Engineering Enzyme Systems by Recombination

Thesis by

Devin Lee Trudeau

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2014

(Defended December 11, 2013)

ACKNOWLEDGEMENTS

I’d like to thank my advisor Frances Arnold for seeing me through four years of failure and success. Frances has continually challenged my ideas and work, which has taught me how to be a better scientist, thinker, and bioengineer. In her lab I’ve had the freedom to pursue many different ideas, and interact with people of diverse personalities and research interests.

The Arnold lab has been my home these last few years. I’d like to thank Florence Mingardon and Yousong Ding for being my two mentors as I was learning the art of protein engineering. Toni Lee has been a great collaborator on our endoglucanase project, and I’m happy that our efforts were as synergistic as our enzymes. Matthew Smith and Indira Wu have both been invaluable sources of advice on cellulase engineering. Sabrina Sun was a pleasure to work with over the course of two summers and a fourth-year project. Joseph Meyerowitz was my collaborator on the aldehyde tolerance project. Sabine Bastian kept the lab running smoothly, and has provided excellent advice on directed evolution methodology. Jackson Cahn, Ardemis Boghossian, Phil Romero, Ryan Lauchli, John McIntosh, Martin Engqvist, Chris Farwell, and Jane Wang have been great feedback for my (mostly crazy) ideas— and good friends.

I’d like to thank Frank Chang and the rest of my biofuels collaborators in the Ho and Chao labs at Academia Sinica. I’d like to thank my committee members Professors David Tirrell, Thomas F. Miller III, Mikhail Shapiro, Steve Mayo, and Rob Phillips for their time and guidance during my candidacy and defense.

Lastly I would like to thank my parents and brothers for their support and keeping me sane.

ABSTRACT

Homologous recombination is a source of diversity in both natural and directed evolution. Standing genetic variation that has passed the test of natural selection is combined in new ways, generating functional and sometimes unexpected changes. In this work we evaluate the utility of homologous recombination as a protein engineering tool, both in comparison with and combined with other protein engineering techniques, and apply it to an industrially important enzyme: *Hypocrea jecorina* Cel5a.

Chapter 1 reviews work over the last five years on protein engineering by recombination. Chapter 2 describes the recombination of *Hypocrea jecorina* Cel5a endoglucanase with homologous enzymes in order to improve its activity at high temperatures. A chimeric Cel5a that is 10.1 °C more stable than wild-type and hydrolyzes 25% more cellulose at elevated temperatures is reported. Chapter 3 describes an investigation into the synergy of thermostable cellulases that have been engineered by recombination and other methods. An engineered endoglucanase and two engineered cellobiohydrolases synergistically hydrolyzed cellulose at high temperatures, releasing over 200% more reducing sugars over 60 h at their optimal mixture relative to the best mixture of wild-type enzymes. These results provide a framework for engineering cellulolytic enzyme mixtures for the industrial conditions of high temperatures and long incubation times.

In addition to this work on recombination, we explored three other problems in protein engineering. Chapter 4 describes an investigation into replacing enzymes with complex cofactors with simple cofactors, using an *E. coli* enolase as a model system. Chapter 5 describes engineering broad-spectrum aldehyde resistance in *Saccharomyces cerevisiae* by evolving an alcohol dehydrogenase simultaneously for activity and promiscuity. Chapter 6 describes an attempt to engineer gene-targeted hypermutagenesis into *E. coli* to facilitate continuous *in vivo* selection systems.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

NOMENCLATURE

ADH. Alcohol dehydrogenase.

AID. Activation induced cytidine deaminase.

BSA. Bovine serum albumin.

CBHI. Cellobiohydrolase I.

CBHII. Cellobiohydrolase II.

CBM. Cellulose binding module.

Cel5a. Family 5 cellulase (endoglucanase II).

Cel6a. Family 6 cellulase (cellobiohydrolase II).

Cel7a. Family 7 cellulase (cellobiohydrolase I).

ChR. Channelrhodopsin.

CV. Column volume.

Da. Dalton.

2R-DHIV. R-2,3-dihydroxyisovalerate.

DMSO. Dimethyl sulfoxide.

DNPH. 2,4-Dinitrophenylhydrazine.

EcYfaw. Rhamnonate dehydrase (YfaW) from *Escherichia coli*.

EGII. Endoglucanase II.

Fe-S. Iron-sulfur cluster.

GzYfaW. Rhamnonate dehydratase (YfaW) from *Gibberella zeae.*

4-HBA. 4-hydroxybenzaldehyde.

HjCel5a. Cel5a from *Hypocrea jecorina.*

5-HMF. 5-hydroxymethyl furfural.

HPLC. High-performance liquid chromatography.

IPTG. Isopropyl β-D-1-thiogalactopyranoside.

KIV. 2-ketoisovalerate.

LB. Lysogeny (Luria) broth.

MWCO. Molecular weight cut-off.

OD600. Optical density at 600 nm.

PDB. Protein databank.

PdCel5a. Cel5a from *Penicillium decumbens.*

PIPES. Piperazine-N,N′-bis(2-ethanesulfonic acid)

PgCel5a. Cel5a from *Phialophora sp. G5.*

PpCel5a. Cel5a from *Penicillium pinophilum*.

PpYfaW. Rhamnonate dehydratase (YfaW) from *Penicillium pinophilum.*

SCA. Semicarbazide.

SD-Ura. Synthetic defined media without uracil.

SOB. Super-optimal broth.

StYfaw. Rhamnonate dehydratase (YfaW) from *Salmonella typhimurium*.

T_{A50}. Temperature with 50% maximal activity.

T_m. Melting temperature.

T7RNAP. RNA polymerase from T7 phage.

WT. Wild-type.

YPD. Yeast extract, peptone, dextrose media.

Chapter 1

INNOVATION BY HOMOLOGOUS RECOMBINATION

A modified version of this chapter appears in: Trudeau D.L., Smith M.A., Arnold F.H. (2013) "Innovation by homologous recombination", *Current Opinion in Chemical Biology*, and is published with permission from Elsevier.

1.1 Abstract

Swapping fragments among protein homologs can produce chimeric proteins with a wide range of properties, including properties not exhibited by the parents. Computational methods that use information from structures and sequence alignments have been used to design highly functional chimeras and chimera libraries. Recombination has generated proteins with diverse thermo- and mechanical stability, enzyme substrate specificity, and optogenetic properties. Linear regression, Gaussian processes, and support vector machine learning have been used to model sequence-function relationships and predict useful chimeras. These approaches enable engineering of protein chimeras with desired functions, as well as elucidation of the structural basis for these functions.

1.2 Introduction

An important source of the genetic variation that underlies evolution by natural selection is homologous recombination, whereby new sequences are generated by exchange of related segments of genes and genomes. This occurs in diverse processes such as sex, horizontal gene transfer, and V(D)J recombination in the immune system. Researchers have long argued the benefits of recombination (and why sex evolved), which include increasing the fitness variation of a population and enabling removal of deleterious alleles[1].

The costs and benefits of recombination have been studied at the level of individual proteins, particularly as a search strategy for directed evolution[2]. In a pioneering 1998 study, Pim Stemmer and colleagues showed that DNA shuffling (which generates new genetic sequences by both random mutation and recombination) of four cephalosporinases increased resistance to the antibiotic moxalactam by ~270 fold, almost two orders of magnitude more than what was attained with random mutagenesis alone[3]. Recombination has been used since then in a large number of directed evolution efforts, and many groups have contributed to an understanding of how functional and structural properties of recombined, or 'chimeric', proteins depend on factors such as the number and sequence identity of the parents, choice and number of recombination sites, and measures of structural disruption.

In this review, we expand on the topics covered in a previous review from 2007[4] and discuss important new developments that address two key questions: 1) What functional variation can arise from recombining proteins that share related or similar structures? And, 2) what methods might facilitate creation of recombined proteins with predictable properties?

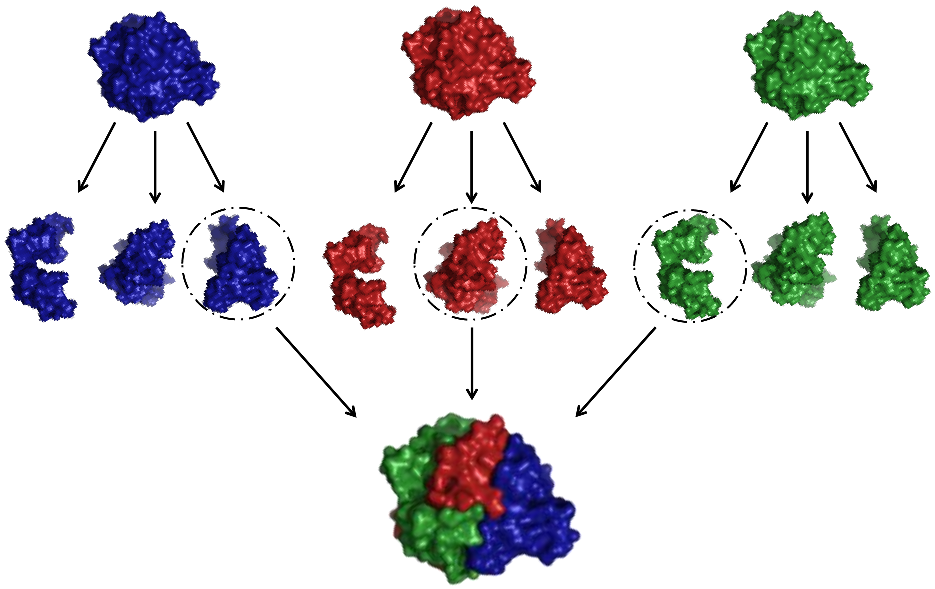


Figure 1-1. Recombination swaps sequence elements from related proteins to create novel chimeras whose properties can differ from the parents'. Parent sequences can be chosen based on structure or sequence alignments, and crossover locations can be chosen to minimize structural and functional disruption. The resulting chimeric proteins can contain dozens of mutations from their closest parents.

1.3 Structural and sequence information facilitates recombination

It is straightforward to recombine genes using DNA shuffling and related methods, as long as there are sufficient stretches of DNA sequence identity to promote crossovers. The more divergent the parent sequences, however, the more difficult it is for methods like DNA shuffling just to generate crossovers. Furthermore, shuffling more divergent sequences introduces more mutations and more structural disruption in the protein, with the consequence that many of the resulting chimeras are non-functional. Juxtaposing elements from different parent proteins can introduce steric clashes or disrupt favorable interactions, resulting in chimeras that do not fold or function. Nonetheless, homologous mutations (mutations chosen from homologous sequences) are significantly less disruptive than random mutations[5; 6].

Judicious choice of recombination sites (crossover locations) can mitigate mutation-induced disruption. Minimizing structural disruption enriches the fraction of folded and functional proteins in a given chimera library[7; 8]. A further advantage is that libraries with fixed crossover locations can be constructed by any number of methods for assembling DNA fragments. Although the sequence space is dramatically reduced when crossovers are fixed, the fitness landscape can be sampled and searched quite efficiently using machine-learning methods[9; 10; 11].

The SCHEMA method for choosing crossover locations to make a high-quality library of chimeric proteins uses a simple metric to assess disruption. The SCHEMA disruption energy E is the sum of all broken contacts in a chimera. Two amino acids are in contact if they are within a certain distance of one another (e.g. 4.5Å) in the structure. If a chimera inherits a contacting pair that is not present in a parent sequence, that contact is said to be broken. This assumes that new contacts are deleterious far more often than they are beneficial. Chimeras are more likely to fold and function if they have fewer broken contacts, and therefore a lower SCHEMA energy E.

Protein crossovers can be chosen to minimize a chimera library's average SCHEMA energy, given constraints on other parameters, such as the size of a recombined element or the average desired mutation level. In recent years, this laboratory has used SCHEMA recombination to make chimeric cytochrome P450s[10], cellulases[12; 13; 14; 15], and arginases[16] that have much higher levels of mutation (sometimes 100 mutations or more) than what is attainable using DNA shuffling or random mutagenesis. The Silberg and Suh labs have recently applied SCHEMA recombination to the capsid protein of adeno-associated virus (AAV), creating chimeras of AAV serotypes 2 and 4 with over 100 mutations relative to each parent[17]. They found that chimeric AAV structural integrity and infectivity were correlated with low SCHEMA energies, suggesting that this metric can be important for recombination of higher-order molecular assemblies like viruses.

Several laboratories have considered whether there might be better metrics than counting broken contacts for predicting whether a given chimera will fold and function and for designing libraries of shuffled proteins. For example, Maranas and coworkers developed the Famclash algorithm, which uses a multiple sequence alignment to predict amino acid interactions based on pair-wise conservation of charge, volume, and hydrophobicity[18]. Another scoring function from the same group, S2, combines conservation of amino acid properties with a SCHEMA-like contact metric[19]. Bailey-Kellogg and coworkers generalized the structural contact idea to multi-residue interactions using a weighted hypergraph model[20]. Residues within 8Å were defined as interacting, and their interaction score was based on evolutionary conservation in a multiple sequence alignment. This metric could predict functionality in a published beta-lactamase chimera library. None of these proposed metrics, however, have been tested in library design.

A chimera library should have a high fraction of functional chimeras with high sequence diversity. Since there is a trade-off between these properties, a good library design optimizes this trade-off (*i.e.* the library is "Pareto optimal"). Bailey-Kellogg and coworkers developed computational algorithms to predict library functionality and diversity and find the chimera libraries that are Pareto optimal[21]. When tested on published purE family proteins and beta-lactamase chimeras, this approach was reported to give better library designs than simply setting a minimum on fragment size.

1.4 Exploring the limits of homologous recombination

Even with the crossovers chosen to minimize average disruption, chimeras of highly divergent parents usually have high levels of structural disruption because they have high levels of mutation. Romero *et al.* [8] developed a random field model parameterized with experimental data from eight SCHEMA libraries to investigate how parent sequence identity and number of crossovers affect the fraction of chimeras that are expected to be folded and functional. Parent sequence identity is the most important factor, but the number of crossovers also contributes to disruption, with more crossovers leading to greater disruption, on average. Choosing crossover locations to minimize disruption can improve the library significantly. Merely choosing contiguous sequence elements, however, also captures and retains many local contacts that would be broken if homologous mutations were made individually rather than taken in blocks from parent proteins. Thus the conservative nature of recombination comes from both the conservative nature of the individual homologous mutations and conservation of their local interactions in a sequence block.

To enable recombination of distant parent sequences, one can relax the constraint that shuffled sequence elements be contiguous in primary sequence and instead shuffle elements that are contiguous in the three-dimensional structure, thereby conserving even more local interactions. Smith *et al.*[22] recently described such a 'non-contiguous recombination' design method. Amino acids are modeled as nodes in a graph, and edges are placed between nodes when SCHEMA contacts exist between their corresponding amino acids. Graph partitioning algorithms are then used to find the optimal division of amino acids into recombining blocks. Smith *et al*. designed a chimera that takes about half its barrel structure from a fungal beta-glucosidase and half from a bacterial beta-glucosidase that is only 41% identical. The resulting chimera had 144 mutations relative to the closest parent (out of 474 amino acids) and was folded and catalytically active (although its activity was lower than that of the parents, it was readily recovered by directed evolution.). The x-ray crystal structure showed that blocks from each parent retained their original, parental structures; in other words, the recombined protein was a true structural chimera of the two parent proteins. Noncontiguous recombination has also been tested on fungal cellobiohydrolase I's (CBHI)[23], where 32 of 35 chimeras constructed by total gene synthesis were active cellulases, despite having an average of 83 mutations relative to the closest parent.

