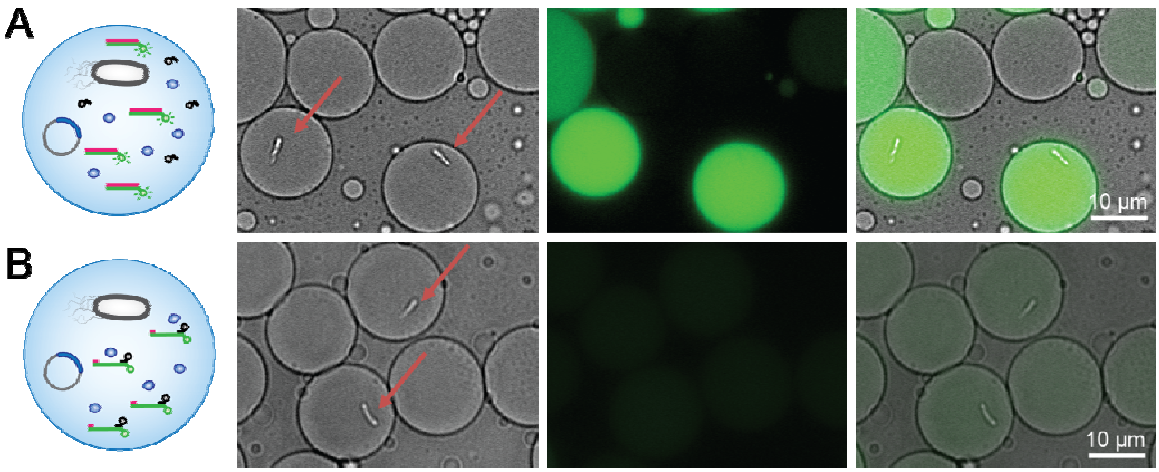


Figure 4.8. Delivery of functional polymerase enzymes to w/o droplets by encapsulation of bacterial cells. Water-in-oil (w/o) emulsion droplets provide artificial compartments where polymerase variants can be evaluated on an individual basis for a desired activity. After formation in a microfluidic device, the w/o emulsions are heated to lyse the *E. coli*, releasing the polymerase into the droplet. The droplets are then incubated under conditions that allow for primer extension. Active variants that extend a primer-template complex with NTPs yield a fluorescence signal, while inactive variants remain quenched. From left to right the panels show a cartoon depiction of the droplet, a brightfield micrograph with arrows indicating *E. coli* cells, a fluorescence micrograph of the same field of view, and an overlay of the brightfiel and fluorescence images. A) *E. coli* cells expressing the 9°N-QGLK polymerase. B) *E. coli* cells expressing the 9°N-L polymeraes.

Having successfully demonstrated that our assay functioned in emulsion droplets using polymerase enzymes delivered by bacteria, the next challenge was to recover droplets with functional polymerase and isolate their plasmid DNA. The fluorescence signal from the polymerase activity assay was used to partition droplets based on polymerase activity using FACS. Since FACS requires that samples be in an aqueous phase, we passed the w/o droplets through a second flow focusing microfluidics device to generate water-in-oil-in-water (w/o/w) emulsions (Figure 4.9). The size and fluorescence characteristics of the droplets were assessed by flow cytometry and w/o/w droplets were found to be stable for days at room temperature. Droplets with strong fluorescence were collected and the solution was extracted to recover the plasmid DNA from the bacteria. Successful transformation of these plasmids into *E. coli* generated a new pool of variants that are enriched in a desired activity. The fraction of droplets that display strong fluorescence could be used during a selection to monitor the enrichment progress through successive rounds and determine if changes in the selective pressure are needed. Selective pressure can be adjusted by modulating parameters such as the length of the template strand or the time of incubation at the polymerase permissive temperature.

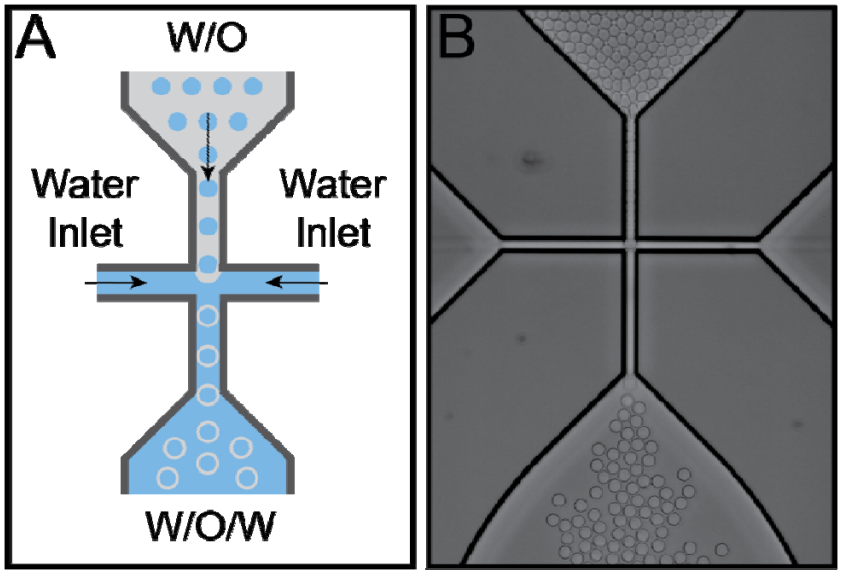


Figure 4.9. Microfluidics formation of water-in-oil-in-water (w/o/w) droplets. A) A cartoon depiction of the flow-focusing junction inside the microfluidic device used for w/o/w formation. The microchannels of microfluidics device have a native glass surface with no coating. The w/o emulsion enters the top of the flow-focusing junction, while a bulk aqueous phase enters from both sides. Optimization of solution flow rates and surfactant concentrations enables conversation of the w/o droplets to w/o/w droplets. B) Bright field micrograph of the flow-focusing junction of the microfluidic device with stable w/o/w droplets produced at the bottom of the image.

With each step in our selection scheme validated, we performed a single round mock selection for RNA synthesis using the 9°N and 9°N-QGLK polymerases to measure population enrichment per round of selection. Only the 9°N-QGLK polymerase can synthesize RNA. In addition to the point mutations required for function, we engineered the 9°N-QGLK plasmid with a unique restriction site to monitor enrichment after selection (Figure 4.10). We induced expression of both bacterial strains in liquid culture, and just prior to w/o droplet formation we mixed the strains at ratios of 1:100, 1:1,000, and 1:10,000 with the 9°N-QGLK strain present in lower abundance. The bacteria were to be encapsulated in w/o emulsions with the primer-template complex, buffer, and NTPs. The plasmid and polymerase were released from the *E. coli* by heat, and allowed to extend the DNA primer with RNA. The w/o droplets were examined by bright field and fluorescence microscopy to

ensure a proper distribution of functional w/o compartments. Next, the w/o droplets were encapsulated into w/o/w droplets and sorted by FACS. The selected population was recovered, plasmid DNA was extracted, and transformed into *E. coli* to generate a new population.

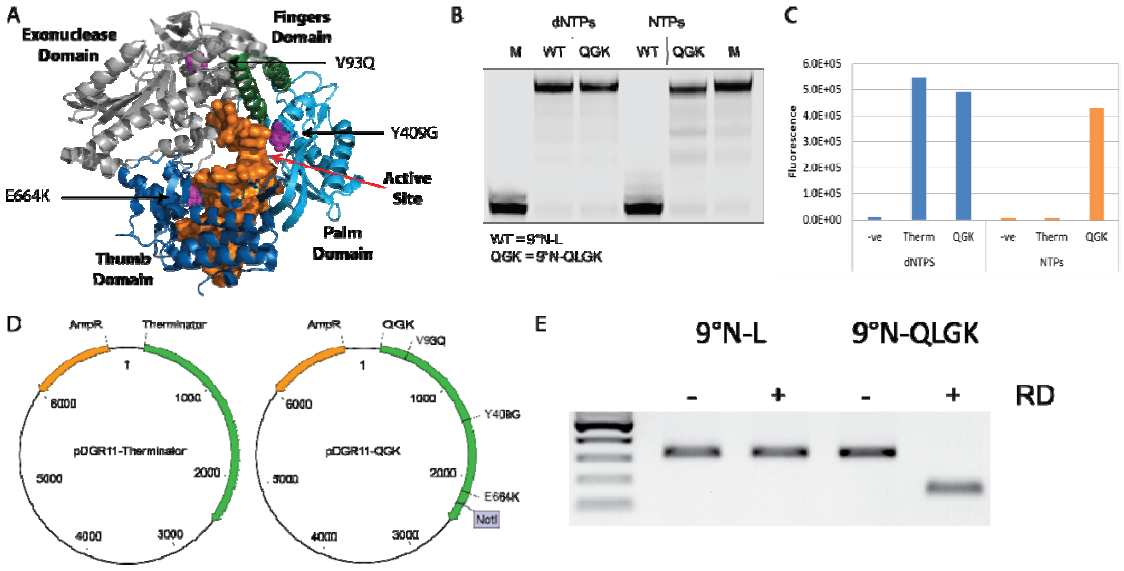
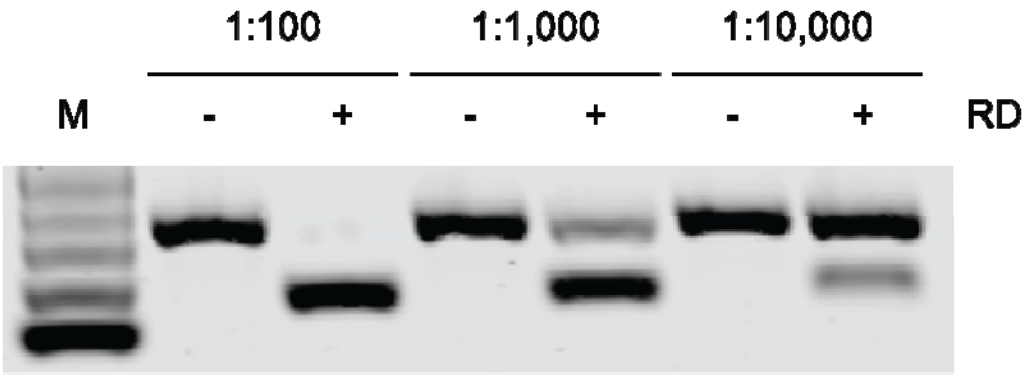


Figure 4.10. Single round mock selection constructs. A) Crystal structure of 9°N polymerase bound to a DNA template. Three key mutations required to enable RNA synthesis are highlighted. B) PAGE gel showing primer extension reactions for 9°N-L and 9°N-QGLK polymerases using dNTP and NTP substrates. C) Fluorescence PAA results for primer extension reactions for 9°N-L and 9°N-QGLK polymerases using dNTP and NTP substrates. D) Vector deisgn for both polymerase constructs. The only differences between the two plasmids are the three point mutations and the inclusion of a unique restriction site in the 9°N-QGLK coding region. E) Agarose gel of PCR products from the segment of the coding region surrounding the NotI restriction site. PCR products were run with and without restriction digest (RD) using the NotI restriction enzyme.