Do protein parents really need to be homologous, i.e. evolutionarily related, or can proteins accommodate structurally compatible swaps from parents whose overall structures are different? Because homologous parents generally exhibit much greater sequence identity and therefore less mutational disruption upon recombination, we might expect structural similarity to be insufficient for successful recombination, at least on average. Recent experiments from the Höcker laboratory illustrate chimeras constructed by combining structurally similar blocks taken from unrelated proteins. Noting that the $(\beta\alpha)_5$-flavodoxin-like fold from bacterial response regulator CheY is structurally similar to half of the $(\beta\alpha)_8$-barrel fold from imidazole glycerol phosphate synthase (HisF), Bharat *et al*. replaced this half of HisF with CheY[24]. This resulted in a stable protein with a $(\beta\alpha)_8$-like fold (save for an additional β strand inside the $(\beta\alpha)_8$-barrel) and 81 mutations (out of 253) from the closest parent, HisF (see **Figure 1-2**). Further engineering using Rosetta design introduced five mutations at the interface between the two pieces that allowed the extra β strand to be removed, resulting in a more natural $(\beta\alpha)_8$-barrel fold[25]. The HisF-CheY chimera could be engineered to bind a phosphorylated substrate by targeted mutation at two residues known to confer binding in the related HisA protein. Half of the $(\beta\alpha)_8$-barrel from HisF could also be recombined with the $(\beta\alpha)_5$-flavodoxin-like fold from nitrite response regulator NarL to make a stable $(\beta\alpha)_8$-barrel fold[26].

Zheng *et al*. created an algorithm to assist site-directed swapping of a fragment from one protein into another, with the only constraint being *local* sequence or structure similarity, as measured by sequence identity or topological similarity[27]. This approach to identifying swappable elements of proteins whose overall folds are different has not yet been tested experimentally.

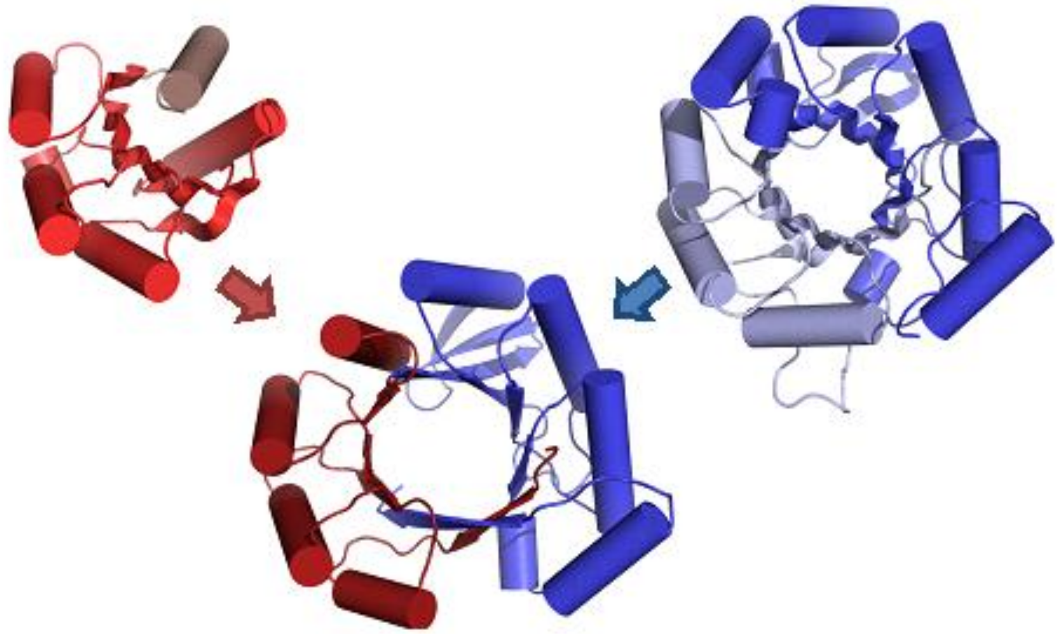


Figure 1-2. **Recombination of structurally similar elements from unrelated proteins.** Bharat *et al.* (2008) used the (βα)$_5$-flavodoxin-like fold from bacterial response regulator CheY (top left) to replace half of the (βα)$_8$-barrel fold from imidazole glycerol phosphate synthase HisF (top right). This created a stable (βα)$_8$-barrel-like fold, with an extra β strand inside the barrel— a (β$_9$α$_8$)-barrel. Further mutation at the interface could remove this extra β strand to make a more natural (βα)$_8$-barrel[24; 25].

1.5 Recombination promotes innovation

Recombination can generate libraries with a high fraction of folded proteins and a high level of mutational diversity. Experiments have shown that the chimeric proteins can also exhibit a range of properties, including properties not exhibited by any of the parents. Thus recombination is both conservative and innovative. Here we cover three properties that have been investigated in recent work: stability, enzyme substrate spectrum, and optogenetic properties of membrane rhodopsins.

1.5.1 Stability

Stability is one of the most important protein properties; it is necessary for folding and function, promotes evolvability by allowing new mutations that are required for function but might be too destabilizing to accumulate, and is important for almost any application. To create highly stable fungal cellulases, Heinzelman and coworkers used SCHEMA to recombine five class I cellobiohydrolases (CBHI) from *Talaromyces emersonii, Chaetomium thermophilum, Thermoascus aurantiacus, Hypocrea jecorina,* and *Acremonium thermophilum*[14]. They cloned and expressed a sample set of 32 chimeras consisting of single block substitions between homologous enzymes. As shown in **Figure 1-3**, these sample chimeras exhibited significant variation in thermostability, including higher and lower than the parent enzymes. Heinzelman *et al.* then combined stabilizing blocks to create chimeras that were both highly thermostable and more active than the parent enzymes at their respective optimum temperatures. SCHEMA recombination was also used to make class II cellobiohydrolases[12] and family 48 cellulases[15] that were more thermostable and more active than their respective parents. Romero and Stone *et al.*[16] used SCHEMA and Gaussian process machine learning tools[9] (see Modeling section, below) to create chimeras of human Arginases I and II (61% sequence identity) with longer half-lives at 37°C, which is important for therapeutic applications.

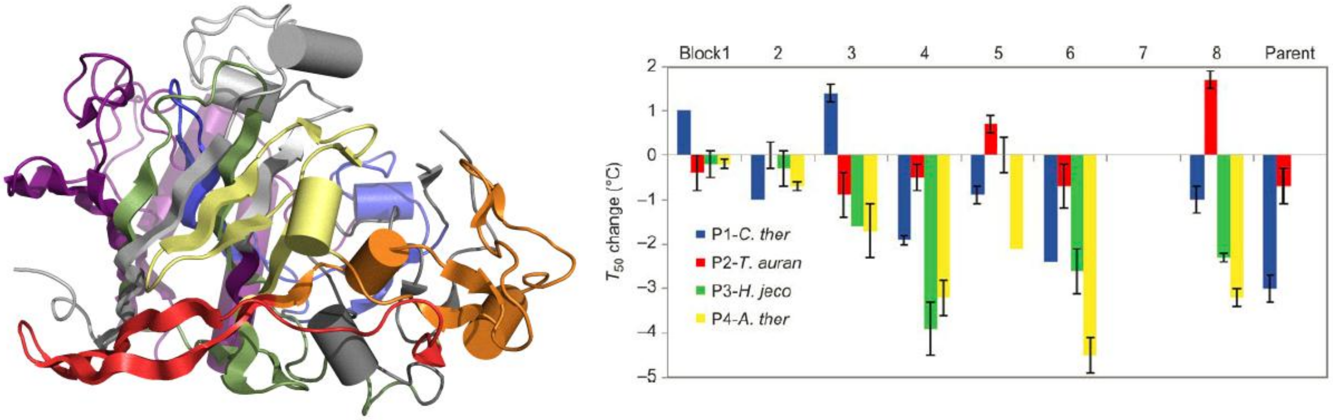


Figure 1-3. **Thermostability contributions of recombined blocks can be determined using linear regression of data from a sample set of chimeras.** Heinzelman *et al.* (2010) made a chimera library of class I cellobiohydrolases (CBHI), with parent enzymes from *T. emersonii, C. thermophilum, T. aurantiacus, T. reesei,* and *A. thermophilum.* Recombination sites chosen by SCHEMA generated the blocks shown in different colors on the *T. emersonii* CBHI structure (left). Individual blocks make different contributions to thermostability relative to blocks from *T.emersonii* CBHI (right). Thermostabilizing blocks were combined to make thermostable chimeras. Modified from reference 14.

Another interesting property is *mechanical* stability, important for proteins in tissue extracellular matrices, spider silk, and other biomaterials. The Li lab explored the structural basis of mechanical stability by recombining structural elements from two homologous immunoglobulin domains (I27 and I32) from the muscle protein titin[28; 29]. Using atomic force microscopy to test the mechanical stability of the different chimeras, the Li lab correlated stability with specific sequence and structure elements. Recombination has also been used by Billings *et al*. and Lu *et al*. to explore mechanical stability of immunoglobulin domains[30; 31], and by Ng *et al*. to explore mechanical stability of fibronectin type III domains[32].

1.5.2 Enzyme substrate spectrum

Recombination can generate large and sometimes quite unexpected changes in enzyme activity on non-native substrates, including the ability to accept new substrates. Clouthier *et al.* looked at the ability of three chimeras of TEM-1 and PSE-4 beta-lactamases (43% identity) to hydrolyze five different cephalosporins[33]. Although the chimera activities on each substrate were usually intermediate between the activities of the parent enzymes, one of only three they studied was almost twice as active on the clinically important antibiotic cefotaxime as the most active parent.

Focused chimeragenesis that targets structural elements in a substrate binding pocket could help transfer a catalytic activity from an enzyme that is difficult to express or manipulate into a more amenable fold. For example, Chen *et al.* transferred short peptide sequences (three to six amino acid residues) in the substrate recognition pocket of the *Diploptera punctate* (cockroach) cytochrome P450 CYP4C7 into the well-studied cytochrome P450 BM3[34]. They reported that the chimeras exhibited increased activity on farnesol and decreased activity on fatty acids, as well as different hydroxylation and epoxidation products from farnesol. Similarly, Campbell *et al.* replaced three substrate-binding loops from *Pyrococcus furiosus* alcohol dehydrogenase D with those from a human aldose reductase homolog[35]. The resulting chimera retained the extreme thermostability of its thermophilic parent, but also gained the human parent's activity on glyceraldehyde and bias towards using NADP(H) as cofactor.

Van Beek *et al*. swapped a substrate-binding subdomain of thermostable phenylacetone monooxygenase with corresponding elements from homologs that accept a broader range of substrates, a cyclohexanone monooxygenase and a steroid monooxygenase[36]. These Baeyer-Villiger monooxygenases are potential industrial biocatalysts. The resulting two chimeras were more stable than the parent cyclohexanone monooxygenase and steroid monooxygenase and exhibited broad substrate ranges, with higher activity and enantioselectivity than their parents on selected substrates.

In a more library-based approach, Jones shuffled six loop regions from serine proteases of the subtilisin family into a Savinase framework[37]. The loops were selected for their known functional importance in substrate binding, metal ion binding, and catalysis. He found chimeric proteases with increased activity on and specificity towards each of four tested colorimetric peptide substrates, including two substrates that Savinase hydrolyzes poorly.

1.5.3 Optogenetic properties

Optogenetics enables researchers to control individual neurons by light activation of heterologously-expressed microbial opsins[38]. This technology provides an unprecedented ability to control and interrogate neuronal behavior; however, it is constrained by the photocurrent characteristics of the available opsins. These characteristics include activation wavelength and kinetics, and ion permeability. Recent studies have shown that these properties can be tuned by recombination of homologous opsins. *Chlamydomonas reinhardtii* channelrhodopsin-2 (ChR2) is commonly used for membrane depolarization in optogenetics[39]. The photocurrent of its paralog channelrhodopsin-1 (ChR1) is too low to depolarize neurons. However, ChR1 has the advantage of having maximal activation at a lower light frequency, lower desensitization after stimulus, and faster on/off kinetics. Wang *et al.* looked for structural determinants of these properties by making single crossover chimeras between ChR2 and ChR1, targeting loops between predicted alpha helices. They found that the wavelength activation profiles, as well as the desensitization profiles of the chimeras, were intermediate between the two parents. Most of this variation was found to come from the fifth transmembrane helix, particularly the Y226(ChR1)/N187(ChR2) site. Two chimeras that were similar to ChR2 but with improved properties were found: one had broader activation wavelength sensitivity and lower desensitization, and one had very fast on/off kinetics and small desensitization. Li *et al.* and Wen *et al.* also found similar results when they recombined ChR2 and ChR1[40; 41].

To create a red-shifted opsin for combinatorial control of neuron activation, Yizhar *et al.* recombined ChR2 and Channelrhodopsin-1 from *Volvox carteri* (VChR1), which was known to have a redshift of over 70nm compared to ChR2, but also low expression and weak photocurrents[38]. Yizhar *et al.* made single crossover recombinants of VChR1 and ChR1 and measured expression and photocurrent in HEK cells. By replacing the first two alpha helices of VChR1 with the corresponding ones from ChR1, they were able to increase VChR1 expression and photocurrent while retaining its large redshift. Interestingly, this chimera had a slower deactivation rate than either parent, a property that is not optimal for control of neurons. However, this rate could be improved by introducing two mutations known to improve deactivation rate in ChR2. With this new chimeric opsin, Yizhar *et al.* were able to explore neuronal control of social behavior in mice.

1.6 Modeling and predicting desired chimeras

The ability to identify the sequences of the most desirable chimeras in a given family using data modeling approaches contributes greatly to the utility of recombination as a protein engineering tool. Recent experiments have shown that researchers can design and construct a small sample set of chimera sequences (perhaps only a few dozen), characterize their properties, and use the data to predict the chimera family members that have the most desirable property profiles. This approach makes great use of rapid, inexpensive gene synthesis to make highly informative sample sets and test predicted chimeras.

The large-scale ‘recombinational fitness landscape’[8] is characterized by a high degree of additivity that correlates with mutations being in conserved parental structural contexts (as opposed to new interfaces generated by recombination), which is exactly what SCHEMA recombination attempts to maximize. That the landscape is largely additive means that relatively simple models can be used to build sequence-function models and predict the properties of chimeras that have not yet been tested. Linear regression can be used, for example, to predict highly stable chimeras from small sample data sets from SCHEMA and noncontiguous recombination libraries[42]. This approach has generated a variety of stable, active enzymes[10; 12; 13; 14; 15; 16; 23].

Modeling by linear regression requires a relatively small sample set because chimera sequences are much reduced compared to the whole protein (chimera sequences are described at the block rather than amino acid level). However, the contributions of individual mutations cannot be identified unless they are made separately, as Heinzelman *et al*. did to uncover a single highly stabilizing mutation in a fungal cellobiohydrolase II block[13].

Romero *et al.*[9] recently used a new class of Bayesian machine-learning tools called Gaussian processes to sample and model the fitness landscape. Their methods can be used to both design maximally-informative sample sets and predict improved sequences. With a structure-based kernel function to describe how sequences are expected to co-vary (i.e. it does not assume simple additivity, but includes pair-wise interactions between residues), their methods can be used to investigate the contributions of individual mutations, and also combine chimera data with information on single mutations. Romero and coworkers found good predictive ability for cytochrome P450 thermostability, catalytic activity on non-native substrates, and ligand binding affinity. Moreover, the model was able to predict the thermostabilities of cytochrome P450s that had mutations not present in the chimera library, and also predicted a new cytochrome P450 variant that was more thermostable than any previously engineered variant.

How transferrable is information gained from one chimera library to another? Buske *et al.* developed a predictive model based on support vector machine learning[11]. When trained on data from a SCHEMA library generated by recombining three bacterial cytochrome P450s, their model could predict the properties of sequences generated by DNA shuffling of human P450s.

1.7 Conclusions

Adaptation requires variation. Homologous mutations have passed the test of natural selection for compatibility with parental fold and function, and new combinations of homologous substitutions can generate new functional diversity. A growing body of experimental data attests to this dual conservative and innovative nature of recombination. The reduced size and overall additive structure of the recombinational fitness landscape, at least for some properties, make it amenable to searches using machine-learning. Making use of the information already inherent in the products of evolution by natural evolution, recombination is a useful tool for protein engineering and promises further insights into the sequence and structure determinants of protein function.

Chapter 2

CHIMERAGENESIS OF FUNGAL ENDOGLUCANASE II REVEALS EPISTATIC CONSTRAINTS TO THERMOSTABILIZATION BY RECOMBINATION

2.1 Abstract

Recombination is an efficient way of using natural protein variation to create chimeras with diverse properties. However, recombination can also create non-functional chimeras by breaking beneficial amino acid interactions and introducing steric clashes between amino acids. Studies have employed algorithms like SCHEMA to choose recombination breakpoints that minimize these non-favorable amino acid interactions in chimera libraries and thereby improve the fraction of folded and functional chimeras.

We wished to explore quantitatively how non-favorable amino acid interactions between recombined structural subunits affect chimera thermostability, and how successful recombination break point optimization is in reducing these effects. To do this, we used the SCHEMA recombination algorithm to design a chimera library with four fungal endoglucanases as parents, including the industrially-relevant *Hypocrea jecorina* endoglucanase II (HjCel5a). This library had the lowest predicted average number of non-favorable amino acid interactions (or "disruption score") of all previously designed SCHEMA recombination libraries.

We experimentally evaluated a maximally informative test set of this chimera library and found that the chimeras had highly diverse thermostabilities that could be modeled using linear regression. Unlike other SCHEMA libraries, this library had a substantially non-additive component, which could be accounted for by including the disruption score as a parameter in the regression modeling. The effect of disruption was to decrease thermostability by an average of 0.9 °C, resulting in chimera average thermostability being decreased by 11 °C. Despite the effect of disruption score on chimera thermostability, the improved linear regression model

facilitated construction of an HjCel5a mutant, which was over 10 °C more stable than any parent and released 25% more cellobiose at its optimum temperature. These results highlight the importance of accounting for non-favorable amino acid interactions when modeling chimeras, and that these effects can be minimized (but not avoided) by computational chimera library design methods.

2.2 Introduction

Over the last 15 years, homologous protein recombination has been used to engineer properties as diverse as substrate specificity, thermo- and mechanostability, and optogenetic characteristics[43]. An important finding from these studies is that introduction of structurally incompatible amino acids by recombination can result in a high fraction of unfolded or inactive chimeras, and that judicious choice of recombination breakpoints may be needed to create a functional chimera library[7; 8; 18; 19].

The SCHEMA algorithm is one approach that has been developed to improve the folded and functional fraction of chimera libraries[7]. SCHEMA scores a chimera by a simple metric: two amino acids form a contact if their heavy chain atoms are within 4.5 Å of each other in a reference crystal structure, and if a contacting pair is present in chimera but not in any parent, the contact is said to be disrupted[7]. The “disruption score” (E) is the sum of disrupted contacts. Chimera library designs that minimize this disruption score have been found to be enriched in folded and functional variants[8; 10; 12; 15; 16; 44]. Disruption score and mutation level are correlated[21], and therefore library functionality and diversity have inherent trade-offs.

Recently this laboratory has been able to obtain near-optimal trade-offs between functionality and diversity[22]. A protein of interest can be modeled as a graph of interacting amino acids, and graph partitioning algorithms can be used to find the (nearly) optimal partitioning of these amino acids[45]. Since these amino acid partitions are generally non-contiguous in primary sequence, gene synthesis is used to create the predicted chimeras. This approach allows chimera libraries to be made with very low average disruption scores (<E> < 25), while maintaining high average mutation levels (<m> > 50)[23].

How successful are these chimera library designs with low predicted disruption scores at reducing the effects of disrupted amino acid contacts? In particular, in these libraries how amenable to improvement are useful properties like thermostability? We set out to address this question by applying structure-guided recombination to explore thermostability of endoglucanase II (Cel5a), an enzyme that cleaves intrachain β-glucosyl bonds in cellulose. Cel5a constitutes over 55% of endoglucanase activity in the industrially important fungus *Hypocrea jecorina*[46]. Thermostabilized Cel5a would allow high-temperature degradation of cellulose synergistic with other engineered thermostable cellulases[12; 44; 47; 48].

In this study, we recombined *Hypocrea jecorina* Cel5a with three homologues from thermophilic fungi (*Phialophora sp. G5, Penicillium decumbens, Penicillium pinophilum*). This chimera library had the lowest disruption score of any SCHEMA library created thus far, while retaining a high mutation level. The library was enriched in active and stable chimeras, many of which were more stable than any parent. Stabilizing single mutations from the most stable chimeras combined to create an HjCel5a mutant which was more stable and hydrolyzed cellulose more efficiently at high temperatures. Computational analysis of the chimera library found that disrupted amino acid contacts had a significant negative contribution to the thermostability of chimeras, and that accounting for these interactions improved predictability of improved chimeras.

2.3 Non-contiguous recombination of fungal endoglucanase II

Based on the structure of *Hypocrea jecorina* Cel5a (HjCel5a) (PDB code:3QR3)[49] we used graph partitioning[50] to find optimal breakpoints for recombination with three other thermostable cellulases from related fungi: Cel5a from *Phialophora sp. G5* (PgCel5a)[51], Cel5a from *Penicillium decumbens* (PdCel5a)[52], and Cel5a from *Penicillium pinophilum* (PpCel5a)[53]. The four parents have pairwise amino acid identities ranging from 60% to 73% and optimum temperatures from 60 °C to

63 °C. The recombination scheme is shown in **Figure 2-1**, which defines a library with average disruption score <E> of 12.1 and average mutation level <m> of 55.4.

The eight-block, four-parent chimera library defined by the blocks indicated in **Figure 2-1** has 65,536 members. Assaying even 1% of this library would be extremely time-consuming. However, assuming that each block contributes additively to the thermostability of a chimera[10], a linear regression model can be created that predicts the effect of substituting each block into a parent of interest:

$$T_{A50} = a_0 + \sum_i \sum_j a_{ij} x_{ij}$$

In this model, a_0 is the TA50 of an arbitrary parent (*e.g.* HjCel5a), a_{ij} is the effect of substituting block *i* from parent *j*, and x_{ij} is either 1 or 0, depending on whether the block is present. The linear regression model for the library investigated here has 25 parameters, one for a_0, and 24 for each a_{ij} (8 blocks from 3 parents). This model requires the thermostability of at least 25 chimeras (including parents) to be evaluated in order to not be rank deficient. In principle, any combination of blocks can be used in a test set, as long as each block appears at least once. However, to account for non-additive behavior in the library arising from possible interactions between blocks, we chose chimeras that also had maximal mutual information between each block[16].

We synthesized 25 chimeras, appending to the N-terminus of each the cellulose binding module (CBM) from *H. jecorina* Cel5a, as well as a C-terminal His_6 tag for purification. We expressed the chimeras in *Saccharomyces cerevisiae* and purified them using column chromatography. 23 out of 25 chimeras were catalytically active, and we measured their thermostabilities by finding the temperature at which the enzyme loses half of its activity relative to that at its optimum temperature over a 2 h reaction (the "T_{A50}")[23]. We measured T_{A50}'s for all the functional chimeras on crystalline cellulose (Avicel). The test set, shown in grey in **Figure 2-2A**, exhibits a range of thermostability values, from which a regression model can be built.

2.4 Adapting linear regression to account for non-additive block interactions

Linear regression was applied to this initial test set to model block contributions to thermostability. This model was used to inform the design of a second test set to explore potentially stabilizing blocks (shown in black in **Figure 2-2A**). Eighteen additional chimeras that had high mutual information, and that were predicted to be more stable than the parent enzymes, were chosen for expression and characterization. Three chimeras from this set were slightly more stable than any parent (shown in red in **Figure 2-2A**).

To find the chimera with the highest predicted thermostability in the library, we repeated linear regression on the new data set containing both the original and optimized test set chimeras. The R^2 value for this data set was only 0.73, suggesting that the block contributions to chimera stability had significant non-additive components. However, when we added the disruption score of chimeras as an additional parameter to the model (which is known to improve linear regression[15]), we found that the R^2 value increased to 0.92, with each disrupted contact reducing thermostability by an average of 0.91 °C. This model, shown below, adds $a_1 * E_c$ as an additional parameter, where E_c is the disruption score (E) of a chimera, and a_1 is a "disruption penalty", a thermostability decrease associated with the disrupted contacts.

$$T_{A50} = a_0 + a_1 * E_c + \sum_i \sum_j a_{ij} x_{ij}$$

The linear regression model is shown in **Figure 2-2B**, and the predicted contributions of each block to thermostability in the HjCel5a background are shown in **Figure 2-3** (both without (**A**) and with (**B**) disruption score taken into account). In the revised model, fewer blocks are predicted to be stabilizing with respect to HjCel5a.

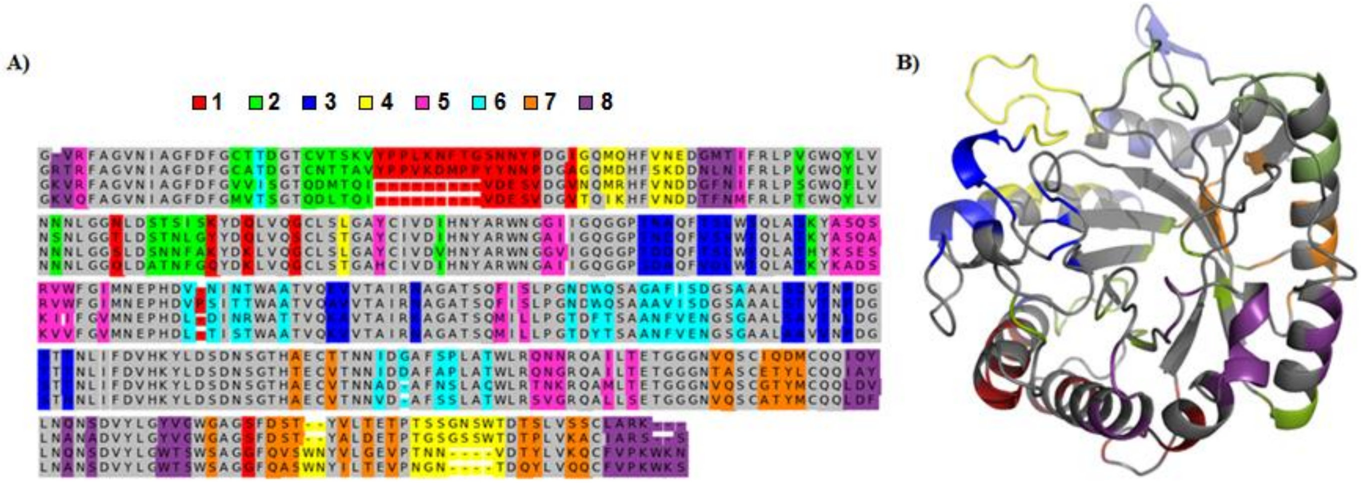

A)
🟥 1 🟩 2 🟦 3 🟨 4 🟪 5 🟦 6 🟧 7 🟪 8
G-VRFAGVNIAGFDFGCTTDGTCVTSKVYPPLKNFTGSNNYPDGIGQMQHFVNEDGMTIFRLPVGWQYLV
GRTRFAGVNIAGFDFGCATDGTCNTTAVYPPVKDMPPYYNNPDGAGQMDHFSKDDNLNIFRLPVGWQYLV
GKVRFAGVNIAGFDFGVVISGTQDMTQI---------VDESVDGVNQMRHFVNDDGFNIFRLPSGWQFLV
GKVQFAGVNIAGFDFGMVTSGTQDLTQI---------VDESVDGVTQIKHFVNDDTFNMFRLPTGWQYLV
NNNLGGNLDSTSISKYDQLVQGCLSLGAYCIVDIHNYARWNGGIIGQGGP[illegible]QF[illegible]W[illegible]QLA[illegible]KYASQS
NSNLGGTLDSTNLGYYDQLVQSCLSTGAYCIVDIHNYARWNGAIIGQGGP[illegible]QF[illegible]W[illegible]QLA[illegible]KYASQA
NNNLGGSLDSNNFAKYDKLVQGCLSLGAYCIVDVHNYARWNGGVIGQGGP[illegible]QF[illegible]W[illegible]QLA[illegible]HYKSES
NNNLGGQLDATNFGQYDKLVQGCLSTGAHCIVDIHNYARWNGAIIGQGGP[illegible]QF[illegible]W[illegible]QLA[illegible]KYKADS
RVWFGIMNEPHDV NINTWAATVQ[illegible]VTAIR[illegible]AGATSQFISLPGNDWQSAGAFISDGSAAAL[illegible]V[illegible]N[illegible]DG
RVWFGIMNEPHDVPSITTWAATVQAVVTAIR[illegible]AGATSQFISLPGNDWQSAAAVISDGSAAAL[illegible]V[illegible]N[illegible]DG
KI-FGVMNEPHDL-DINRWATTVQ[illegible]VTAIR[illegible]AGATSQMILLPGTDFTSAANFVENGSGAAL[illegible]AV[illegible]N[illegible]DG
KVVFGVMNEPHDL-TISTWAATVQ[illegible]VTAIR[illegible]AGATSQMILLPGTDYTSAANFVENGSGAAL[illegible]V[illegible]N[illegible]DG
[illegible]T[illegible]NLIFDVHKYLDSDNSGTHAECTTNNIDGAFSPLATWLRQNNRQAILTETGGGNVQSCIQDMCQQIQY
[illegible]T[illegible]NLIFDVHKYLDSDNSGTHTECVTNNIDDAFAPLATWLRQNGRQAILTETGGGNTASCETYLCQQIAY
[illegible]T[illegible]NLIFDVHKYLDSDNSGTHAECVTNNAD-AFNSLACWLRTNKRQAMLTETGGGNVQSCGTYMCQQLDV
[illegible]T[illegible]NLIFDVHKYLDSDNSGTHAECVTNNVD-AFSSLATWLRSVGRQALLSETGGGNVQSCATYMCQQLDF
LNQNSDVYLGYVGWGAGSFDST--YVLTETPTSSGNSWTDTSLVSSCLARK---
LNANADVYLGYVGWGAGSFDST--YALDETPTGSGSSWTDTPLVKACIARS--S
LNQNSDVYLGWTSWSAGGFQVSWNYVLGEVPTNN-----VDTYLVKQCFVPKWKN
LNANSDVYLGWTSWSAGGFQASWNYILTEVPNGN-----TDQYLVQQCFVPKWKS
B)

Figure 2-1. Cel5a recombination scheme. A) Sequence alignment of Cel5a homologues from *H. jecorina, P. pinophilum, P. decumbens,* and *Philiaphora G5*. Each of the eight blocks is highlighted by a different color, and the conserved residues are in grey. **B**) The x-ray crystal structure of *H. jecorina* Cel5a from PDB 3QR3[49]. Recombined structural blocks are colored to match the sequence alignment.