Enrichment is determined as the fraction of the pool that has become enriched in the 9°N-QGLK strain as a result of RNA synthesis activity and FACS selection. The gene encoding region of the plasmid population was amplified by PCR and subjected to a restriction enzyme digest. The cut and uncut fragments were run on an agarose gel. The ratio of cut to uncut DNA following PCR and restriction digest

was determined by densitomitry. Based on this single round mock selection we observed an average enrichment factor of ~1000-fold (Figure 4.11 and Table 4.2), which is consistent with literature using similar approaches (162).



Figre 5.11. Analysis of mock selection Output. PCR amplification product from the coding region of the selected pools run on an agarose gel with and without restriction digest using NotI restriction site.

Percentage of "active" QGLK variants in starting population	Average number of E. coli cells per emulsion droplet	Percentage of "active" QGLK varianst in selected population	Enrichment (n-fold)
1.0%	0.1	100%	>100
0.1%	0.1	76.7%	767
0.01%	0.1	15.8%	1580

Table 4.2. Quantification of enrichment from mock selection. Enrichment values are derived from the increae in the percentage of active 9°N-QGLK variants after one round of selection.

While results from the mock selection indicate a potential for strong enrichment, the simple binary design is a poor reflection of the background that would be observed with a complex library of variants To ensure that our microfluidics-based selection would scale to a library of polymerase variants, we performed a polymerase selection starting with a pool of ~8,000 unique variants. First, we used cassette mutagenesis to create a combinatorial library of all possible variants at amino acid positions 409, 485 and 664 of the 9°N DNA polymerase. Three long oligonuceotides were generated synthetically (IDT, gBlocks) such that

codons at positions 409, 485 and 664 were completely degenerate (Figure 4.13). Using overlapping PCR the three oligonucleotides were combined to generate one large fragment that was cloned into a plasmid containing 9°N using restriction digestion and ligation. During library creation we noted a significant number of unanticipated mutations present in the library. We attributed these mutations to our PCR strategy but noted that they resulted in a significant fraction of the pool expressing truncated protein. To improve the quality of the library, we subjected the pool to one round of selection where a DNA template and primer were coupled with standard dNTPs for the PAA. Following this pre-selection step we observed that almost all of the clones generated full-length protein. Additionally, sequencing indicated that a significant variation of mutations were present at the three desired positions, with some additional mutations spread throughout the palm, finger, and thumb domains.

Next, we used our pre-selected library and performed one round of selection with a PAA that included NTPs for RNA extension. By this point we had identified that the V93Q mutation in 9°N-QGLK was not required for RNA extension so we knew *a priori* that at least one variant in the library, the 9°N-GLK mutant, should function for RNA extension. However, it was also possible that other variants would display similar RNA extension activity. Following one round of selection we sequenced several clones and among several other sequences we identified the 9°N-GLK mutant. The ability to identify an active variant from a population of unique polymerases offers a more robust test of our methodology and suggested that our selection was capable of significant enrichment from a library of variants.

The original goal for development of our selection approach was to identify polymerases with enhanced activity for TNA polymerization. The best TNA polymerase identified to date, 9°N-L, suffers from sequence biases and is only

capable of transcribing complex libraries of TNA when manganese ions are in the reaction. The most striking challenge observed for TNA synthesis is the ability to incorporate tCTP when the polymerase encounters dGTP in the template (103). To overcome this limitation we performed a selection for TNA transcription in the absence of manganese ions. Our previously generated, and pre-selected library served as a promising starting point for this selection. A485L is a key mutation that has proven important for the ability of 9°N to synthesize TNA and in combination with mutations at positions 409 and 664 is also important for RNA synthesis. Not only do these three positions play a pivotal role in substrate specificity, but mutations at these positions are well tolerated by 9°N. We hypothesized that other combinations of residues at these three positions could help to enhance TNA synthesis.

Using our emulsion-based polymerase evolution strategy, we encapsulated a population of *E. coli* expressing our pre-selected library along with our PAA and TNA triphosphates in a w/o emulsion. The selection proceeded as described for the NTP selection above to enrich for polymerase variants capable of TNA extension. The template used for the selection contained only a single dG residue (excluding the primer region). After one round of selection, the sorted variants were amplified at the DNA level by cloning recovered plasmids into new *E. coli* cells. We obtained individual isolates by growing a portion of the cultures on solid media and picking individual colonies. Individual colonies were grown to recover the plasmid DNA for sequencing and to test their functional activity. For protein expression the clones were grown in deep-well plates, induced with IPTG, lysed, and centrifuged to separate the lysate from the cellular debris. Variants were tested for their ability to extend a primer using tNTPs. One variant in particular, carrying the mutations A485R

and E664I and retaining a Y at position 409 showed improved ability to read through sequential G residues in the template compared to 9°N-L (Figure 4.12).

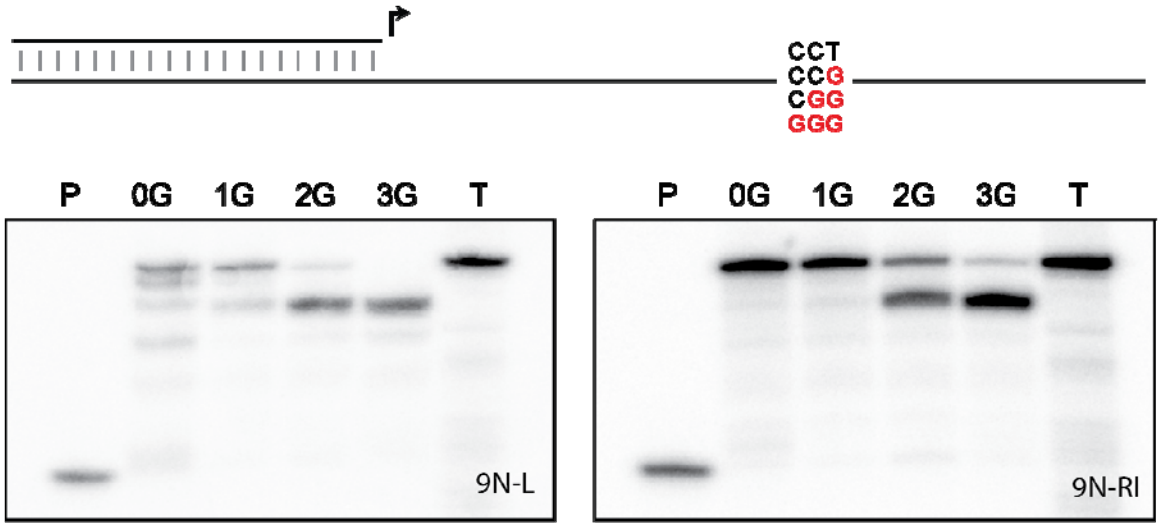


Figure 4.12. Screening functional polymerases. After selection, an enriched pool of polymerase variants are transformed into E. coli and plated onto agar plates supplemented with ampicilin. A) Individual colonies are then chosen, grown in liquid media supplemented with ampicilin and polymerase expression is induced during log growth phase with the addition of Isopropyl β-D-1-thiogalactopyranoside (IPTG). B) Following four hours of protein expression the cells are lysed by the addition of lysozyme for one hour followed by heating at 75°C for 15 minutes. The cellular debris is removed from the soluble polymerase enzymes by centrifugation. C) Polymerase variants are screened using the fluorescence reporter assay in 96-well plates.

Discussion

While simple screening approaches have yielded many useful polymerases, the capacity to push beyond modulation of polymerase activity and reach towards completely altered function is likely to require significant engineering and increasingly larger searches of sequence space. *In vitro* selections offer the ability to query larger libraries than would ever be possible with screening alone. Successful enzyme selection methodologies share the same two requirements: First, they must establish a strong link between the enzymatic function being selected (phenotype) and the genetic information responsible for the function (genotype). Second, they

must enable partitioning of active from inactive variants, often identifying only a small number of positive hits from a large population of variants. Emulsion droplet microfluidics technology is being established as a general tool for engineering enzymes with novel or improved activities. Our selection represents the first application of this technology for polymerases. Key to our technology is a novel, fluorescence-based assay to identify functional polymerases in monodisperse w/o compartments with sub-nanoliter reaction volumes. The compartments are formed in high-throughput using commercial emulsion droplet microfluidics devices, and recovery of functional variants is achieved with the aid of FACS. This technology aims to increase the number and types of selections that can be performed by enabling the use of XNA templates of any length or complexity in the selection step. This advance will open the door for identification of polymerases that can faithfully replicate new forms of genetic polymers.

In addition to their roles in the storage and flow of genetic information, nucleic acids have great potential for the development of high-affinity ligands (as aptamers), catalysts (DNAzymes and RNAzymes), nanostructures, and nanomaterials. However, applications based on natural nucleic acids are limited by their narrow chemical diversity and poor biological stability. There is increasing interest in the ability to generate modified nucleic acids to expand this chemical diversity and for use as expanded genetic alphabets. As sophisticated new chemistries are generated the demand grows for advances in polymerase engineering to enable enzymatic synthesis, replication and evolution of these unnatural polymers. The new sequence and chemical space that can be explored by combined advances in chemistry and biology are likely to be a fruitful source of novel nucleic acid therapeutics, aptamers and enzymes with useful applications in medicine, biotechnology, nanotechnology and material science.

Here we have taken another step towards these goals by identifying new mutations that enhance the ability to generate TNA polymerase with increased processivity and fidelity. The ability to code and decode sequence-defined genetic polymers, like TNA, provides access to many non-biological applications that will benefit materials science, nanotechnology, and molecular medicine. For example, TNA is highly resistant to nuclease degradation, making it a stable scaffold for future diagnostic and therapeutic applications. In addition, because TNA has the ability to evolve in response to imposed selection constraints, it could also be used to enhance our understanding of why nature chose RNA as the molecular basis of life's genetic material (79, 163). Previous work in this area has been limited by the absence of polymerases that could be used to study alternative chemistries of life. As this paradigm is now changing, we may soon discover that many different types of genetic polymers exhibit the characteristic signatures of heredity and evolution two important hallmarks of life (113).