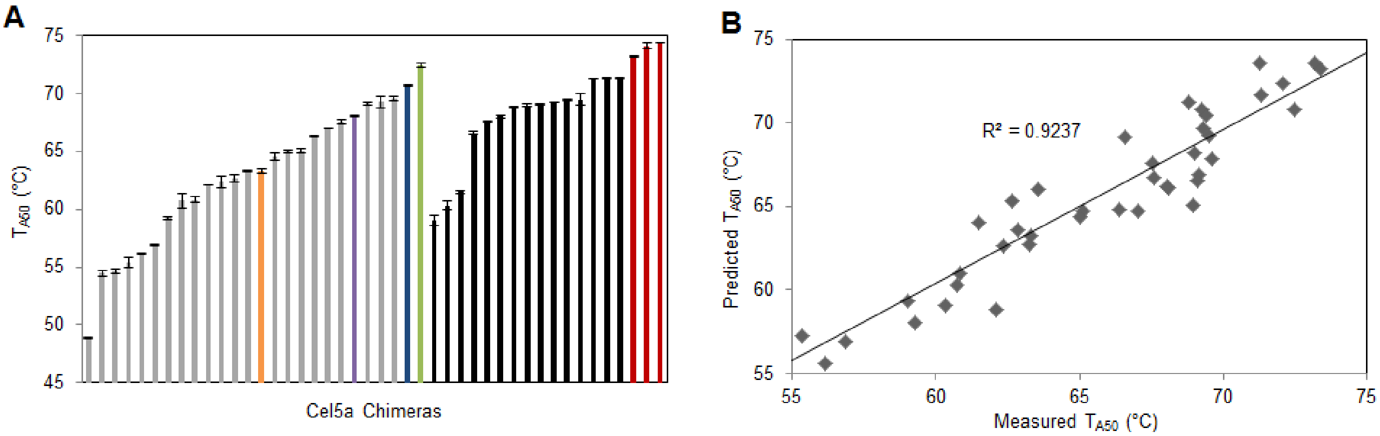


Figure 2-2. Linear regression modeling of a maximally informative subset of the Cel5a recombination library. A) Measured thermostabilities of the 22 active chimeras in initial test set (grey), second optimized test set (black), four parental enzymes PdCel5a, PpCel5a, HjCel5a, and PgCel5a (orange, purple, blue, and green, respectively), and three chimeras with thermostabilities higher than any parent (red). Thermostability was measured using the T_{A50} assay. **B)** Linear regression model of chimera thermostability.

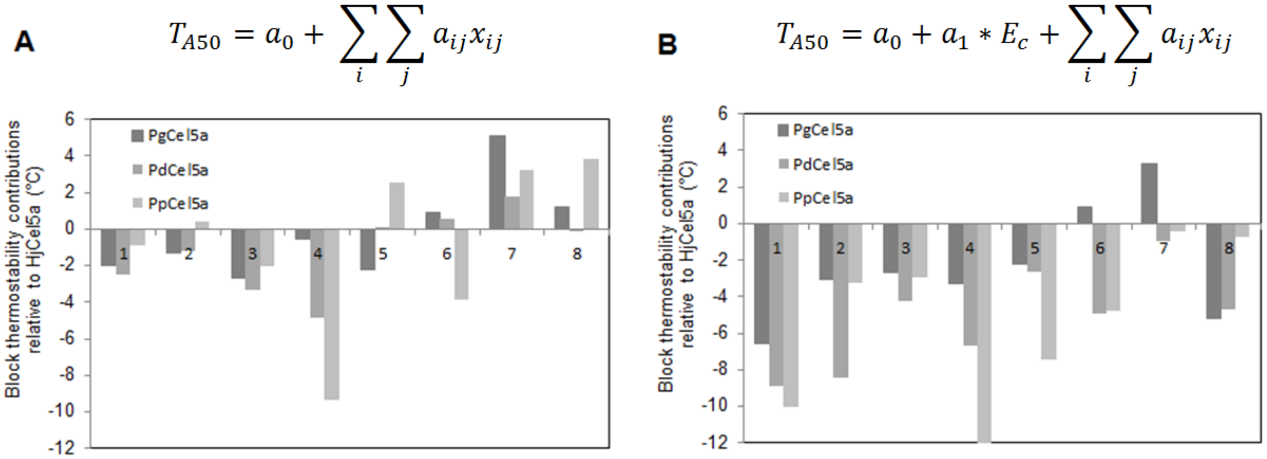


Figure 2-3. Block contributions to thermostability depend on disruption score. Block contribution to thermostability predicted by linear regression, without (**A**) and with (**B**) disruption score as a parameter. Including disruption score reduces number of stabilizing blocks from eight to two.

2.5 Regression modeling predicts a highly stable chimera

The revised linear regression model predicts that only two blocks are stabilizing relative to HjCel5a: blocks 6 and 7 from PgCel5a (+1.0 °C and +3.3 °C, respectively). Combining these two blocks allowed us to create chimera 110, which is 4.3 °C more stable and 18 mutations away from its closest parent, HjCel5a. Since chimera blocks may contain both stabilizing and destabilizing point mutations, we introduced each of the 18 single amino acid mutations in these blocks into HjCel5a individually and tested the thermostabilities of the mutant enzymes.

As shown in **Figure 2-4**, nine of the eighteen mutations were stabilizing, and seven were destabilizing. Two had no effect. We next combined all stabilizing mutations (save for T233V, which compromised activity slightly) to create chimera 110F. This chimera had a stability increase (as measured by T_{A50}) of 10.1 °C relative to HjCel5a (**Figure 2-5A**). Its optimal temperature was also increased by ~10 °C, and its activity was not compromised by thermostabilization.

To evaluate the improvement of 110F over industrially relevant time scales, we compared its activity to that of HjCel5a over 60 h hydrolyses at 60 °C and 70 °C. As shown in **Figure 2-5B**, 110F displays more activity at both temperatures. Importantly, it remains highly active at 70 °C over a 60 h period, while wild-type HjCel5a is nearly inactive at this temperature.

2.6 Disruption penalties vary among SCHEMA libraries

The linear regression model fit of $R^2 = 0.73$ was the lowest seen for any SCHEMA recombination library, but when disruption score was added as a parameter the model improved to $R^2 = 0.92$. We were interested in whether this was a general occurrence in recombination libraries. We analyzed the thermostability models for all previous SCHEMA libraries (cytochrome P450[10], Cel48[15], cellobiohydrolase I[23], and cellobiohydrolase II[12]), both with and without disruption score as a parameter. A recombination library of bacterial endoglucanases (abbreviated bEGII) from a paper in preparation by Chang *et al.* was also analyzed[54]. The analysis is presented in **Table 2-1.** The

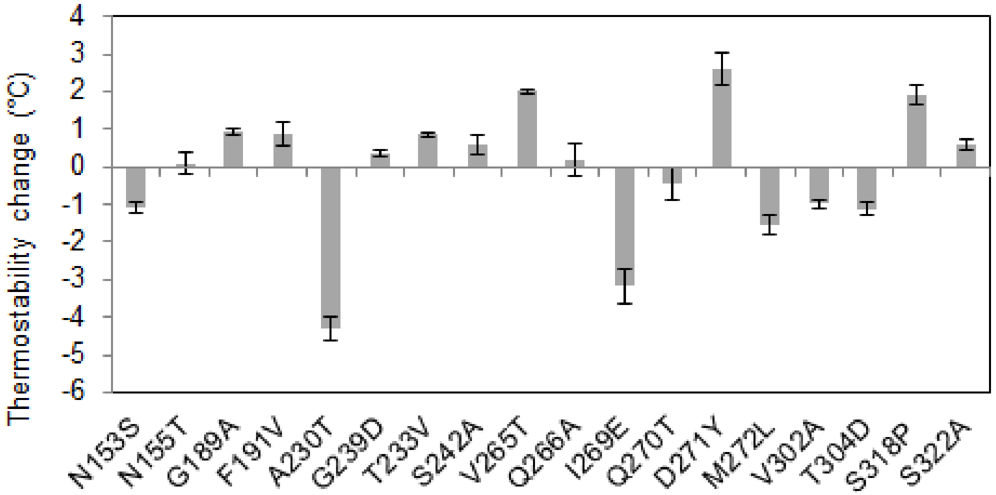


Figure 2-4. Thermostability effects of single mutations from stabilizing blocks in HjCel5a. Point mutations from thermostabilizing blocks were introduced into HjCel5a and their thermostabilities were evaluated using the T_{A50} assay.

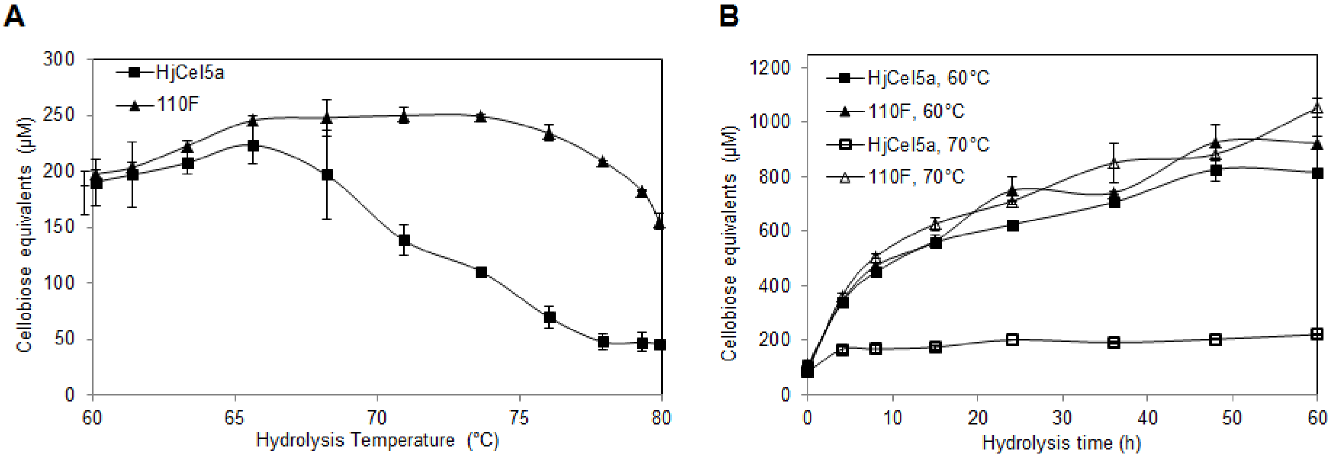


Figure 2-5. A highly stable Cel5a chimera. A) Total cellobiose equivalents released after 2 h Avicel hydrolysis at temperatures ranging from 60 to 80 °C with HjCel5a and 110F. **B)** Total cellobiose equivalents released after 60 h Avicel hydrolysis at 60 °C and 70 °C with HjCel5a and 110F.

Table 2-1: Disruption penalty varies between SCHEMA libraries. Linear regression models were analyzed for previously investigated chimera libraries. Parameters relating to disruption penalty for each library are listed.

	P450	Cel48	CBHI	CBHII	EGII	bEGII
Chimera measurements	44	60	42	58	48	16
Model parameters	17	17	25	17	25	9
Disruption penalty	-0.14 °C	-0.29 °C	-0.66 °C	-0.08 °C	-0.91 °C	-1.40 °C
Average disruption score	33.4	31	24.8	15.7	12.1	9.5
Total number of contacts	2293	2531	2111	1809	1670	2794
Number of amino acids	466	631	441	361	327	306
R^2 without disruption penalty	0.84	0.82	0.87	0.86	0.73	0.67
R^2 with disruption penalty	0.88	0.88	0.92	0.86	0.92	0.87
F-test P-value for disruption penalty	P<2E-16	P<.0168	P<0.00964	P<0.2734	P<5.94E-07	P<0.0001
Reference	Li *et al*. (2007)[10]	Smith *et al.* (2012)[15]	Smith *et al.* (2013)[23]	Heinzelman *et al.* (2009)[12]	This study	Chang *et al.* (in preparation)[54]

linear regression models for all libraries, except for CBHII, were improved by adding disruption score as a parameter, as evaluated by the F-test ($P < 0.05$). For these libraries, disruption penalty was a more important parameter than any individual block substitution. Disruption penalties were -0.14, -0.29, -0.66, -0.08, and -1.40 °C for P450, Cel48, CBHI, CBHII, and bEGII, respectively.

2.7 A highly thermostable fungal endoglucanase II chimera

These results demonstrate that SCHEMA-guided structure-based recombination can be used to create a thermostable fungal-derived endoglucanase II (Cel5a). The best variant, 110F, has enhanced activity at high temperatures relative to wild-type HjCel5a. This thermostable Cel5a is compatible with other cellulases engineered by this group and others to work efficiently and synergistically at high temperatures[47; 48; 55].

Recent work has shown that HjCel5a is amenable to a variety of stabilization approaches. These include consensus design, core and helix stabilization by computational design, and energy minimization using the FoldX and Rosetta force fields (Lee *et al*., in preparation). Each of these methods was able to find multiple amino acid substitutions, which increased thermostability with an overall success rate of 5-20%. Thermostabilization by recombination performed favorably, finding nine stabilizing mutations, five of which were not found by any other method (F191V, T233V, S242A, V265T, and S322A). G189A and G293A were identified by consensus design[56], whereas D271Y and S318P were identified by FoldX[57]. Building the thermostability model required the construction and evaluation of no more than 45 chimeras, which was comparable to the number of mutations screened for the rational design methods.