Experimental

Materials

DNA oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA), and purified by denaturing polyacrylamide gel electrophoresis followed by electroelution and ethanol precipitation. Oligonucleotide concentrations were determined by UV absorbance using a NanoDrop spectrophotometer. NTPs and dNTPs were purchased from Sigma (St. Louis, MO). TNA triphosphates (tNTPs) were obtained by chemical synthesis as previously described (ref). Hen egg lysozyme was purchased from Sigma. Fluorinated oil HFE-7500 was purchased from 3M Novec, USA, and the fluorosurfactant and microfluidic chips were purchased from Dolomite, UK.

Polymerase Library Generation

The library of 9°N polymerase variants was generated by replacing the region coding for the finger thumb and palm domains of the protein with a DNA cassette containing mutations of interest. The triple saturation mutagenesis cassette, where amino acid positions 409, 485, and 664 were randomized, was created using three gBlock fragments purchased from IDT. These double stranded DNA fragments were chemically synthesized with all three positions of the desired codons having a random distribution of A, C, G, and T. The fragments were designed to have overlapping constant regions and defined restriction sites for cloning into a protein expression vector. The fragments were first amplified by PCR using three sets of unique primers (P1.For, P1.Rev, P2.For, P2.Rev, P3.For, P3.Rev) and the high fidelity AccuPrime polymerase (Life Technologies). An optimized number of PCR cycles were determined by qPCR analysis to minimize excessive amplification. 15 ng of each fragment was then pooled into a 100 µL PCR using the outermost forward (P1.For) and reverse primers (P3.Rev) to generate the full length DNA cassette. Following restriction digestion, the library was ligated into a protein expression plasmid backbone containing the remainder of the 9°N polymerase coding region.

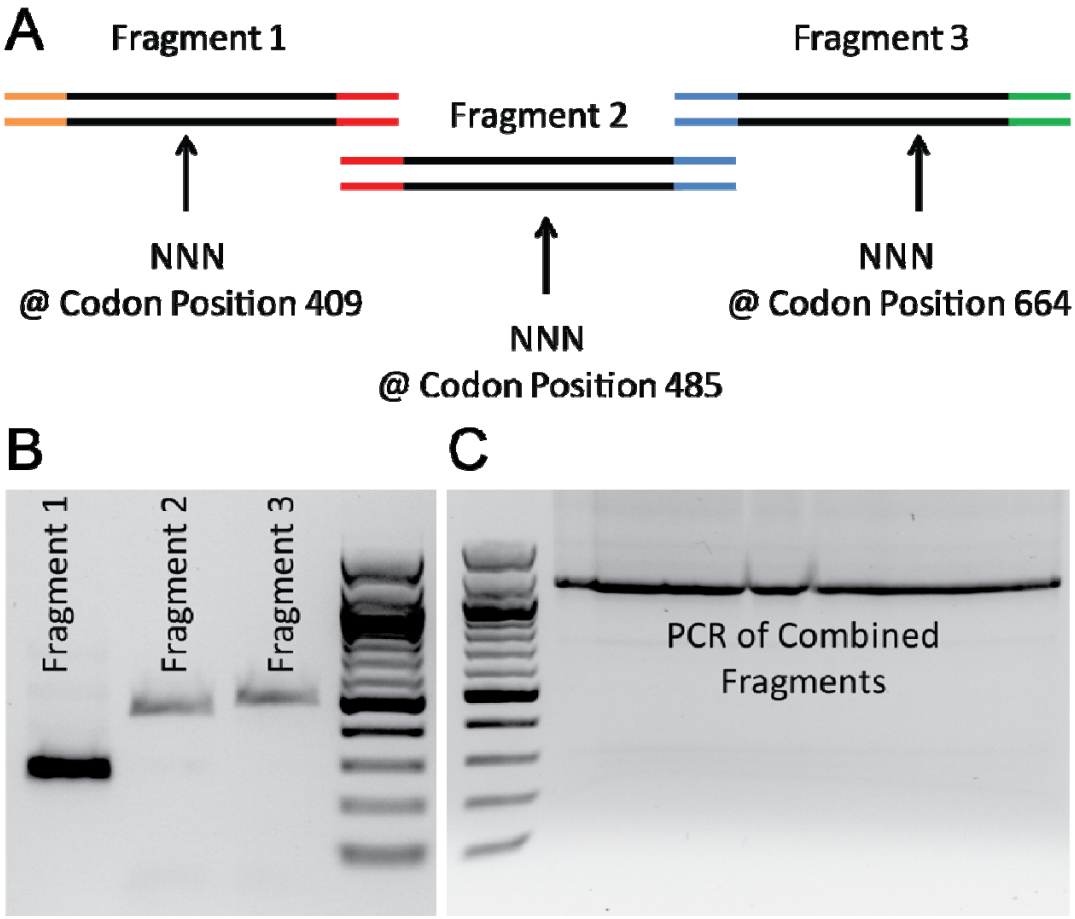


Figure 4.13 Polymerase library creation. A) Three gBlock dsDNA fragments were purchased fron IDT with fully degenerate codons at positions 409, 485 and 664 of the 9°N polymerase. Fragment two contains a 5′ region that is conserved with the 3′ end of fragment one and a 3′ region that is conserved with the 5′ end of fragment three. B) Each of the three fragments was individually amplified to generate sufficient quantities. C) The three fragments were pools and combined into a single PCR reaction using the forward primer for fragment one and the reverse primer for fragment three. The full length fragment was cloned into a protein expression vector containing the 9°N polymerase by restriction digestiong, ligation and transformation.

Microfluidic Droplet Generation

All microfluidic devices for monodisperse emulsion formation were purchased from Dolomite, UK and designs are available from their website. Syringe pumps and 1/16” OD fluorinated ethylene propylene (FEP) tubing with 0.01” ID (Idex 1478-20) was used to transport fluids through to microfluidic chips, and from the chip outlet to collection vessels. All fluid connections off chip were formed using 1/16’’ Upchurch

fitting connectors. The formation of water-in-oil single emulsions was performed using a quartz glass microfluidic device with a single inlet flow focusing junction geometry of 14 x 17 μM with a hydrophobic/fluorophilic coating (Cat. C000525G, Dolomite, UK). The device was connected by FEP tubing through a top interface linear connector (Cat. 3000109, Dolomite, UK) to syringes (100 μL, 500 μL SGE glass syringes, 2500 μL Hamilton Gastight syringe or 3 mL plastic syringe (Becton-Dickinson, Madrid, Spain)), which were driven by either an NE1002x syringe infusion pumps (New Era Pump Systems Inc., USA) or a pump manifold of neMESYS low pressure syringe pumps (Cetoni Gmbh, Germany) with accompanying control software. Carrier fluid was filtered using a 0.2 μm inline syringe filter, while the aqueous phase was filtered using an inline 10 μm frit filter. Droplet generation was monitored using a Nikon eclipse TS100 microscope with 20x ELWD Nikon objective and captured using a QIclick 12 bit monochrome CCD camera (QImaging, BC Canada). Flow rates were adjusted based on visual inspection with an average rate of 5 μL/min for the aqueous phase and 12 μL/min for the carrier oil. These flow rates yielded droplets with an average diameter of 14 ± X μm (~1 pL volume). A low viscosity fluorinated oil (HFE-7500, 3M USA) containing 1% (w/w) picosurf surfactant (Dolomite, UK) was used as the carrier fluid.

The formation of water-in-oil-in-water double emulsions was performed using a quartz glass microfluidic device with a single inlet flow focusing junction geometry of 14 x 17 μM (Cat. 3200136, Dolomite, UK). The water-in-oil emulsion and aqueous carrier phase were delivered to the device using syringes connected in the same fashion as described above for single emulsion formation. The water-in-oil emulsion was slowly drawn into a 250 μl SGE glass syringe, mounted into an infusion pump in a vertical position and left to settle for at least 30 minutes prior to delivery. Carrier fluid (25 mM NaCl, 1% Tween-80) was filtered using a 0.2 μm inline syringe filter,

while the water-in-oil emulsion was filtered using an inline 10 μm frit filter. Flow rates were adjusted based on visual inspection with an average rate of 1 μL/min for the single emulsion and 8 μL/min for the carrier aqueous phase.

Polymerase expression

Individual polymerase variants were tested by growing a clonal population of XL-1 blue *E. coli* carrying a plasmid encoding the polymerase of interest in Luria Broth (LB) supplemented with ampicillin (100 μg mL^{-1}). Cultures were grown at 37°C with shaking at 240 rpm and protein expression was induced by adding IPTG to a final concentration of 1 mM when the culture reached an OD600 of 0.6. Induced cultures were grown for 3 hours at 37°C with shaking to express protein. Following protein expression the cells were pelleted, the media was removed, and the cells were suspended in lysozyme buffer [50 mM TrisHCl pH 8.0, 300 mM NaCl, 0.1% (v/v) Triton X-100, 0.1 mg/ml lysozyme] with hen egg lysozyme. Cellular debris was removed by centrifugation for 15 min at 13,000 rpm and the supernatant was used directly as a source of polymerase for extension activity assays. Protein expression was confirmed by SDS-PAGE analysis and coomassie blue staining.

Cell compartmentalization in droplets

Cell populations were grown and polymerase variants were expressed as described above. After expression, a 2 mL aliquot of culture was centrifuged for 5 min (2,000 rcf) and the supernatant discarded. The cells were washed three times with 1x ThermoPol buffer [20 mM Tris-HCl, 10 mM $(NH4)_2SO4$, 10 mM KCl, 2 mM MgSO4, 0.1% Triton X-100, pH 8.8] (New England Biolabs Inc., Massachusetts, USA). After each was cells were centrifuged for 5 min (2,000 rcf) and the supernatant discarded. The bacterial pellet was taken up in 500 μL 1x ThermoPol buffer and the A600nm was determined. Cells were diluted to enable encapsulation at occupancies of 0.1 cells per droplet, according to the assumption that 1 mL *E. coli*

suspension at A600 of 1.0 contain 5 x 10^8 cells. Just prior to emulsification the cells were mixed with the PAA (see section below). The w/o emulsion was collected under a layer of mineral oil in an Eppendorf tube. Subsequently the water-in-oil emulsion was transformed into double emulsion as described under device operation.

Microscopy

Images were collected using a brightfield microscope (Eclipse TE300, Nikon) equipped with a a Hamamatsu Orca 3CCD camera using a 60×, 1.32 NA, oil-immersion objective lens and Immersion Oil Type DF (Cargille Laboratories) imaging medium. QED InVivo 3.2 (Media Cybernetics) was used to collect images, which were processed with Photoshop CS4 (Adobe) or ImageJ (NIH) software.

Microfluidic droplet generation was monitored using a Nikon eclipse TS100 inverted microscope with either a 10x, 0.3 NA Plan fluor, or 20x, 0.45 NA ELWD S Plan Fluor, Nikon objectives and captured using a QIclick 12 bit monochrome CCD camera (QImaging, BC Canada).

Flow cytometric analysis of double emulsion droplets

Water-in-oil–inwater double emulsion droplets were diluted into 150 mM NaCl and subjected to flow cytometric analysis (FACSCalibur, BD Biosciences). The sample was excited with a 488 nm argon laser and the emission was detected using a 530 ± 15 nm band-pass filter. Double emulsion populations were gated on logFSC/logSSC. Fluorescent readout was obtained from more than 15,000 droplets for each measurement. Cytometr software (Cell Quest, BD Biosciences) was used for data analysis.

Fluorescence-activated droplet sorting

Prior to sorting droplets using a fluorescence-activated cell sorter (FACS), the aqueous carrier phase (1% w/w Tween 80 in 25 mM NaCl) was exchanged for a solution of 25 mM NaCl to reduce the presence of surfactant in the aqueous phase.

Samples were sorted in a BD FACSAria (BD Biosciences) using PBS as sheath fluid. A set-up with a 70 μm nozzle was chosen to give an average sort rate of 5,000 – 8,000 events per second. The threshold trigger was set on side scatter. The sample was excited with a 488 nm argon laser and the emission was detected using a 530 ± 15 nm band-pass filter. The double emulsion population was gated from other populations in the sample on logFSC/logSSC

DNA recovery and transformation

Sorted samples were de-emulsified by extraction with ~2 volumes of Picobreak (Dolomite, UK) which contains 1H,1H,2H,2H-perfluorooctanol (PFO). After addition of Picobreak, the samples were vortexed followed by centrifugation (15 seconds, 2,000 rcf) to attain phase separation. The top, aqueous layer containing the plasmid DNA of interest was recovered. The bottom layer was re-extracted with 1 volume of molecular grade water to improve recovery yields. The plasmid DNA was concentrated from the combined aqueous layers using a spin column (DNA Clean & Concentrator™-5, Zymo Research), eluting with molecular grade water (10 μL). The DNA Clean & Concentrator™-5 also facilitates removal of protein from the sample. Electrocompetent *E. coli* cells (50 μL, β-10 *E. coli* cells NEB, USA) were transformed with 5 μL of purified DNA by applying one electric pulse of 1.80 kV (using an *E. coli* Pulser Cuvette, 0.1 cm electrode; Bio-Rad MicroPulser). Sterile S.O.C Medium (500 μL, Invitrogen) was added immediately after pulsing and the sample was grown at 37 °C with shaking at 240 rpm for 30 minutes before plating on LB agar containing ampicillin (100 μg mL^{-1}) followed by incubation at 37 °C overnight. Plasmid recovery efficiency was determined by comparison of the number of sorted droplets to the number of colonies obtained after transformation and plating. When large numbers of colonies were obtained, dilution plating was used to estimate the number of successful transformants.

Polymerase activity assay

Polymerase assays for selection were performed using an unlabeled DNA primer, a template with a fluorophore label at the 5' end and a quencher probe labeled with a quencher dye at the 3' end. The primer-template complex was annealed in 1x ThermoPol buffer by heating for 5 min at 95°C and cooling for 5 min at 4°C. The concentration of primer and template and quencher strands were 2, 1 and 3 μM respectively. Nucleotide triphosphates (100 μM final) were added to the reaction after primer annealing. Bacterial cells expressing polmyerase variants were added just prior to emulsification. Following emulsification, the reactions were incubated at 90°C for 5 minutes to lyse cells, followed by 55°C for the indicated amount of time.

Polymerase assays monitored by gel electrophoresis were carried out with a DNA primer labeled at the 5’ end with an IR800 dye. The primer-template complex was annealed in 1x ThermoPol buffer [20 mM Tris-HCl, 10 mM $(NH4)_2SO4$, 10 mM KCl, 2 mM MgSO4, 0.1% Triton X-100, pH 8.8] by heating for 5 min at 95°C and cooling for 5 min at 4°C. The concentration of primer and template strands were 1 and 2 μM respectively. The polymerase and nucleotide triphosphates (100 μM final) were added to the reaction after primer annealing. Following addition of all components, the reactions were incubated at 55°C for the indicated amount of time. Upon completion, reactions were quenched by addition of 10-fold stop buffer [1x Tris-boric acid buffer, 20 mM EDTA, 7 M urea, pH 8]. Samples were denatured by incubating at 90°C for 5 minutes prior to separation by denaturing PAGE and visualization of the IR800 dye using a LICOR Oddysey CLx imager.

Table 4.3. DNA primers and templates sequences.

Name	DNA Sequence (5' -> 3')
Lib.409.NNN	GAACGTGAACTGGCGCGCCGTCGTGGCGGTTATGCGGGCGGTTATGTGA AAGAACCGGAACGTGGCCTGTGGGATAACATTGTGTATCTGGATTTTCGT AGCCTG**NNN**CCGAGCATTATTATCACCCACAATGTGAGCCCGGATACCC TGAACCGTGAAGGCTGCAAAGAATATGATGTGGCGCCGGAAGTGGGCCA TAAATTCTGCAAAGATTTCCCGGGCTTTATT
Lib.485.NNN	AAGATTTCCCGGGCTTTATTCCGAGCCTGCTGGGCGATCTGCTCGAGGAA CGCCAGAAAATCAAACGCAAAATGAAAGCGACCGTTGATCCGCTGGAAA AAAAACTGCTGGATTATCGTCAGCGC**NNN**ATTAAAATTCTGGCCAACAGC TTCTATGGCTATTATGGTTATGCGAAAGCGCGTTGGTATTGCAAAGAATG CGCGGAAAGCGTGACCGCGTGGGGCCGTGAATATATCGAAATGGTGATC CGCGAGCTCGAAGAAAAATTCGGCTTCAAAGTGCTGTATGCGGATACCG ATGGCCTGCATGCGACCATTCCGGGTGCGGATGCGGAAACCGTGAAAAA AAAAGCGAAAGAATTCCTGAAATACATCAATCCGAAACTGCCGGGCCTGC TGGAACTGGAATATGAAGGCTTTTATGTGCGTGGCTTTTTCGTGACCAAA AAAAAATACGCGGTGATCGATGAAGAAGGCAAAATTACCACCCGTGGCCT GGAA
Lib.664.NNN	AATACGCGGTGATCGATGAAGAAGGCAAAATTACCACCCGTGGCCTGGA AATTGTGCGTCGTGATTGGAGCGAAATTGCGAAAGAAACCCAGGCGCGT GTGCTGGAAGCGATTCTGAAACATGGCGATGTGGAAGAAGCGGTGCGTA TTGTTAAAGAAGTGACCGAAAAACTGAGCAAATATGAGGTACCGCCGGAA AAACTGGTGATTCAT**NNN**CAAATTACCCGTGATCTGCGTGATTATAAAGC GACCGGTCCGCATGTGGCGGTGGCAAAACGTCTGGCAGCGCGTGGCGT GAAAATTCGTCCGGGCACCGTGATTAGCTATATTGTGCTGAAAGGCAGCG GCCGCATTGGCGATCGTGCGATTCCGGCGGATGAATTTGATCCGACCAA ACATCGTTATGATGCGGAATATTATATCGAAAACCAGGTGCTGCCGGCGG TGGAACGTATTCTGAAAGCGTTTGGCTATCGTAAAGAAGATCTGCGCTAT C
P1.For	AACTGGCGCGCCGTCGTGGCGGTTATGCGG
P1.Rev	CGTTCCTCGAGCAGATCGCCCAGCAGGCTCGGAATAAAG
P2.For	ATCTGCTCGAGGAACGCCAGAAAATCAAACGC
P2.Rev	TTCCAGGCCACGGGTGGTAATTTTGC
P3.For	AATACGCGGTGATCGATGAAG
P3.Rev	GATAGCGCAGATCTTCTTTACGATAGCC
PBS2	GACACTCGTATGCAGTAGCC
ST.1G.Cy5	/5Cy3/ACAACCATACTCTCCTCATCACTATTCAACTTACAATCGATACAAC CTTATAATCCACATGGCTACTGCATACGAGTGTC
ST.1G.FAM	/56FAM/ACAACCATACTCTCCTCATCACTATTCAACTTACAATCGATACAA CCTTATAATCCACATGGCTACTGCATACGAGTGTC
QP13.Iowa	ACAACCATACTCT/3IABkFQ/
QP16.Iowa	ACAACCATACTCTCCT/3IABkFQ/
QP20.Iowa	ACAACCATACTCTCCTCATC/3IABkFQ/
QP13.BHQ	ACAACCATACTCTCCT/3BHQ_1/
QP16.BHQ	ACAACCATACTCTCCTCATC/3BHQ_1/
QP20.BHQ	ACAACCATACTCTCCTCATC/3BHQ_1/
QP20.BHQ	ACAACCATACTCTCCTCATC/3BHQ_1/

References

66. Endy D (2005) Foundations for engineering biology. *Nature* 438:449–453.

67. Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet* 6(7):533–543.

68. Yang Z, Chen F, Alvarado JB, Benner SA (2011) Amplification, mutation, and sequencing of a six-letter synthetic genetic system. *J Am Chem Soc* 133(38):15105–15112.

69. Malyshev DA, et al. (2012) Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc Natl Acad Sci U S A* 109(30):12005–12010.

70. Moran S, Ren RXF, Rumney S, Kool ET (1997) Difluorotoluene, a nonpolar isostere for thymine, codes specifically and efficiently for adenine in DNA replication. *J Am Chem Soc* 119(8):2056–2057.

71. Switzer C, Moroney SE, Benner SA (1989) Enzymatic incorporation of a new base pair into DNA and RNA. *J Am Chem Soc* 111(21):8322–8323.

72. Piccirilli JA, Krauch T, Moroney SE, Benner SA (1990) Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* 343(6253):33–37.

73. Malyshev D a, et al. (2014) A semi-synthetic organism with an expanded genetic alphabet. *Nature* 509(7500):385–388.

74. Marlière P, et al. (2011) Chemical evolution of a bacterium's genome. *Angew Chemie - Int Ed* 50(31):7109–7114.

75. El-Sagheer AH, Sanzone AP, Gao R, Tavassoli A, Brown T (2011) Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in Escherichia coli. *Proc Natl Acad Sci U S A* 108(28):11338–11343.