Figure 2-6 shows the five novel mutations mapped onto the HjCel5a structure[49]. These mutations were not predicted to be significantly stabilizing by conventional protein design methods[58], so the basis for these mutations is not obvious. F191 is found in an alpha helix, in a hydrophobic pocket adjacent to the beta barrel. It is possible that the mutation to valine results

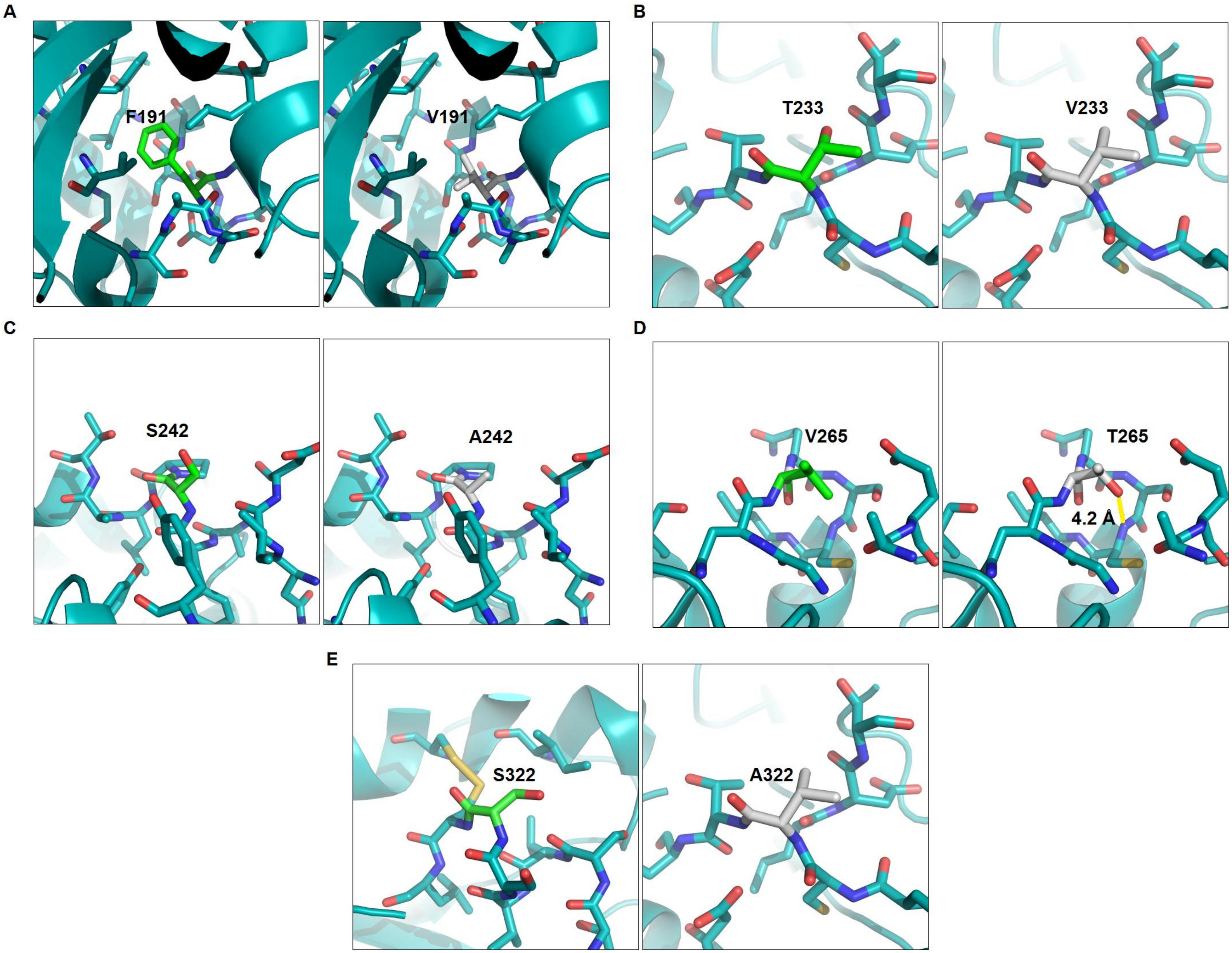

A
F191
V191
B
T233
V233
C
S242
A242
D
V265
T265
4.2 Å
E
S322
A322

Figure 2-6 (Previous page). Thermostabilizing mutations introduced by recombination are conservative. Mutations F191V (**A**), T233V (**B**), S242A (**C**), V265T (**D**), and S322A (**E**) are shown mapped onto the structure of HjCel5a (PDB 3QR3). Residues within 5 A are shown as sticks. Unchanged residues are colored blue, the wild-type residues at mutated sites are colored green, and the mutated residues are colored grey. Oxygen atoms are colored red, and nitrogen atoms are colored blue. F191V shrinks a residue in a hydrophobic pocket. T233V, S242A, and S322A replace small hydrophilic residues on the surface of the protein with small hydrophobic residues. V265T may form a new hydrogen bond at the N-terminus of an alpha helix (shown in yellow).

in more favorable packing. T233 is found in a loop region on the surface of the enzyme and does not appear to form any hydrogen bonds. It is unclear how mutating T233 into V improves stability. S242 is present at the N-terminus of an alpha helix. Mutations at this residue have been implicated in stabilizing the helix dipole[58], but mutation to alanine is unlikely to have this effect. V265 is also present at the N-terminus of an alpha helix, and mutation to threonine may help stabilize the helix by forming a new hydrogen bond to the nitrogen on the backbone of residue 268. Explicit design for helix stabilization did not find this mutation, even though it was one of the most stabilizing mutations found by chimeragenesis (+ 2.0 ºC)[58]. S322 is present in an alpha helix near the C-terminus of the protein. The amino acid is solvent exposed, and does not appear to participate in any hydrogen bonds. It is unclear how mutation to alanine improves thermostability for this residue.

2.8 Non-additive block interactions can constrain engineering by recombination

Compared to previous recombination libraries explored in this laboratory, the Cel5a library had few (two) stabilizing blocks. Without factoring in disruption score, there were eight blocks that were predicted to be thermostabilizing (**Figure 2-4A**). However, when disruption score was added as a parameter, most of these blocks were predicted to be destabilizing or neutral (**Figure 2-4B**). Therefore, in this library, the destabilizing effect of non-favorable amino acid contacts introduced by recombination (-0.91 ºC per contact) significantly compromised thermostability of the library. Indeed, since the average disruption score of chimeras in the library was 12.1, amino acid disruptions reduced average library thermostability by an estimated 11 ºC. This effect was smaller but still present in many other SCHEMA recombination libraries that have been investigated previously, with disruption penalties predicted to range from -0.14 ºC to -1.40 ºC.

Disrupted amino acid contacts are a fundamentally non-additive, or epistatic, phenomenon, since they arise from interactions between recombined blocks. Molecular epistasis has a variety of causes, including stability thresholds, synergism, and suppressor mutations[59], and is a complicating factor for both engineering enzymes[60] and studying evolution[61] because it can lead to highly rugged fitness landscapes. In fact, epistasis has recently been shown to play a major role in molecular evolution[62].

This work highlights the importance of modeling disrupted amino acid contacts for chimeragenesis studies, and that they can account for the majority of non-additive effects seen in a library (R^2 increases from 0.73 to 0.92 when disruption score is taken into account). The possibly large and inhibitory effect of disrupted contacts for thermostability suggests that chimera libraries should be designed to mitigate their effect. **Figure 2-5** shows the mutation level vs. disruption score for different chimera library designs for the Cel5a parents recombined here. The library investigated in this study aimed to have as many mutations as possible, and lies near the top left of the mutation-disruption curve (indicated in red). However, by sacrificing only a few mutations, disruption score could be more than halved, suggesting that libraries that reduce average disruption score while accepting a lower average mutation level may be beneficial.

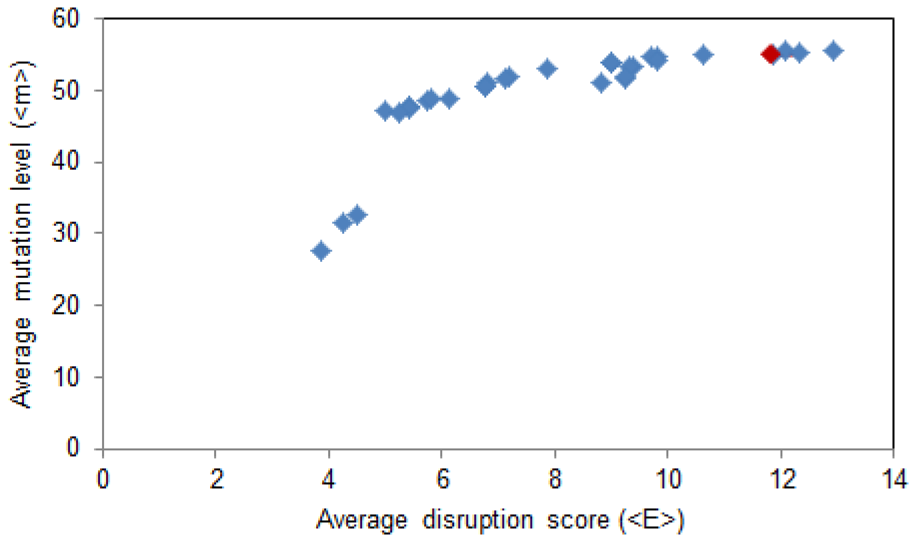


Figure 2-7. Average mutation level and disruption score for Cel5a recombination libraries designed by SCHEMA. The recombination library investigated in this study is shown in red.

2.9 Methods

General methods are described in Appendix 1.

Cel5A Plasmid Construction

Genes encoding *Hypocrea jecorina* Cel5a, *Phialophora sp. G5* Cel5a, *Penicillium decumbens* Cel5a, and *Penicillium pinophilum* Cel5a were synthesized with *S. cerevisiae* codon bias (DNA 2.0, Menlo Park, CA), and cloned into the yeast secretion vector YEp352/PGK91-1-αss as described previously [47; 48]. Each gene had a C-terminal linker and carbohydrate binding module from *H. jecorina* Cel5a. Sequences of all genes are in Appendix 2.

SCHEMA guided structure-based recombination

Gene sequences of *Hypocrea jecorina* Cel5a, *Phialophora sp. G5* Cel5a, *Penicillium decumbens* Cel5a, and *Penicillium pinophilum* Cel5a were aligned using the MUSCLE multiple sequence alignment software[63].The structure of the catalytic domain of *H. jecorina* Cel5a (PDB structure 3QR3 chain A) [49] was used to build a map of amino acid contacts. A contact is defined as two amino acids having at least one non-hydrogen atom within 4.5 Å of each other. Libraries that minimized the average number of SCHEMA contacts in the resulting chimeras were designed using graph partitioning as described[23] (code can be found at http://cheme.che.caltech.edu/groups/fha/software.htm). A library design was chosen with an average SCHEMA energy (number of disrupted contacts) of 12.1 and an average of 55.4 amino acid mutations from the closest parent. The C-terminal linker and carbohydrate binding module from *H. jecorina* Cel5a was appended to each chimera.

Optimal Experimental Design.

We used the Submodular Function Optimization Matlab toolbox[64] to choose chimeras that had both low SCHEMA disruption and maximal mutual information between the sampled chimeras and the rest of the library, as described[16].

Chimera Library Construction

Chimeras were constructed from 500bp DNA fragments via overlap extension PCR, as described previously[65]. The DNA fragments ("gBlocks") were synthesized by Integrated DNA Technologies (San Jose, CA). Codons were optimized for yeast expression using Gene Designer software from DNA 2.0 (Menlo Park, CA)[66]. Genes were cloned into the YEp352/PGK91-1-αss vector using Gibson assembly[67].

Enzyme Purification

YEp352/PGK91-1-αss vectors containing Cel5a chimeras were transformed into the BY4742 Δkre2 strain of yeast (BY4742; Mat a; his3D1; leu2D0; lys2D0; ura3D0; YDR483w::kanMX4) obtained from EUROSCARF (Frankfurt, Germany). Yeast colonies expressing Cel5a with C-terminal His$_6$ tag were grown at 30 °C: first overnight in 5mL SD-Ura medium, then expanded into 50 mL SD-Ura (+50 µg/mL kanamycin) medium for 24 h, and then expanded into 1 L YPD (+50 µg/mL kanamycin) medium for an additional 48 h. Cultures were centrifuged at 4500x*g* for 20 min, and the supernatant was filtered with 0.2 µ m PES filter unit from Nalgene (VWR, Radnor, PA). Protein was loaded onto 5 mL HisTrap columns and purified using an ÄKTAxpress chromatography system (GE Healthcare, Pittsburgh, PA). Purified cellulases were buffered-exchanged to 50mM sodium acetate buffer pH 5.0 using Vivaspin 20 ultrafiltration spin tubes (GE Healthcare, Pittsburgh, PA). Protein concentrations were determined using A280, with theoretical extinction coefficients found using ProtParam on the ExPASy server[68].

T_{A50} Thermostability Measurements

100 µL samples in 50mM sodium acetate buffer, pH 5.0 containing 0.2 µM Cel5a and 1% (w/v) Avicel were incubated at a range of temperatures for 2 h. A modified Park-Johnson reducing sugar assay was used to measure activity [69]; briefly, reaction mixtures were spun at 1000 g for 5 min to remove Avicel. 50 µL of supernatant was removed and transferred to a mixture of 100 µL ferricyanide reagent (0.5 g/L $K_3Fe(CN)_6$, 34.84 g/L K_2HPO_4, pH 10.6) and 50 µL carbonate-cyanide reagent (5.3 g/L Na_2CO_3, 0.65 g/L KCN). The reaction was heated at 95 ºC for 15 min in an Eppendorf Mastercycler, and then cooled on ice for 5 min. 180 µL of the reaction was removed and mixed with 90 µL ferric iron solution (2.5 g/L $FeCl_3$, 10 g/L polyvinyl pyrrolidone, 2 N H_2SO_4). After 2 min, absorbance at 595 nm was taken, using solutions of 0 µM to 300 µM cellobiose as standards.

T_{A50} was determined by plotting activities against the temperature using Matlab (Mathworks, Natick, MA) and fitted using 4-parameter sigmoidal curves. The T_{A50} value is the temperature at which enzyme activity is halfway between optimal activity and no activity. Reported values were averaged from at least two independent measurements.

Cellulase Activity Measurements

All cellulase activity measurements were conducted in 50 mM sodium acetate buffer, pH 5.0. To determine activity-temperature profiles of Cel5A, samples containing 0.2 µM of purified Cel5a and 1% (w/v) Avicel were incubated at 60 and 70 °C for 60 h. After hydrolysis, the reaction supernatants were sampled for reducing sugar concentrations via Nelson–Somogyi assay, using cellobiose as the reducing sugar standard [70; 71]: 50 µL of reaction solution was added to 40 µL carbonate-tartrate solution (144 g/L Na_2SO_4, 12 g/L potassium tartrate tetrahydrate, 24 g/L Na_2CO_3, 16 g/L $NaHCO_3$) and 10 µL copper solution (180 g/L Na_2SO_4, 20 g/L $CuSO_4.5H_2O$) and heated to 95 ºC for 15 min in an Eppendorf Mastercycler. The reaction was placed on ice for 5 min and then mixed with 50 μL arsenomolybdate solution (50 g/L $(NH_4)_2MoO_4$, 1.5 N H_2SO_4, 6 g/L NaH_2AsO_4). After mixing, absorbance at 520 nm was read, using 0 to 2mM cellobiose solutions as standards.

2.10 Supplementary information

Figure S2-1 shows example raw data for thermostability determination. **Figure S2-2** shows the linear regression model without including contact disruption as a parameter. **Table S2-1** shows thermostabilities and contact disruption of chimeras described in this study, as measured by the T_{A50} assay. **Table S2-2** shows thermostabilities of point mutations.