76. Krueger AT, Peterson LW, Chelliserry J, Kleinbaum DJ, Kool ET (2011) Encoding phenotype in bacteria with an alternative genetic set. *J Am Chem Soc* 133(45):18447–18451.

77. Pochet S, Kaminski PA, Van Aerschot A, Herdewijn P, Marlière P (2003) Réplication in vivo et ex vivo de l'analogue hexitol des acides nucléiques. *Comptes Rendus - Biol* 326(12):1175–1184.

78. Orgel L (2000) A Simpler Nucleic Acid. *Sciecnce* 290(5495):1306–1307.

79. Eschenmoser A (1999) Chemical etiology of nucleic acid structure. *Science (80-)* 284(5423):2118–2124.

80. Wilds CJ, Wawrzak Z, Krishnamurthy R, Eschenmoser A, Egli M (2002) Crystal Structure of a B-Form DNA Duplex Containing (L)-r-Threofuranosyl (3'-2')

Nucleosides: A Four-Carbon Sugar Is Easily Accommodated into the Backbone of DNA. *Science (80-)* 124(46):13716–13721.

81. Yang YW, Zhang S, McCullum EO, Chaput JC (2007) Experimental evidence that GNA and TNA were not sequential polymers in the prebiotic evolution of RNA. *J Mol Evol* 65(3):289–295.

82. Steitz TA (1999) DNA Polymerases: Structural Diversity and Common Mechanisms. *J Biol Chem* 274(25):17395–17398.

83. Braithwaite DK, Ito J (1993) Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res* 21(4):787–802.

84. Ito J, Braithwaite DK (1991) Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res* 19(15):4045–4057.

85. Cann IKO, Ishino Y (1999) Archaeal DNA replication: Identifying the pieces to solve a puzzle. *Genetics* 152(4):1249–1267.

86. Moon AF, et al. (2007) The X family portrait: Structural insights into biological functions of X family polymerases. *DNA Repair (Amst)* 6(12):1709–1725.

87. Sale j. E, Lehmann AR, Woodgate R (2012) Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat Rev Mol Cell Biol* 13(3):141–152.

88. Kornberg A (1960) Biological Synthesis of Deoxyribonucleic Acid. *Science (80-)* 131(3412):1503–1508.

89. Ollis DL, Brick P, Hamlin R, Xuong NG, Steitz TA (1985) Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature* 313(6005):762–766.

90. Mönttinen H a M, Ravantti JJ, Stuart DI, Poranen MM (2014) Automated structural comparisons clarify the phylogeny of the right-hand-shape polymerases. *Mol Biol Evol* 31(10):2741–2752.

91. Beard WA, Wilson SH (2000) Structural design of a eukaryotic DNA repair polymerase: DNA polymerase beta. *Mutat Res* 460(3-4):231–244.

92. Werner F, Grohmann D (2011) Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* 9(2):85–98.

93. Klenow H, Henningsen I (1970) Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. *Proc Natl Acad Sci U S A* 65(2):168–175.

94. Lawyer FC, et al. (1993) High level expression, purification and enzymatic characterization of full length Thermus aquaticus DNA polymerase and truncated a foom deficient in 5’ to 3' exonuclease activity. *J Biol Chem* 2(4):275–287.

95. Lundberg KS, et al. (1991) High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene* 108(1):1–6.

96. Kong H, Kucera RB, Jack WE (1993) Characterization of a DNA polymerase from the hyperthermophile archaea Thermococcus litoralis. *J Biol Chem* 268(3):1965–1975.

97. Gardner AF, Jack WE (2002) Acyclic and dideoxy terminator preferences denote divergent sugar recognition by archaeon and Taq DNA polymerases. *Nucleic Acids Res* 30(2):605–613.

98. Gardner AF, Jack WE (1999) Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res* 27(12):2545–2553.

99. Ichida JK, Horhota A, Zou K, McLaughlin LW, Szostak JW (2005) High fidelity TNA synthesis by Therminator polymerase. *Nucleic Acids Res* 33(16):5219–5225.

100. Chaput JC, Szostak JW (2003) TNA synthesis by DNA polymerases. *J Am Chem Soc* 125(31):9274–9275.

101. Ebert MO, Mang C, Krishnamurthy R, Eschenmoser A, Jaun B (2008) The structure of a TNA-TNA complex in solution: NMR study of the octamer duplex derived from α-(L)-threofuranosyl-(3’-2')-CGAATTCG. *J Am Chem Soc* 130(45):15105–15115.

102. Schöning K, et al. (2000) Chemical etiology of nucleic acid structure: the alpha-threofuranosyl-(3’-->2') oligonucleotide system. *Science (80-)* 290(5495):1347–1351.

103. Dunn MR, et al. (2015) DNA Polymerase-Mediated Synthesis of Unbiased Threose Nucleic Acid (TNA) Polymers Requires 7-Deazaguanine To Suppress G:G Mispairing during TNA Transcription. *J Am Chem Soc* 137(12):4014–4017.

104. Yu H, Zhang S, Dunn MR, Chaput JC (2013) An efficient and faithful in vitro replication system for threose nucleic acid. *J Am Chem Soc* 135(9):3583–3591.

105. Ichida JK, et al. (2005) An in vitro selection system for TNA. *J Am Chem Soc* 127(9):2802–2803.

106. Horhota A, et al. (2005) Kinetic analysis of an efficient DNA-dependent TNA polymerase. *J Am Chem Soc* 127(20):7427–7434.

107. Zhang S, Chaput JC (2013) Synthesis and enzymatic incorporation of a-l-threofuranosyl adenine triphosphate (tATP). *Bioorganic Med Chem Lett* 23(5):1447–1449.

108. Chaput JC, Ichida JK, Szostak JW (2003) DNA polymerase-mediated DNA synthesis on a TNA template. *J Am Chem Soc* 125(4):856–857.

109. Cadwell RC, Joyce GF (1992) Randomization of genes by PCR mutagenesis. *Genome Res* 2(1):28–33.

110. Leung DW, Chen E, Goeddel D V (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique* 1:11–15.

111. Skelly J V, Edwards KJ, Jenkins TC, Neidle S (1993) Crystal structure of an oligonucleotide duplex containing G.G base pairs: influence of mispairing on DNA backbone conformation. *Proc Natl Acad Sci U S A* 90(3):804–808.

112. Staiger N, Marx A (2010) A DNA polymerase with increased reactivity for ribonucleotides and C5-modified deoxyribonucleotides. *ChemBioChem* 11(14):1963–1966.

113. Pinheiro VB, et al. (2012) Synthetic Genetic Polymers Capable of Heredity and Evolution. *Science (80-)* 336(6079):341–344.

114. Chen T, Romesberg FE (2014) Directed polymerase evolution. *FEBS Lett* 588(2):219–229.

115. Sauter KBM, Marx A (2006) Evolving thermostable reverse transcriptase activity in a DNA polymerase scaffold. *Angew Chemie - Int Ed* 45(45):7633–7635.

116. Lutz S, Burgstaller P, Benner SA (1999) An in vitro screening technique for DNA polymerases that can incorporate modified nucleotides. Pseudothymidine as a substrate for thermostable polymerases. *Nucleic Acids Res* 27(13):2792–2798.

117. Aposhian H V., Kornberg A (1962) Enzymatic Synthesis of Deoxyribonucleic Acid. *J Biol Chem* 237(2):519–525.

118. Summerer D, Marx A (2002) A molecular beacon for quantitative monitoring of the DNA polymerase reaction in real-time. *Angew Chemie - Int Ed* 41(19):3620–3622.

119. Dorjsuren D, et al. (2009) A real-time fluorescence method for enzymatic characterization of specialized human DNA polymerases. *Nucleic Acids Res* 37(19):1–12.

120. Griep MA (1995) Fluorescence recovery assay: a continuous assay for processive DNA polymerases applied specifically to DNA polymerase III holoenzyme. *Anal Biochem* 232(2):180–189.

121. Yu L, Hu G, Howells L (2002) Fluorescence-based, high-throughput DNA polymerase assay. *Biotechniques* 33(4):938–941.

122. Tveit H, Kristensen T (2001) Fluorescence-based DNA polymerase assay. *Anal Biochem* 289(1):96–98.

123. Reetz MT, Carballeira JD (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc* 2(4):891–903.

124. Li Y, Mitaxov V, Waksman G (1999) Structure-based design of Taq DNA polymerases with improved properties of dideoxynucleotide incorporation. *Proc Natl Acad Sci U S A* 96(17):9491–9496.

125. Summerer D, Rudinger NZ, Detmer I, Marx A (2005) Enhanced fidelity in mismatch extension by DNA polymerase through directed combinatorial enzyme design. *Angew Chemie - Int Ed* 44(30):4712–4715.

126. Gloeckner C, Sauter KBM, Marx A (2007) Evolving a thermostable DNA polymerase that amplifies from highly damaged templates. *Angew Chemie - Int Ed* 46(17):3115–3117.

127. Siegmund V, Santner T, Micura R, Marx A (2012) Screening mutant libraries of T7 RNA polymerase for candidates with increased acceptance of 2'-modified nucleotides. *Chem Commun* 48(79):9870–9872.

128. Sweasy JB, Loeb LA (1993) Detection and characterization of mammalian DNA polymerase beta mutants by functional complementation in Escherichia coli. *Proc Natl Acad Sci U S A* 90(10):4626–4630.

129. Hathaway TR (1996) Human Immunodeficiency Virus Reverse Transcriptase. *J Biol Chem* 271(9):4872–4878.

130. Camps M, Naukkarinen J, Johnson BP, Loeb LA (2003) Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proc Natl Acad Sci U S A* 100(17):9727–9732.

131. Patel PH, Kawate H, Adman E, Ashbach M, Loeb LA (2001) A Single Highly Mutable Catalytic Site Amino Acid Is Critical for DNA Polymerase Fidelity. *J Biol Chem* 276(7):5044–5051.

132. Patel PH, Loeb LA (2000) DNA polymerase active site is highly mutable: evolutionary consequences. *Proc Natl Acad Sci U S A* 97(10):5095–5100.

133. Suzuki M, Baskin D, Hood L, Loeb LA (1996) Random mutagenesis of Thermus aquaticus DNA polymerase I: concordance of immutable sites in vivo with the crystal structure. *Proc Natl Acad Sci U S A* 93(18):9670–9675.

134. Glick E, Vigna KL, Loeb LA (2001) Mutations in human DNA polymerase eta motif II alter bypass of DNA lesions. *EMBO J* 20(24):7303–7312.

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

136. Smith GP, Petrenko VA (1997) Phage Display. *Chem Rev* 97(96):391–410.

137. Brunet E, Chauvin C, Choumet V, Jestin J-L (2002) A novel strategy for the functional cloning of enzymes using filamentous phage display: the case of nucleotidyl transferases. *Nucleic Acids Res* 30(9):e40.

138. Jestin J, Kristensen P, Winter G (1999) A Method for the Selection of Catalytic Activity Using Phage Display and Proximity. *Angew Chemie - Int Ed* 38(8):1124–1127.

139. Xia G, et al. (2002) Directed evolution of novel polymerase activities: mutation of a DNA polymerase into an efficient RNA polymerase. *Proc Natl Acad Sci U S A* 99(10):6597–6602.

140. Fa M, Radeghieri A, Henry AA, Romesberg FE (2004) Expanding the substrate repertoire of a DNA polymerase by directed evolution. *J Am Chem Soc* 126(6):1748–1754.

141. Vichier-Guerre S, Ferris S, Auberger N, Mahiddine K, Jestin JL (2006) A population of thermostable reverse transcriptases evolved from Thermus aquaticus DNA polymerase I by phage display. *Angew Chemie - Int Ed* 45(37):6133–6137.

142. Leconte AM, Chen L, Romesberg FE (2005) Polymerase evolution: Efforts toward expansion of the genetic code. *J Am Chem Soc* 127(36):12470–12471.

143. Yang G, Withers SG (2009) Ultrahigh-throughput FACS-based screening for directed enzyme evolution. *ChemBioChem* 10(17):2704–2715.

144. Tawfik DS, Griffiths AD (1998) Man-made cell-like compartments for molecular evolution. *Nat Biotechnol* 16(7):652–656.

145. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A* 98(8):4552–4557.

146. Anna SL, Bontoux N, Stone HA (2003) Formation of dispersions using "flow focusing" in microchannels. *Appl Phys Lett* 82(3):364–366.

147. Umbanhowar PB, Prasad V, Weitz D a. (2000) Monodisperse emulsion generation via drop break off in a coflowing stream. *Langmuir* 16(2):347–351.

148. Nisisako T, Okushima S, Torii T (2005) Controlled formulation of monodisperse double emulsions in a multiple-phase microfluidic system. *Soft Matter* 1(1):23.

149. Miller OJ, et al. (2006) Directed evolution by in vitro compartmentalization. *Nat Methods* 3(7):561–570.

150. Ghadessy FJ, et al. (2004) Generic expansion of the substrate spectrum of a DNA polymerase by directed evolution. *Nat Biotechnol* 22(6):755–759.

151. Ong JL, Loakes D, Jaroslawski S, Too K, Holliger P (2006) Directed Evolution of DNA Polymerase, RNA Polymerase and Reverse Transcriptase Activity in a Single Polypeptide. *J Mol Biol* 361(3):537–550.

152. Ramsay N, et al. (2010) CyDNA: Synthesis and replication of highly Cy-Dye substituted DNA by an evolved polymerase. *J Am Chem Soc* 132(14):5096–5104.

153. Cozens C, Pinheiro VB, Vaisman A, Woodgate R, Holliger P (2012) A short adaptive path from DNA to RNA polymerases. *Proc Natl Acad Sci U S A* 109(21):8067–8072.

154. Griffiths AD, Tawfik DS (2006) Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol* 24(9):395–402.

155. Holtze C, et al. (2008) Biocompatible surfactants for water-in-fluorocarbon emulsions. *Lab Chip* 8(10):1632–1639.

156. Garstecki P, Fuerstman MJ, Stone HA, Whitesides GM (2006) Formation of droplets and bubbles in a microfluidic T-junction-scaling and mechanism of break-up. *Lab Chip* 6(3):437–446.

157. Teh S-Y, Lin R, Hung L-H, Lee AP (2008) Droplet microfluidics. *Lab Chip* 8(2):198–220.

158. Riche CT, Zhang C, Gupta M, Malmstadt N (2014) Fluoropolymer surface coatings to control droplets in microfluidic devices. *Lab Chip* 14(11):1834–41.

159. Baret J-C, et al. (2009) Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab Chip* 9(13):1850–1858.

160. Bernath K, et al. (2004) In vitro compartmentalization by double emulsions: Sorting and gene enrichment by fluorescence activated cell sorting. *Anal Biochem* 325(1):151–157.

161. Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5’----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A* 88(16):7276–7280.

162. Zinchenko A, et al. (2014) One in a million: Flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Anal Chem* 86(5):2526–2533.

163. Joyce GF (2002) The antiquity of RNA-based evolution. *Nature* 418(6894):214–221.

REFERENCES

1. Baines IC, Colas P (2006) Peptide aptamers as guides for small-molecule drug discovery. *Drug Discov Today* 11(7-8):334–341.

2. Takahashi TT, Austin RJ, Roberts RW (2003) mRNA display: Ligand discovery, interaction analysis and beyond. *Trends Biochem Sci* 28(3):159–165.

3. Molek P, Strukelj B, Bratkovic T (2011) Peptide phage display as a tool for drug discovery: Targeting membrane receptors. *Molecules* 16(1):857–887.

4. Olson CA, et al. (2012) Single-round, multiplexed antibody mimetic design through mRNA display. *Angew Chemie - Int Ed* 51(50):12449–12453.

5. Matochko WL, et al. (2012) Deep sequencing analysis of phage libraries using Illumina platform. *Methods* 58(1):47–55.

6. Turunen L, Takkinen K, Söderlund H, Pulli T (2009) Automated panning and screening procedure on microplates for antibody generation from phage display libraries. *J Biomol Screen* 14(3):282–293.

7. Cung K, et al. (2012) Rapid, multiplexed microfluidic phage display. *Lab Chip* 12(3):562–565.

8. Baggio R, et al. (2002) Identification of epitope-like consensus motifs using mRNA display. *J Mol Recognit* 15(3):126–134.

9. Colwill K, Gräslund S (2011) A roadmap to generate renewable protein binders to the human proteome. *Nat Methods* 8(7):551–558.

10. Katzen F, Chang G, Kudlicki W (2005) The past, present and future of cell-free protein synthesis. *Trends Biotechnol* 23(3):150–156.

11. Wong I, Lohman TM (1993) A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions. *Proc Natl Acad Sci U S A* 90(12):5428–5432.

12. Wilson DS, Keefe AD, Szostak JW (2001) The use of mRNA display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci U S A* 98(7):3750–3755.

13. Kapust RB, Waugh DS (2000) Controlled intracellular processing of fusion proteins by TEV protease. *Protein Expr Purif* 19(2):312–318.

14. Keefe AD, Wilson DS, Seelig B, Szostak JW (2001) One-step purification of recombinant proteins using a nanomolar-affinity streptavidin-binding peptide, the SBP-Tag. *Protein Expr Purif* 23(3):440–446.

15. Yu H, Jiang B, Chaput JC (2011) Aptamers can discriminate alkaline proteins with high specificity. *ChemBioChem* 12(17):2659–2666.

16. Xu D, Shi H (2009) Composite RNA aptamers as functional mimics of proteins. *Nucleic Acids Res* 37(9):1–9.

17. Yu H, Zhang S, Chaput JC (2012) Darwinian evolution of an alternative genetic system provides support for TNA as an RNA progenitor. *Nat Chem* 4(3):183–187.

18. Raffler N a., Schneider-Mergener J, Famulok M (2003) A novel class of small functional peptides that bind and inhibit human alpha-thrombin isolated by mRNA display. *Chem Biol* 10(1):69–79.

19. Crawley JTB, Zanardelli S, Chion CKNK, Lane D a. (2007) The central role of thrombin in hemostasis. *J Thromb Haemost* 5:95–101.

20. Jerabek-Willemsen M, Wienken CJ, Braun D, Baaske P, Duhr S (2011) Molecular interaction studies using microscale thermophoresis. *Assay Drug Dev Technol* 9(4):342–353.

21. Sato AK, Viswanathan M, Kent RB, Wood CR (2006) Therapeutic peptides: technological advances driving peptides into development. *Curr Opin Biotechnol* 17(6):638–642.

22. Goodwin D, Simerska P, Toth I (2012) Peptides As Therapeutics with Enhanced Bioactivity. *Curr Med Chem* 19(26):4451–4461.

23. McGregor DP (2008) Discovering and improving novel peptide therapeutics. *Curr Opin Pharmacol* 8(5):616–619.

24. Williams BA, et al. (2009) Creating protein affinity reagents by combining peptide ligands on synthetic DNA scaffolds. *J Am Chem Soc* 131(15):17233–17241.

25. R Core Team R: A Language and Environment for Statistical Computing. Available at: http://www.r-project.org.

26. Jackson RJ, Hellen CUT, Pestova T V (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* 11(2):113–127.

27. Sonenberg N, Hinnebusch AG (2009) Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell* 136(4):731–745.

28. Shatsky IN, Dmitriev SE, Terenin IM, Andreev DE (2010) Cap- and IRES-independent scanning mechanism of translation initiation as an alternative to the concept of cellular IRESs. *Mol Cells* 30(4):285–293.

29. Spriggs KA, Stoneley M, Bushell M, Willis AE (2008) Re-programming of translation following cell stress allows IRES-mediated translation to predominate. *Biol Cell* 100(1):27–38.

30. Johannes G, Carter MS, Eisen MB, Brown PO, Sarnow P (1999) Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F

concentrations using a cDNA microarray. *Proc Natl Acad Sci U S A* 96(23):13118–13123.

31. Roberts RW, Szostak JW (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc Natl Acad Sci U S A* 94(23):12297–12302.

32. Salehi-Ashtiani K, Lupta A, Litovchick A, Szostak JW (2006) A Genomewide Search for Ribozymes. *Science (80-)* 313(5794):1788–1792.

33. Kasowski M, et al. (2010) Variation in transcription factor binding among humans. *Science (80-)* 328(5975):232–235.

34. Korbel J, Urban A, Affourtit J, Godwin B (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science (80-)* 13(5849):350–358.

35. Gilbert W V, Zhou K, Butler TK, Doudna JA (2007) Cap-Independent Translation Is Required for Starvation-Induced Differentiation in Yeast. *Science (80-)* 217(5842):1224–1227.

36. Baranick BT, et al. (2008) Splicing mediates the activity of four putative cellular internal ribosome entry sites. *Proc Natl Acad Sci U S A* 105(12):4733–4738.