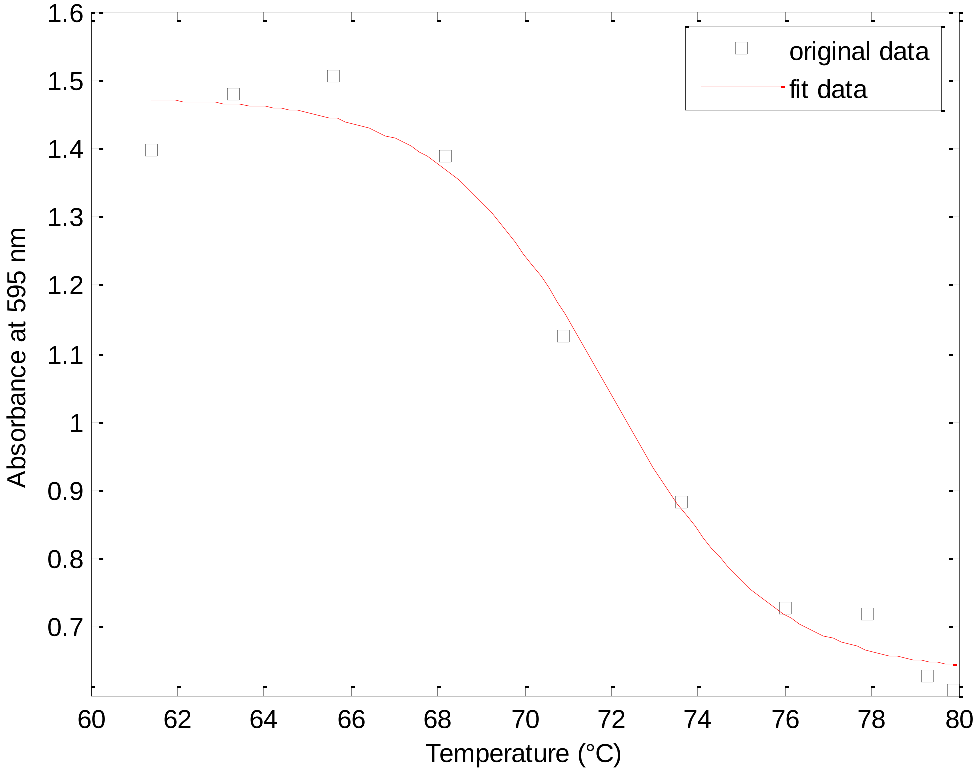


Figure S2-1. Example raw data used to determine T_{A50}. Curve is fit to a Boltzmann four parameter sigmoidal function.

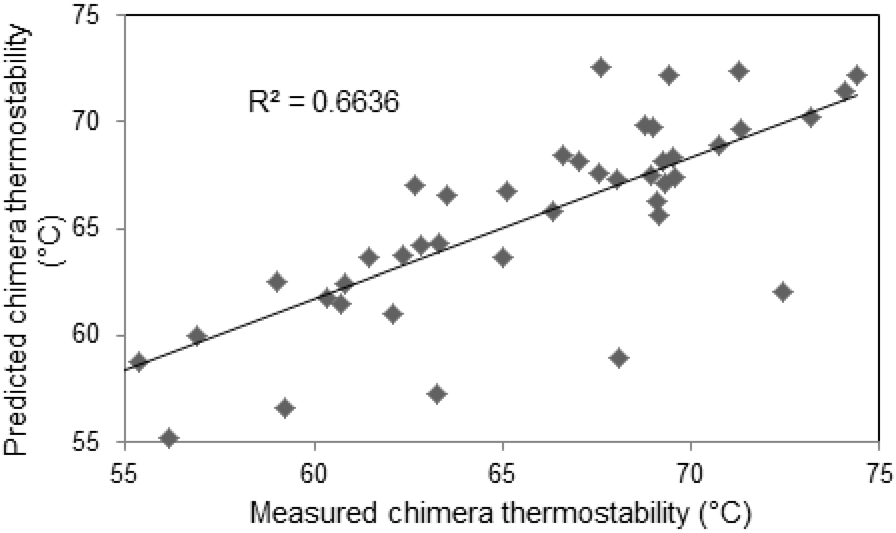


Figure S2-2. Linear regression for chimera thermostability without using SCHEMA disruption as a parameter.

Table S2-1. Thermostabilities and contact disruption scores of Cel5a chimeras. First set is the Cel5a parents, second set is the initial test set, third group is an optimized test set of predicted stabilizing blocks, and forth group are chimeras designed for optimized stability.

Chimera	T_{A50} (°C)	E	Notes
00000000	72.0	0	
11111111	72.4	0	
22222222	63.2	0	
33333333	68.1	0	
00012032	69.1	9	
00031021	54.7	9	
01200030	67.0	7	
03011110	69.6	9	
03110301	62.1	9	
10310232	63.3	9	
11010323	67.5	3	
22030130	59.3	6	
20320310	56.9	8	
23111331	55.4	9	
23121233	62.4	6	
32321133	60.8	9	
33103312	65.1	9	
33113333	69.3	4	
33212131	62.7	9	
33231313	56.2	7	
11311330	54.5	7	
20333123	60.7	8	
10203103	62.8	12	
00130002	48.8	9	
13101033	66.4	9	
31311011	65.0	8	
01003013	73.2	10	
01003213	69.4	14	
01013113	68.8	13	
31013113	67.6	17	
01003113	71.3	11	
00000003	71.3	5	
01013213	69.0	16	
01000000	69.4	2	
00000010	74.1	2	

00003000	68.9	11	Decreased activity
00002000	69.3	3	
03000000	66.6	4	
01000000	69.1	5	
00000013	74.4	6	
12002010	60.3	17	
12002013	59.0	21	
2000000	61.5	8	
13002010	63.5	11	
13002013	68.0	15	
00000100	73.4	0	Decreased activity
00000113	78.9	7	
00000110	75.6	2	

Table S2-1. Thermostabilities of HjCel5a single point mutants. Wild-type T_{A50} is 72.0 ºC.

HjCel5a mutant	ΔT_{A50} (°C)
N153S	-1.10 +/- 0.14
N155T	0.09 +/- 0.27
G189A	0.92 +/- 0.10
F191V	0.89 +/- 0.32
A230T	-4.29 +/- 0.33
T233V	0.87 +/- 0.09
G239D	0.34 +/- 0.06
S242A	0.58 +/- 0.26
V265T	2.01 +/- 0.06
Q266A	0.17 +/- 0.44
I269E	-3.18 +/- 0.44
Q270T	-0.44 +/- 0.43
D271Y	2.60 +/- 0.41
M272L	-1.54 +/- 0.28
V302A	-0.99 +/- 0.10
T304D	-1.12 +/- 0.18
S318P	1.89 +/- 0.26
S322A	0.58 +/- 0.13

Chapter 3

A SYNERGISTIC SET OF ENGINEERED THERMOSTABLE FUNGAL CELLULASES ACCELERATES HIGH-TEMPERATURE CELLULOSE HYDROLYSIS

3.1 Abstract

A major obstacle to the widespread use of cellulose as a source of renewable fuels and chemicals is the difficulty in converting cellulose into soluble sugars for fermentation, as the cellulases used to catalyze cellulose hydrolysis are slow and expensive. One possible solution is to engineer cellulases that are more thermostable, allowing higher activity at higher reaction temperatures. We have previously combined directed evolution, rational design, and structure-based recombination to engineer thermostable fungal cellobiohydrolases Cel6a and Cel7a. Here we describe the creation of the most stable known fungal endoglucanase, a derivative of *Hypocrea jecorina* (anamorph *Trichoderma reesei*) Cel5a, by combining mutations isolated from chimera studies, consensus design, and other computational methods. The engineered endoglucanase is 17 °C more thermostable than *H. jecorina* Cel5a and hydrolyzes 50% more cellulose over 60 h at its optimum temperature. A set of thermostabilized cellulases (Cel5a, Cel6a, Cel7a) synergistically hydrolyzes cellulose at an optimum performance temperature of 70 °C, with total sugar production three times greater than the of wild-type enzymes at their optimum temperature of 60 °C, over 60 h incubations.

3.2 The utility of thermostable cellulase mixtures

Cellulases engineered for increased thermostability can reduce lignocellulose biomass degradation times and costs, facilitating the use of this feedstock for biofuels and specialty chemicals[72]. Thermostable cellulases can have increased cellulolytic activity at higher temperatures and remain active for longer at these temperatures[73; 74]. Moreover, biomass degradation at elevated temperatures reduces cooling costs following pre-treatment and reduces the risk of microbial contamination[72].

Effective cellulose degradation requires four cellulase activities: cellobiohydrolases I and II processively hydrolyze opposite ends (reducing and non-reducing, respectively) of the cellulose chain, endoglucanases cleave intrachain bonds, and beta-glucosidases break down cellobiose molecules released by other cellulases[75]. Over the last four years, this lab has engineered thermostable class I cellobiohydrolases (Cel7a)[44; 47] and class II cellobiohydrolases (Cel6a)[12; 48] using a combination of SCHEMA recombination, rational design, and directed evolution. As reported by Wu and Arnold[48], combining thermostabilized Cel6a and Cel7a increases the amount of released cellobiose by approximately 80% over a 60 h incubation, relative to the wild-type Cel6a and Cel7a mixture, when each mixture operates at its optimum temperature (70 °C for engineered and 60 °C for wild-type). Since fungal beta-glucosidases with optimum temperatures above 70 °C are already known[76], the final step to creating a thermostable cellulolytic enzyme mixture was to engineer a thermostable fungal endoglucanase that retains high catalytic activity.

The wild-type class II endoglucanase Cel5a accounts for up to 12% of the total secreted cellulase and 55% of the endoglucanase activity in the industrial fungal strain *Hypocrea jecorina* (anamorph *Trichoderma reesei*)[77; 78; 79]. HjCel5a has an optimum activity at 64 °C measured over 2 h incubations, and exhibits significantly decreased activity at 70 °C, making it incompatible with a thermostable cellulase mixture and a prime target for protein engineering to increase thermostability.

3.3 Engineering the most stable known fungal endoglucanase

To create a thermostable HjCel5a we combined stabilizing mutations identified from homologous recombination (Chapter 2) and various computational approaches[58]. In Chapter 2 we reported the creation of an HjCel5a (called 110F) with an optimal temperature of 74 °C. Lee *et al*. reported the creation of an HjCel5a (called s13pt4) with an optimal temperature between 75 and 78 °C using various computational methods[58]. Here we combined all the thermostabilizing mutations from these studies that did not compromise activity. When two suitable mutations were at the same site, we chose the more thermostabilizing of the two.

The mutations identified by chimeragenesis were F191V, T233V, and V265T. Thermostabilizing mutations E53D, T57N, S79P, T80E, V101I, S133R, N155E, G189S, G239E, G293A, and S309W were described by Lee *et al*.[58]. D271Y and S318P were identified in both studies. A list of the combined mutations is shown in **Table 3-1**, and their locations in HjCel5a are shown in **Figure 3-1**.

The resulting HjCel5a variant (OptCel5a) has an optimal temperature of 81.1 °C when used to hydrolyze crystalline cellulose (Avicel) for 2 h (**Figure 3-2A**). This makes OptCel5a more than 17 °C more thermostable than wild-type HjCel5a, more than 7 °C more stable than the 110F variant, and 3 °C more stable than the s13pt4 variant. It is the most stable fungal endoglucanase reported.

To investigate the long-term activity of OptCel5a, we tested its activity over 60 h, at both 60 °C and 70 °C and in comparison with HjCel5a. OptCel5a had highest activity at 70 °C, hydrolyzing over 50% more cellulose than HjCel5a at its optimal temperature of 60 °C (**Figure 3-2B**). OptCel5a is therefore compatible with the previously engineered thermostable Cel6a and Cel7a, which both have an optimum 60 h temperature at 70 °C[48].

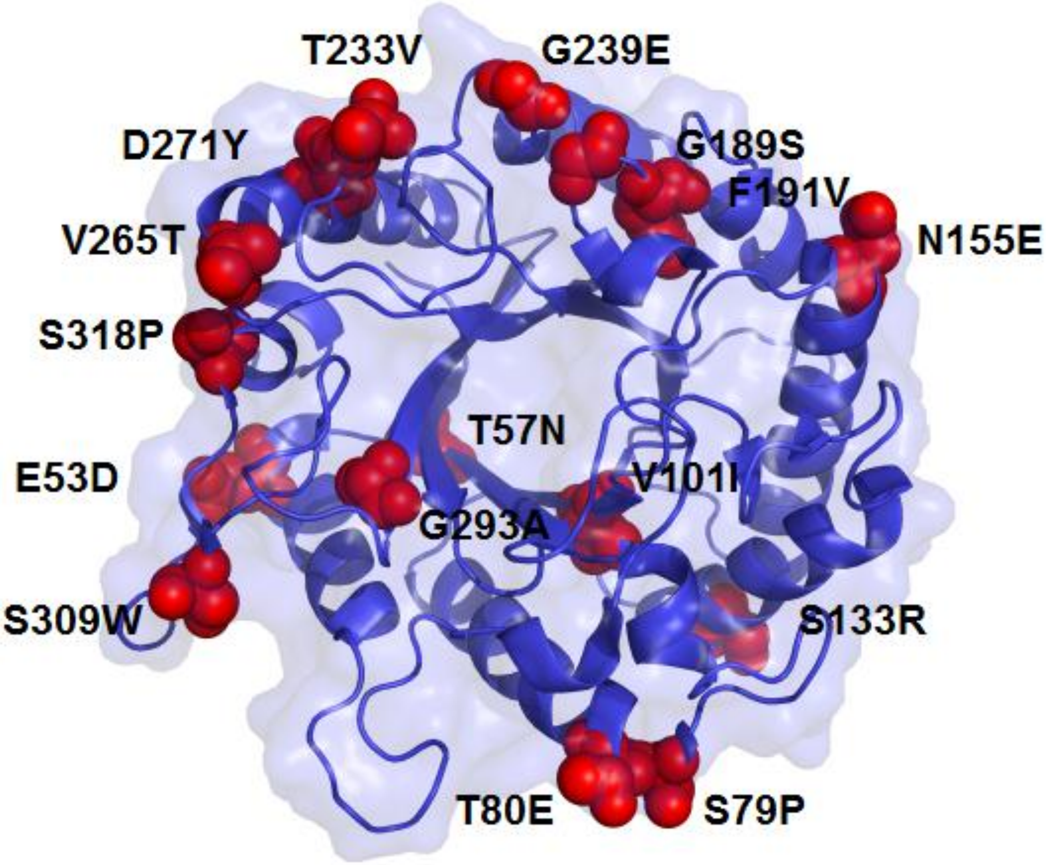


Figure 3-1. Location of stabilizing mutations on the HjCel5a crystal structure. Mutations are E53D, T57N, S79P, T80E, V101I, S133R, N155E, G189S, F191V, T233V, G239E, V265T, D271Y, G293A, S309W, and S318P.

Table 3-1. Stabilizing mutations combined to create OptCel5a.

Mutation	Thermostability increase (ºC)	Stabilization method	Source
F191V	0.89	Chimeragenesis	Chapter 2
T233V	0.87	Chimeragenesis	
V265T	2.01	Chimeragenesis	
S318P	3.43	FoldX/Chimeragenesis	Chapter 2
D271Y	2.67	FoldX/Chimeragenesis	Lee *et al.*[58]
S79P	0.29	FoldX	Lee *et al.*[58]
E53D	2.72	Consensus	Lee *et al.*[58]
T57N	1.12	Consensus	
G293A	3.58	Consensus	
V101I	0.12	Core stabilization	Lee *et al.*[58]
N155E	0.54	Helix dipole stabilization	Lee *et al.*[58]
T80E	0.50	Helix dipole stabilization	
S133R	0.44	Helix dipole stabilization	
G239E	0.24	Helix dipole stabilization	
S309W	0.35	Triad ddG	Lee *et al.*[58]
G189S	0.94	Backbone entropy reduction	Lee *et al.*[58]

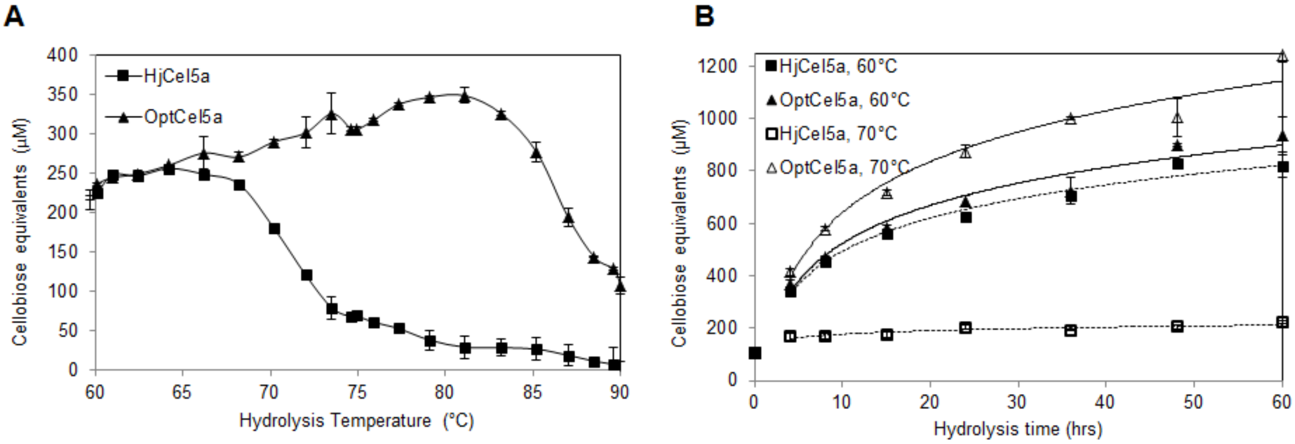


Figure 3-2. A highly stable engineered Cel5a endoglucanase. A) Total cellobiose equivalents released after 2 h Avicel hydrolysis with HjCel5a and OptCel5a. B) Total cellobiose equivalents released after 60 h Avicel hydrolysis at 60 °C and 70 °C with HjCel5a and OptCel5a. Loading was 0.2 µM enzyme and 1% Avicel.

3.4 Evaluating the synergy of engineered thermostable cellulases

It has been known for the last 40 years that endoglucanases and cellobiohydrolases act synergistically to degrade cellulose[75; 80; 81]. We explored the synergy between cellobiohydrolases Cel6a, Cel7a, and endoglucanase Cel5a by comparing mixtures of the wild-type enzymes with engineered-thermostable cellulase mixtures. The engineered-thermostable mixture consists of OptCel5a as the Cel5a variant, 3C6P as the Cel6A[48], and TS8 as the Cel7A[47]. Each of these enzymes has an optimal activity at or greater than 70 °C when measured over 60 h incubations. As our wild-type mixture, we used Cel5a from *H. jecorina*, Cel6A from *H. insolens*, and Cel7A from *T. emersonii*, which are the most thermostable known homologues of each enzyme. These enzymes exhibit an optimal activity of 60 °C over 60 h[48].

In these experiments the total cellulase concentration was fixed at 0.5 µM, and the relative concentrations of each cellulase were varied in steps of 0.1 µM, allowing a ternary synergy diagram to be constructed[81]. Reactions were carried out on Avicel over 60 h at 60 °C for wild-type and 70 °C for engineered enzymes. These conditions were chosen to be consistent with previous synergy studies[48; 81; 82] and industrial conditions of high temperatures and incubation times. As shown in **Figures 3-2A** and **B**, both enzyme mixtures exhibited substantial synergy, with the mixtures more active than any of the enzymes alone. The degree of synergy, obtained by dividing the activity of the mixture by the sum of the activities of the individual cellulases, ranged from 1.0 to 1.6 for the wild-type enzymes, and from 1.0 to 2.1 for the engineered enzymes[79].

In both wild-type and engineered mixtures the highest cellulose hydrolysis activity occurred with relatively small amounts of endoglucanase (10-20% of total mixture), which has been observed in other synergy studies[83]. This small amount of endoglucanase required for an optimal mixture can be explained by the fact that the role of endoglucanase is to produce free ends that can be targets for cellobiohydrolases. Cellobiohydrolases processively hydrolyze along these ends, and are responsible for the bulk of hydrolysis.

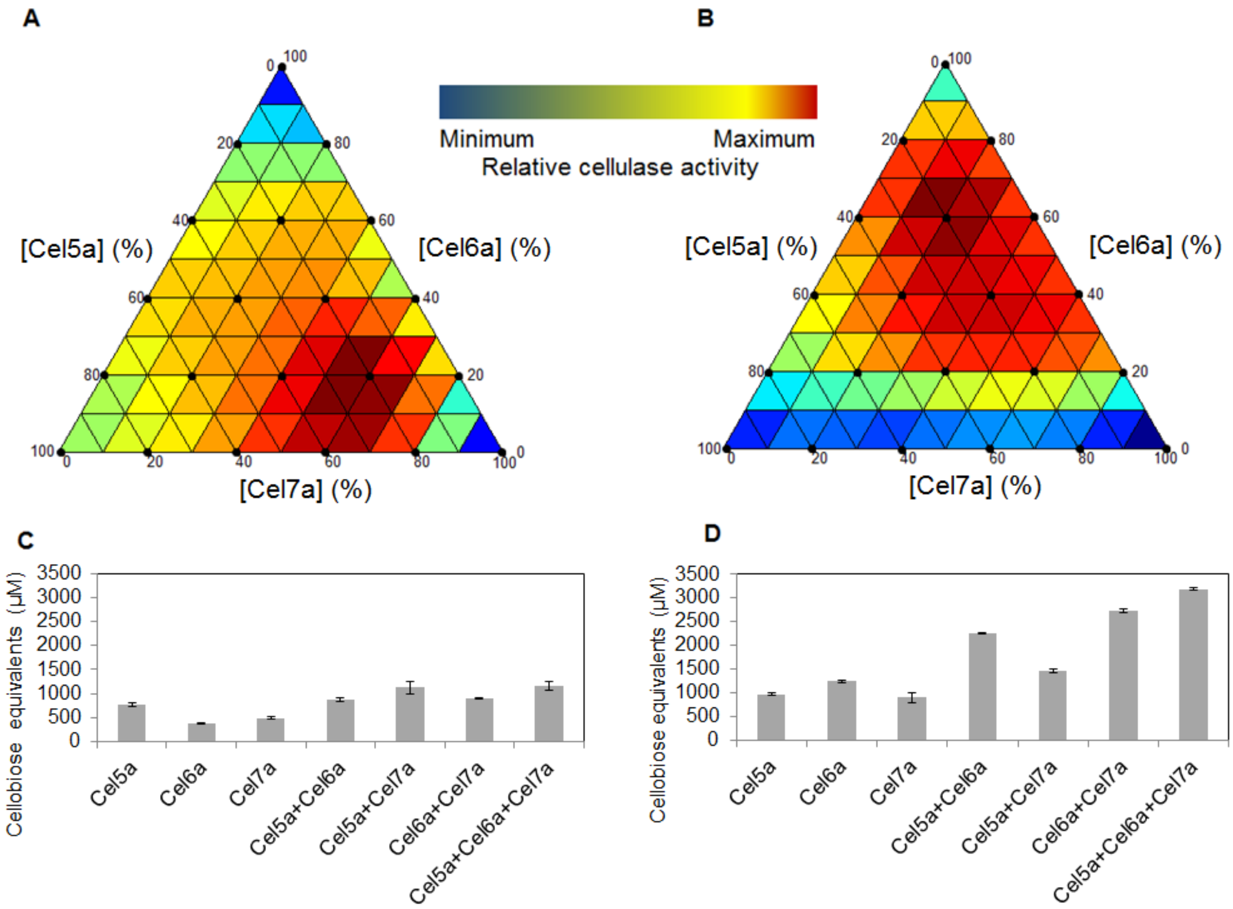


Figure 3-3. Synergistic cellulose hydrolysis by wild-type (A) and engineered-thermostable (B) Cel5a, Cel6A, and Cel7A. A total concentration of 0.5 µM of cellulase and 1% w/v Avicel was used. Each edge indicates the concentration of the labeled cellulase, which ranges from 0% to 100% of the total mixture. Each vertex represents 100% concentration of an individual cellulase, each edge represents a mixture of two cellulases, and the interior of the triangle is a mixture of all three cellulases. Black dots are individual measurements (in duplicate), and colors are arithmetic averages between each point, with red representing maximum activity and blue representing minimum activity. Colors are normalized for each synergism test. The absolute activities of the individual enzymes as well, as the best mixtures for double and triple enzyme combinations, are shown for wild-type (**C**) and engineered thermostable (**D**).

The mixture with highest hydrolysis activity shifted from predominantly Cel7a for wild-type mixtures to predominantly Cel6a concentrations for engineered mixtures. As shown in **Figure 3-2 C** and **D**, this change in optimal enzyme loadings reflected the relative activities of Cel6a and Cel7a in the wild-type and engineered cases [48]. **Figures 3-2 C** and **D** also show the activities of the optimal cellulase mixtures for two and three enzymes. The best mixture of wild-type enzymes in this experiment was over 1.5X better than any of its constituent enzymes, while the optimal mixture of engineered thermostable enzymes was over 2.5X better. The optimal engineered thermostable mixture was also 1.16X better than a mixture containing only engineered Cel6a and Cel7a, with an equal total enzyme concentration.

3.5 An optimized mixture of engineered cellulases accelerates cellulose hydrolysis

We searched the region of maximum activity more closely in steps of 0.04 µM and found the optimal mixture for wild-type to be 0.16: 0.28: 0.56 Cel5a:Cel6a:Cel7a. The optimal engineered thermostable mixture is 0.08:0.56:0.36 Cel5a:Cel6a:Cel7a. We call the optimized engineered thermostable mixture T-PRIMED. We evaluated the activity of T-PRIMED over 60 h at both 60 °C and 70 °C and compared it to the activity of the best wild-type mixture. We ran this assay on 1%, 3%, and 5% Avicel to see the effects of varying cellulose concentrations (**Figure 3-3A,B,C**). T-PRIMED has the highest activity at 70 °C, where it is approximately three times more active than the best mixture of wild-type enzymes at 60 °C. The activity of all cellulase mixtures increased at higher cellulose concentrations, with the activity ratio remaining approximately constant.

We also tested the activity of the mixtures on two industrially relevant lignocellulose substrates: milled corn stover and dilute acid-treated rice straw (**Figure 3-3D**). T-PRIMED had higher activity than wild-type on both substrates, with 1.8X the activity on milled corn stover, and 2.5X the activity on treated rice straw.

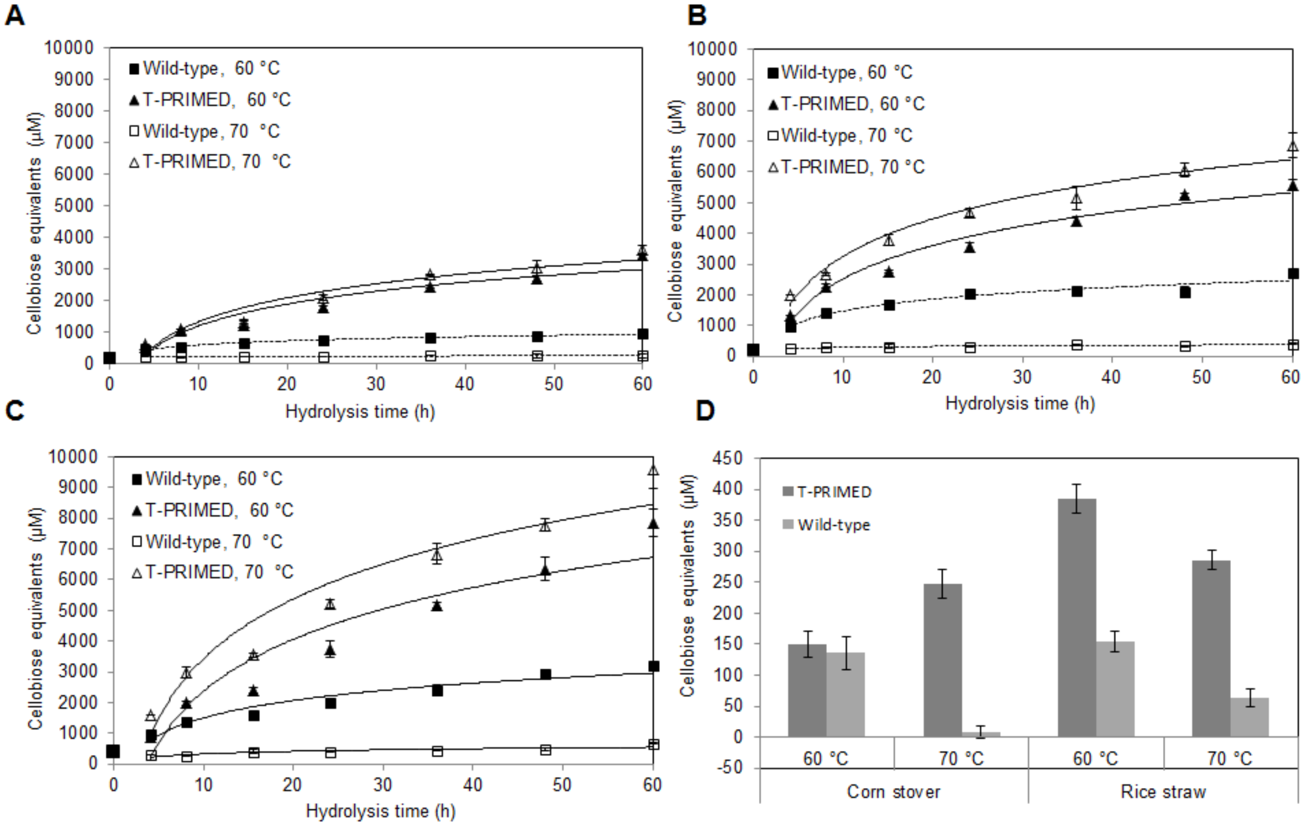


Figure 3-4. Total cellobiose equivalents released during 60 h hydrolysis with wild-type and engineered-thermostable cellulase mixtures at 60 °C and 70 °C, on both 1% (**A**), 3% (**B**), and 5% (**C**) w/v Avicel; and after 60 h hydrolysis on both milled corn stover and dilute-acid treated rice straw (**D**). The wild-type mixture is 0.16:0.28:0.56 Cel5a:Cel6a:Cel7a, and the engineered thermostable mixture is 0.08:0.56:0.36 Cel5a:Cel6a:Cel7a, with a total concentration of 0.5 µM, as described in text.