37. Moss B (2013) Vaccinia Virus : Vaccine for Tool Development. *Science (80-)* 252(5013):1662–1667.

38. Van Eden ME, Byrd MP, Sherrill KW, Lloyd RE (2004) Demonstrating internal ribosome entry sites in eukaryotic mRNAs using stringent RNA test procedures. *RNA* 10(4):720–730.

39. Mitchell SF, et al. (2010) The 5′-7-methylguanosine cap on eukaryotic mRNAs serves both to stimulate canonical translation initiation and to block an alternative pathway. *Mol Cell* 39(6):950–962.

40. Mokrejš M, et al. (2009) IRESite A tool for the examination of viral and cellular internal ribosome entry sites. *Nucleic Acids Res* 38:131–136.

41. Sakharkar MK, Chow VTK, Kangueane P (2004) Distributions of exons and introns in the human genome. *In Silico Biol* 4(4):387–393.

42. Akey JM, et al. (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* 19(5):711–722.

43. Sabeti PC, et al. (2006) Positive natural selection in the human lineage. *Science (80-)* 312(5780):1614–1620.

44. Traynelis SF, et al. (2010) Glutamate Receptor Ion Channels: Structure, Regulation, and Function. *Pharmacol Rev* 62(3):405–496.

45. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.

46. Ji H, et al. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 26(11):1293–1300.

47. Kin T, Ono Y (2007) Idiographica: A general-purpose web application to build idiograms on-demand for human, mouse and rat. *Bioinformatics* 23(21):2945–2946.

48. Thomas PD, et al. (2003) PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res* 13(9):2129–2141.

49. Auerbach RK, et al. (2009) Mapping accessible chromatin regions using Sono-Seq. *Proc Natl Acad Sci U S A* 106(35):14926–14931.

50. Zhu J (2012) Mammalian cell protein expression for biopharmaceutical production. *Biotechnol Adv* 30(5):1158–1170.

51. Swiech K, Picanço-Castro V, Covas DT (2012) Human cells: New platform for recombinant therapeutic protein production. *Protein Expr Purif* 84(1):147–153.

52. Schmidt FR (2004) Recombinant expression systems in the pharmaceutical industry. *Appl Microbiol Biotechnol* 65(4):363–372.

53. Wurm FM (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol* 22(11):1393–1398.

54. Colosimo A, et al. (2000) Review Transfer and Expression of Foreign Genes in Mammalian Cells. *Biotechniques* 29(2):314–331.

55. Yin J, Li G, Ren X, Herrler G (2007) Select what you need: A comparative evaluation of the advantages and limitations of frequently used expression systems for foreign genes. *J Biotechnol* 127(3):335–347.

56. Demain AL, Vaishnav P (2009) Production of recombinant proteins by microbes and higher organisms. *Biotechnol Adv* 27(3):297–306.

57. Jacobs BL, et al. (2009) Vaccinia virus vaccines: Past, present and future. *Antiviral Res* 84(1):1–13.

58. Wellensiek BP, et al. (2013) Genome-wide profiling of human cap-independent translation-enhancing elements. *Nat Methods* 10(8):747–50.

59. Elfakess R, Dikstein R (2008) A translation initiation element specific to mRNAs with very short 5′UTR that also regulates transcription. *PLoS One* 3(8):e3094.

60. Schwer B, Visca P, Vos JC, Stunnenberg HG (1987) Discontinuous transcription or RNA processing of vaccinia virus late messengers results in a 5′ poly(A) leader. *Cell* 50(2):163–169.

61. Fuerst TR, Niles EG, Studier FW, Moss B (1986) Eukaryotic transient-expression system based on recombinant vaccinia virus that synthesizes bacteriophage T7 RNA polymerase. *Proc Natl Acad Sci U S A* 83(21):8122–8126.

62. Hebben M, et al. (2007) High level protein expression in mammalian cells using a safe viral vector: modified vaccinia virus Ankara. *Protein Expr Purif* 56(2):269–278.

63. Elroy-Stein O, Fuerst TR, Moss B (1989) Cap-independent translation of mRNA conferred by encephalomyocarditis virus 5' sequence improves the performance of the vaccinia virus/bacteriophage T7 hybrid expression system. *Proc Natl Acad Sci U S A* 86(16):6126–6130.

64. Drexler I, Heller K, Wahren B, Erfle V, Sutter G (1998) Highly attenuated modified vaccinia virus Ankara replicates in baby hamster kidney cells, a potential host for virus propagation, but not in various human transformed and primary cells. *J Gen Virol* 79:347–352.

65. Büssow K, et al. (2005) Structural genomics of human proteins - target selection and generation of a public catalogue of expression clones. *Microb Cell Fact* 4:21.

66. Endy D (2005) Foundations for engineering biology. *Nature* 438:449–453.

67. Benner SA, Sismour AM (2005) Synthetic biology. *Nat Rev Genet* 6(7):533–543.

68. Yang Z, Chen F, Alvarado JB, Benner SA (2011) Amplification, mutation, and sequencing of a six-letter synthetic genetic system. *J Am Chem Soc* 133(38):15105–15112.

69. Malyshev DA, et al. (2012) Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc Natl Acad Sci U S A* 109(30):12005–12010.

70. Moran S, Ren RXF, Rumney S, Kool ET (1997) Difluorotoluene, a nonpolar isostere for thymine, codes specifically and efficiently for adenine in DNA replication. *J Am Chem Soc* 119(8):2056–2057.

71. Switzer C, Moroney SE, Benner SA (1989) Enzymatic incorporation of a new base pair into DNA and RNA. *J Am Chem Soc* 111(21):8322–8323.

72. Piccirilli JA, Krauch T, Moroney SE, Benner SA (1990) Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* 343(6253):33–37.

73. Malyshev D a, et al. (2014) A semi-synthetic organism with an expanded genetic alphabet. *Nature* 509(7500):385–388.

74. Marlière P, et al. (2011) Chemical evolution of a bacterium's genome. *Angew Chemie - Int Ed* 50(31):7109–7114.

75. El-Sagheer AH, Sanzone AP, Gao R, Tavassoli A, Brown T (2011) Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in Escherichia coli. *Proc Natl Acad Sci U S A* 108(28):11338–11343.

76. Krueger AT, Peterson LW, Chelliserry J, Kleinbaum DJ, Kool ET (2011) Encoding phenotype in bacteria with an alternative genetic set. *J Am Chem Soc* 133(45):18447–18451.

77. Pochet S, Kaminski PA, Van Aerschot A, Herdewijn P, Marlière P (2003) Réplication in vivo et ex vivo de l'analogue hexitol des acides nucléiques. *Comptes Rendus - Biol* 326(12):1175–1184.

78. Orgel L (2000) A Simpler Nucleic Acid. *Sciecnce* 290(5495):1306–1307.

79. Eschenmoser A (1999) Chemical etiology of nucleic acid structure. *Science (80-)* 284(5423):2118–2124.

80. Wilds CJ, Wawrzak Z, Krishnamurthy R, Eschenmoser A, Egli M (2002) Crystal Structure of a B-Form DNA Duplex Containing (L)-r-Threofuranosyl (3'-2') Nucleosides: A Four-Carbon Sugar Is Easily Accommodated into the Backbone of DNA. *Science (80-)* 124(46):13716–13721.

81. Yang YW, Zhang S, McCullum EO, Chaput JC (2007) Experimental evidence that GNA and TNA were not sequential polymers in the prebiotic evolution of RNA. *J Mol Evol* 65(3):289–295.

82. Steitz TA (1999) DNA Polymerases: Structural Diversity and Common Mechanisms. *J Biol Chem* 274(25):17395–17398.

83. Braithwaite DK, Ito J (1993) Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res* 21(4):787–802.

84. Ito J, Braithwaite DK (1991) Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res* 19(15):4045–4057.

85. Cann IKO, Ishino Y (1999) Archaeal DNA replication: Identifying the pieces to solve a puzzle. *Genetics* 152(4):1249–1267.

86. Moon AF, et al. (2007) The X family portrait: Structural insights into biological functions of X family polymerases. *DNA Repair (Amst)* 6(12):1709–1725.

87. Sale j. E, Lehmann AR, Woodgate R (2012) Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat Rev Mol Cell Biol* 13(3):141–152.

88. Kornberg A (1960) Biological Synthesis of Deoxyribonucleic Acid. *Science (80-)* 131(3412):1503–1508.

89. Ollis DL, Brick P, Hamlin R, Xuong NG, Steitz TA (1985) Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature* 313(6005):762–766.

90. Mönttinen H a M, Ravantti JJ, Stuart DI, Poranen MM (2014) Automated structural comparisons clarify the phylogeny of the right-hand-shape polymerases. *Mol Biol Evol* 31(10):2741–2752.

91. Beard WA, Wilson SH (2000) Structural design of a eukaryotic DNA repair polymerase: DNA polymerase beta. *Mutat Res* 460(3-4):231–244.

92. Werner F, Grohmann D (2011) Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* 9(2):85–98.

93. Klenow H, Henningsen I (1970) Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. *Proc Natl Acad Sci U S A* 65(2):168–175.

94. Lawyer FC, et al. (1993) High level expression, purification and enzymatic characterization of full length Thermus aquaticus DNA polymerase and truncated a foom deficient in 5’ to 3' exonuclease activity. *J Biol Chem* 2(4):275–287.

95. Lundberg KS, et al. (1991) High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene* 108(1):1–6.

96. Kong H, Kucera RB, Jack WE (1993) Characterization of a DNA polymerase from the hyperthermophile archaea Thermococcus litoralis. *J Biol Chem* 268(3):1965–1975.

97. Gardner AF, Jack WE (2002) Acyclic and dideoxy terminator preferences denote divergent sugar recognition by archaeon and Taq DNA polymerases. *Nucleic Acids Res* 30(2):605–613.

98. Gardner AF, Jack WE (1999) Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res* 27(12):2545–2553.

99. Ichida JK, Horhota A, Zou K, McLaughlin LW, Szostak JW (2005) High fidelity TNA synthesis by Therminator polymerase. *Nucleic Acids Res* 33(16):5219–5225.

100. Chaput JC, Szostak JW (2003) TNA synthesis by DNA polymerases. *J Am Chem Soc* 125(31):9274–9275.

101. Ebert MO, Mang C, Krishnamurthy R, Eschenmoser A, Jaun B (2008) The structure of a TNA-TNA complex in solution: NMR study of the octamer duplex derived from α-(L)-threofuranosyl-(3’-2')-CGAATTCG. *J Am Chem Soc* 130(45):15105–15115.

102. Schöning K, et al. (2000) Chemical etiology of nucleic acid structure: the alpha-threofuranosyl-(3’-->2') oligonucleotide system. *Science (80-)* 290(5495):1347–1351.

103. Dunn MR, et al. (2015) DNA Polymerase-Mediated Synthesis of Unbiased Threose Nucleic Acid (TNA) Polymers Requires 7-Deazaguanine To Suppress G:G Mispairing during TNA Transcription. *J Am Chem Soc* 137(12):4014–4017.

104. Yu H, Zhang S, Dunn MR, Chaput JC (2013) An efficient and faithful in vitro replication system for threose nucleic acid. *J Am Chem Soc* 135(9):3583–3591.

105. Ichida JK, et al. (2005) An in vitro selection system for TNA. *J Am Chem Soc* 127(9):2802–2803.

106. Horhota A, et al. (2005) Kinetic analysis of an efficient DNA-dependent TNA polymerase. *J Am Chem Soc* 127(20):7427–7434.

107. Zhang S, Chaput JC (2013) Synthesis and enzymatic incorporation of a-l-threofuranosyl adenine triphosphate (tATP). *Bioorganic Med Chem Lett* 23(5):1447–1449.

108. Chaput JC, Ichida JK, Szostak JW (2003) DNA polymerase-mediated DNA synthesis on a TNA template. *J Am Chem Soc* 125(4):856–857.

109. Cadwell RC, Joyce GF (1992) Randomization of genes by PCR mutagenesis. *Genome Res* 2(1):28–33.

110. Leung DW, Chen E, Goeddel D V (1989) A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. *Technique* 1:11–15.

111. Skelly J V, Edwards KJ, Jenkins TC, Neidle S (1993) Crystal structure of an oligonucleotide duplex containing G.G base pairs: influence of mispairing on DNA backbone conformation. *Proc Natl Acad Sci U S A* 90(3):804–808.

112. Staiger N, Marx A (2010) A DNA polymerase with increased reactivity for ribonucleotides and C5-modified deoxyribonucleotides. *ChemBioChem* 11(14):1963–1966.

113. Pinheiro VB, et al. (2012) Synthetic Genetic Polymers Capable of Heredity and Evolution. *Science (80-)* 336(6079):341–344.

114. Chen T, Romesberg FE (2014) Directed polymerase evolution. *FEBS Lett* 588(2):219–229.

115. Sauter KBM, Marx A (2006) Evolving thermostable reverse transcriptase activity in a DNA polymerase scaffold. *Angew Chemie - Int Ed* 45(45):7633–7635.

116. Lutz S, Burgstaller P, Benner SA (1999) An in vitro screening technique for DNA polymerases that can incorporate modified nucleotides. Pseudothymidine as a substrate for thermostable polymerases. *Nucleic Acids Res* 27(13):2792–2798.

117. Aposhian H V., Kornberg A (1962) Enzymatic Synthesis of Deoxyribonucleic Acid. *J Biol Chem* 237(2):519–525.

118. Summerer D, Marx A (2002) A molecular beacon for quantitative monitoring of the DNA polymerase reaction in real-time. *Angew Chemie - Int Ed* 41(19):3620–3622.

119. Dorjsuren D, et al. (2009) A real-time fluorescence method for enzymatic characterization of specialized human DNA polymerases. *Nucleic Acids Res* 37(19):1–12.

120. Griep MA (1995) Fluorescence recovery assay: a continuous assay for processive DNA polymerases applied specifically to DNA polymerase III holoenzyme. *Anal Biochem* 232(2):180–189.

121. Yu L, Hu G, Howells L (2002) Fluorescence-based, high-throughput DNA polymerase assay. *Biotechniques* 33(4):938–941.

122. Tveit H, Kristensen T (2001) Fluorescence-based DNA polymerase assay. *Anal Biochem* 289(1):96–98.

123. Reetz MT, Carballeira JD (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc* 2(4):891–903.

124. Li Y, Mitaxov V, Waksman G (1999) Structure-based design of Taq DNA polymerases with improved properties of dideoxynucleotide incorporation. *Proc Natl Acad Sci U S A* 96(17):9491–9496.

125. Summerer D, Rudinger NZ, Detmer I, Marx A (2005) Enhanced fidelity in mismatch extension by DNA polymerase through directed combinatorial enzyme design. *Angew Chemie - Int Ed* 44(30):4712–4715.

126. Gloeckner C, Sauter KBM, Marx A (2007) Evolving a thermostable DNA polymerase that amplifies from highly damaged templates. *Angew Chemie - Int Ed* 46(17):3115–3117.

127. Siegmund V, Santner T, Micura R, Marx A (2012) Screening mutant libraries of T7 RNA polymerase for candidates with increased acceptance of 2′-modified nucleotides. *Chem Commun* 48(79):9870–9872.

128. Sweasy JB, Loeb LA (1993) Detection and characterization of mammalian DNA polymerase beta mutants by functional complementation in Escherichia coli. *Proc Natl Acad Sci U S A* 90(10):4626–4630.

129. Hathaway TR (1996) Human Immunodeficiency Virus Reverse Transcriptase. *J Biol Chem* 271(9):4872–4878.

130. Camps M, Naukkarinen J, Johnson BP, Loeb LA (2003) Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proc Natl Acad Sci U S A* 100(17):9727–9732.

131. Patel PH, Kawate H, Adman E, Ashbach M, Loeb LA (2001) A Single Highly Mutable Catalytic Site Amino Acid Is Critical for DNA Polymerase Fidelity. *J Biol Chem* 276(7):5044–5051.

132. Patel PH, Loeb LA (2000) DNA polymerase active site is highly mutable: evolutionary consequences. *Proc Natl Acad Sci U S A* 97(10):5095–5100.

133. Suzuki M, Baskin D, Hood L, Loeb LA (1996) Random mutagenesis of Thermus aquaticus DNA polymerase I: concordance of immutable sites in vivo with the crystal structure. *Proc Natl Acad Sci U S A* 93(18):9670–9675.

134. Glick E, Vigna KL, Loeb LA (2001) Mutations in human DNA polymerase eta motif II alter bypass of DNA lesions. *EMBO J* 20(24):7303–7312.

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

136. Smith GP, Petrenko VA (1997) Phage Display. *Chem Rev* 97(96):391–410.

137. Brunet E, Chauvin C, Choumet V, Jestin J-L (2002) A novel strategy for the functional cloning of enzymes using filamentous phage display: the case of nucleotidyl transferases. *Nucleic Acids Res* 30(9):e40.

138. Jestin J, Kristensen P, Winter G (1999) A Method for the Selection of Catalytic Activity Using Phage Display and Proximity. *Angew Chemie - Int Ed* 38(8):1124–1127.

139. Xia G, et al. (2002) Directed evolution of novel polymerase activities: mutation of a DNA polymerase into an efficient RNA polymerase. *Proc Natl Acad Sci U S A* 99(10):6597–6602.

140. Fa M, Radeghieri A, Henry AA, Romesberg FE (2004) Expanding the substrate repertoire of a DNA polymerase by directed evolution. *J Am Chem Soc* 126(6):1748–1754.

141. Vichier-Guerre S, Ferris S, Auberger N, Mahiddine K, Jestin JL (2006) A population of thermostable reverse transcriptases evolved from Thermus aquaticus DNA polymerase I by phage display. *Angew Chemie - Int Ed* 45(37):6133–6137.

142. Leconte AM, Chen L, Romesberg FE (2005) Polymerase evolution: Efforts toward expansion of the genetic code. *J Am Chem Soc* 127(36):12470–12471.

143. Yang G, Withers SG (2009) Ultrahigh-throughput FACS-based screening for directed enzyme evolution. *ChemBioChem* 10(17):2704–2715.

144. Tawfik DS, Griffiths AD (1998) Man-made cell-like compartments for molecular evolution. *Nat Biotechnol* 16(7):652–656.

145. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A* 98(8):4552–4557.

146. Anna SL, Bontoux N, Stone HA (2003) Formation of dispersions using "flow focusing" in microchannels. *Appl Phys Lett* 82(3):364–366.

147. Umbanhowar PB, Prasad V, Weitz D a. (2000) Monodisperse emulsion generation via drop break off in a coflowing stream. *Langmuir* 16(2):347–351.

148. Nisisako T, Okushima S, Torii T (2005) Controlled formulation of monodisperse double emulsions in a multiple-phase microfluidic system. *Soft Matter* 1(1):23.

149. Miller OJ, et al. (2006) Directed evolution by in vitro compartmentalization. *Nat Methods* 3(7):561–570.

150. Ghadessy FJ, et al. (2004) Generic expansion of the substrate spectrum of a DNA polymerase by directed evolution. *Nat Biotechnol* 22(6):755–759.

151. Ong JL, Loakes D, Jaroslawski S, Too K, Holliger P (2006) Directed Evolution of DNA Polymerase, RNA Polymerase and Reverse Transcriptase Activity in a Single Polypeptide. *J Mol Biol* 361(3):537–550.

152. Ramsay N, et al. (2010) CyDNA: Synthesis and replication of highly Cy-Dye substituted DNA by an evolved polymerase. *J Am Chem Soc* 132(14):5096–5104.

153. Cozens C, Pinheiro VB, Vaisman A, Woodgate R, Holliger P (2012) A short adaptive path from DNA to RNA polymerases. *Proc Natl Acad Sci U S A* 109(21):8067–8072.

154. Griffiths AD, Tawfik DS (2006) Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol* 24(9):395–402.

155. Holtze C, et al. (2008) Biocompatible surfactants for water-in-fluorocarbon emulsions. *Lab Chip* 8(10):1632–1639.

156. Garstecki P, Fuerstman MJ, Stone HA, Whitesides GM (2006) Formation of droplets and bubbles in a microfluidic T-junction-scaling and mechanism of break-up. *Lab Chip* 6(3):437–446.

157. Teh S-Y, Lin R, Hung L-H, Lee AP (2008) Droplet microfluidics. *Lab Chip* 8(2):198–220.

158. Riche CT, Zhang C, Gupta M, Malmstadt N (2014) Fluoropolymer surface coatings to control droplets in microfluidic devices. *Lab Chip* 14(11):1834–41.

159. Baret J-C, et al. (2009) Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab Chip* 9(13):1850–1858.

160. Bernath K, et al. (2004) In vitro compartmentalization by double emulsions: Sorting and gene enrichment by fluorescence activated cell sorting. *Anal Biochem* 325(1):151–157.

161. Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5’----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A* 88(16):7276–7280.

162. Zinchenko A, et al. (2014) One in a million: Flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Anal Chem* 86(5):2526–2533.

163. Joyce GF (2002) The antiquity of RNA-based evolution. *Nature* 418(6894):214–221.