3.6 Discussion

We report here the engineering of the most stable reported HjCel5a variant, OptCel5a, and the characterization of its synergy with other engineered thermostable cellulases. This enzyme has an optimal temperature (over 2 h incubations) of 81 ºC, and releases over 1.5X more soluble sugar over 60 h incubations compared to wild-type Cel5a from *Hypocrea jecorina.*

OptCel5a works synergistically with previously reported engineered thermostable cellobiohydrolases I and II[48]. T-PRIMED, an optimized mixture of these enzymes, releases over 3X more soluble sugar over 60 h incubations on crystalline cellulose (Avicel) compared to a similarly optimized wild-type mixture. T-PRIMED is also more active on model cellulose substrates derived from corn stover and rice straw.

The synergy studies presented here on engineered thermostable fungal cellulases extend the results from synergy studies on wild-type fungal cellulases[79; 81; 84; 85; 86; 87; 88; 89; 90]. An optimal mixture of fungal cellulases requires at least three different cellulase activities: endoglucanase, cellobiohydrolase I, and cellobiohydrolase II, and this holds true for both engineered and wild-type mixtures. The engineered endoglucanase reported here increases the activity of a previously reported mixture of engineered cellobiohydrolases[48] by 16 %. The degree of synergism also increases from a maximum of 1.6 for wild-type cellulases to a maximum of 2.1 for engineered thermostable cellulases. These synergy values are typical for reaction of fungal cellulase mixtures on Avicel, which range from 1.3 to 2.2 for *Hypocrea jecorina* cellulases[79]. Although these data suggests a temperature dependent effect on synergy, wild-type mixtures assessed for cellulase activity at 50 ºC, 60 ºC, and 70 ºC have similar synergy values (**Supplementary Figure 3-1**).

In this proof of principle study, we limited our investigation to the synergy of engineered thermostable cellulases on Avicel. Avicel is known to have lower degree of polymerization (DP) than other cellulosic substrates like wood pulp and higher DP than substrates like phosphoric acid swollen cellulose (PASC)[79]. Cellobiohydrolases are more active on substrates with lower DP, due to a higher proportion of chain ends, and therefore the optimum mixture of cellulases will change depending on substrate. Moreover, the results here show that the optimum temperature can change based on substrate as well: T-PRIMED displays an optimum temperature of 70 °C displayed on Avicel and corn stover, while on treated rice-straw it has an optimal activity of 60 °C. This change in optimum temperature reflects the fact that the thermostability of cellulase mixtures can depend on binding to cellulose[74], and binding is likely to change based on composition of the substrate[79].

Synergy is also expected to decrease with hydrolysis time[90] and enzyme loadings[84], two properties that were not investigated here. These results together imply that engineered cellulase mixtures will likely need to be optimized for particular applications. High-throughput approaches for optimizing cellulase mixtures, such as robotic platforms[91] and computationally guided approaches[92], will likely be required.

In summary, we have combined the results of multiple protein engineering efforts to 1) create the most thermostable fungal endoglucanase reported, 2) create the most thermostable set of synergistically-acting cellulases reported to date, and 3) demonstrate an approximately three-fold enhancement in hydrolysis activity on crystalline cellulose for this set compared to a set of wild-type fungal enzymes. Our study demonstrates two important considerations for engineering systems of cellulolytic enzymes. When enzymes work cooperatively, it is necessary to engineer all key components of the system to attain the highest possible improvement. The relative importance of enzymes in these systems can also change, and synergy experiments such as those carried out in this study should be used to find the optimum enzyme mixture.

3.7 Methods

General methods are described in Appendix 1.

Cel5A Plasmid Construction

Genes encoding Cel6A, and Cel7A were cloned into the yeast secretion vector YEp352/PGK91-1-αss as described previously[47; 48]. The gene encoding wild-type Cel5A gene (including its cellulose binding module) was synthesized with *S. cerevisiae* codon optimization (DNA 2.0, Menlo Park, CA). Sequences of all genes are in Supplementary Information.

Enzyme Purification

Yeast colonies expressing Cel5a and Cel6A with C-terminal His_6 tags and Cel7A with an N-terminal His_8 tag were grown at 30 °C: first overnight in 5 mL SD-Ura medium, then expanded into 50 mL SD-Ura (+50 µg/mL kanamycin) medium for 24 h, and then expanded into 1 L YPD (+50 µg/mL) medium for an additional 48 h. Cultures were centrifuged at 4500 *g* for 20 min, and the supernatant was filtered with 0.2 mm PES filter unit from Nalgene (VWR, Radnor, PA). Protein was purified using 5mL HisTrap columns (GE Healthcare, Pittsburgh, PA). Purified cellulases were buffered-exchanged to 50mM sodium acetate buffer pH 5.0 using Vivaspin 20 ultrafiltration spin tubes (GE Healthcare, Pittsburgh, PA). Protein concentrations were determined using A280, with theoretical extinction coefficients found using ProtParam on the ExPASy server[68].

Thermostability Measurements

100 µL samples in 50 mM sodium acetate buffer, pH 5.0 containing 0.2 µM Cel5a and 1% (w/v) Avicel were incubated at a range of temperatures for 2 h in an Eppendorf Mastercycler (Hamburg, Germany). A modified Park-Johnson reducing sugar assay was used to measure activity[69], as described in Chapter 2.

Cellulase Activity Measurements

All cellulase activity measurements were conducted in 50 mM sodium acetate buffer, pH 5.0. Constant temperature was maintained using an Eppendorf Mastercycler (Hamburg, Germany). To determine activity-temperature profiles of Cel5A, samples containing 0.2 µM of purified Cel5a and 1% (w/v) Avicel were incubated at 60 and 70 °C for 60 h. To determine the activity of the Cel5a, Cel6A, and Cel7A mixtures, purified Cel5a, Cel6A, and Cel7A were combined at different ratios to a final concentration of 0.5 µM along, with 1% Avicel in 100 µL and incubated 60 °C and 70 °C for 60 h. After hydrolysis, reaction supernatants were sampled for reducing sugar concentrations via a modified Nelson–Somogyi assay[70; 71], as described in Chapter 2.

Cellulose hydrolysis activity over time to determine optimized engineered and wild-type cellulase mixtures was carried out on 1% and 3% Avicel at 60 °C and 70 °C. Time points were taken at 0 h, 4 h, 8 h, 15 h, 24 h, 36 h, 48 h, and 60 h. Reducing sugar concentration was quantified as above.

Pre-treated lignocellulose from corn stover was obtained as a gift from Alex Nisthal. Pre-treated lignocellulose from rice straw was obtained as a gift from Frank C. J. Chang and prepared according to Hsu *et al.* (2010)[93]. Activity assays were carried out for optimized engineered and wild-type cellulase mixtures on 3% substrate at 60 °C and 70 °C for 60 h. Released cellobiose at the end of 60 h was quantified as above.

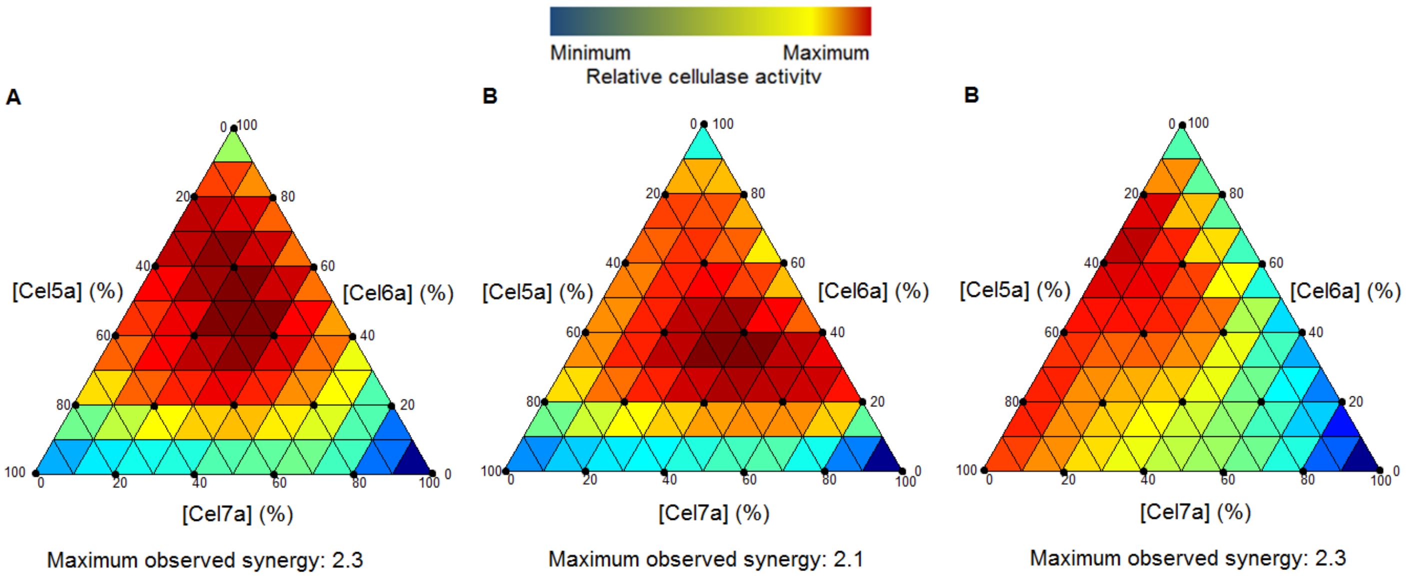
Data analysis

Cellulase activity and thermostability data were plotted using Microsoft Excel (Redmond, WA). Synergy plots were made in Matlab (The Mathworks, Inc., Natick, MA), using the Ternplot package developed by Carl Sandrock.

(http://www.mathworks.com/matlabcentral/fileexchange/2299-ternplot).

3.8 Supplementary Information

Figure S3-1 shows the synergistic activity of wild-type cellulase mixtures incubated at 50 ºC, 60 ºC, and 70 ºC for 24 h on Avicel. The mixtures had similar synergy values (maximum observed synergy of ~2) at these temperatures.



Maximum observed synergy: 2.3

Maximum observed synergy: 2.1

Maximum observed synergy: 2.3

Figure S3-1. Synergistic activity of wild-type cellulase mixtures at 50 °C, 60 °C, and 70 °C. A total concentration of 0.5 µM of cellulase and 1% w/v Avicel was used. Mixtures were incubated at 50 °C (**A**), 60 °C (**B**), and 70 °C (**C**) for 24 h. Each edge indicates the concentration of the labeled cellulase, which ranges from 0% to 100% of the total mixture. Each vertex represents 100% concentration of an individual cellulase, each edge represents a mixture of two cellulases, and the interior of the triangle is a mixture of all three cellulases. Black dots are individual measurements of released soluble sugar (in duplicate), and colors are arithmetic averages between each point, with red representing maximum activity and blue representing minimum activity. Colors are normalized for each synergism test. Maximum observed synergy values are displayed below each plot.

Chapter 4

DIRECTED EVOLUTION OF AN ENOLASE FOR NEXT-GENERATION BIOFUELS

4.1 Abstract

Enzymes often use complex and expensive cofactors to perform both essential cellular reactions and industrially important chemical transformations. The creation of enzymes with simpler and more efficient cofactors would be significant for understanding how nature has optimized enzyme cofactor choice and for improving cost-effective production of fuels and chemicals. In this study we investigated the dehydration of R-2,3-dihydroxyisovalerate (2R-DHIV) into 2-ketoisovalerate (KIV), a reaction involved in branched-chain amino acid biosynthesis as well as industrial isobutanol biosynthesis. In nature this reaction is catalyzed by dihydroxyacid dehydratase (DHAD), an enzyme which contains an oxygen-sensitive iron-sulfur cluster cofactor whose maturation requires a complicated and energetically expensive biosynthetic process. In contrast, members of the enolase family of enzymes catalyze the dehydration of similar substrates, using only a simple magnesium ion cofactor.

In this study we aimed to engineer an enolase (L-rhamnonate dehydratase, YfaW) to replace DHAD, thereby replacing a complicated enzyme by a simpler one. Since this new dehydratase would be predicted to be insensitive to oxygen, as well as cheaper to produce biosynthetically, it could be used to promote cost-effective production of next-generation biofuels such as isobutanol and other higher alcohols. We applied structure-guided saturation mutagenesis, directed evolution, substrate walking, and rational design, screening over 20,000 protein variants and selecting over 10^7 protein variants. We were unable to find any variants that could dehydrate 2R-DHIV. These results

provide an example of the limitations of protein engineering to alter reactivity. We speculate that cofactor choice may be optimal for some natural enzymatic reactions.

4.2 Natural enzymes limit production of biofuels and biochemical

A major obstacle to the widespread adoption of biosynthetic strategies for production of fuels and chemicals as sustainable, affordable, and environmentally-friendly replacements to existing industrial processes is the low production yield from microorganisms. One reason for this is that the natural enzymes used to synthesize biofuels and chemicals have not been optimized for this task, having insufficient reaction rates and inappropriate reaction conditions.

Protein engineering by methods like directed evolution and recombination has made significant progress in improving the stability[12; 44; 48], activity[94], and substrate specificity[95] of natural enzymes as a step towards introducing them into industrial processes. These efforts have focused on engineering the primary sequence of enzymes; however, many enzymes have highly reactive cofactors that dictate their catalytic behavior and properties. For example, many complex metal cofactors are oxygen sensitive and expensive to biosynthesize (such as the iron molybdenum cofactor in nitrogenases[96], or the iron-sulfur cofactor in dehydratases[97]), which can limit the industrial utility of these enzymes as well as inhibit engineering by directed evolution.

Can cofactor choice in natural enzymes be engineered? Recently this lab used directed evolution to switch the cofactor preference of ketol-acid reductoisomerase (KARI) from NADPH to the structurally similar NADH[98; 99]. This cofactor switch allowed KARIs to operate efficiently under anaerobic conditions where NADPH is limiting.

In isobutanol production starting from sugar[100], the limiting step is the dehydration reaction catalyzed by the iron-sulfur (Fe-S) cofactor dependent dehydratase, dihydroxy acid dehydratase (DHAD), which converts 2R-dihydroxy-isovalerate (2R-DHIV) into ketoisovalerate (KIV)[97]. Fe-S biogenesis and integration into proteins is complex and energy-intensive, and the resulting enzyme is difficult to express at high levels; yet, only this class of enzymes is known to catalyze 2R-DHIV dehydration across all organisms.

In contrast, members of the enolase superfamily of enzymes can catalyze related dehydration reactions using only the simple cofactor Mg^{2+} [101]. Therefore, it may be that Fe-S dependent DHAD could be replaced by a simpler Mg^{2+}-dependent dehydratase, which could improve active enzyme concentration and reaction rate. One such dehydratase that we believed was a promising template for engineering is YfaW from *E. coli*, which catalyzes dehydration of the sugar-acid L-rhamnonate[102]. L-rhamnonate has a similar structure to 2R-DHIV, with the major difference being a longer sugar backbone (**Figure 4-1**).

To create an Mg^{2+}-dependent DHAD, we applied a combination of rational protein design and directed evolution to attempt to change the substrate specificity of YfaW. Despite extensive screening, we were unable to find any variants with activity on 2R-DHIV. These results suggest that it may be difficult to replace complex enzymes with simpler enzymes in biological pathways.

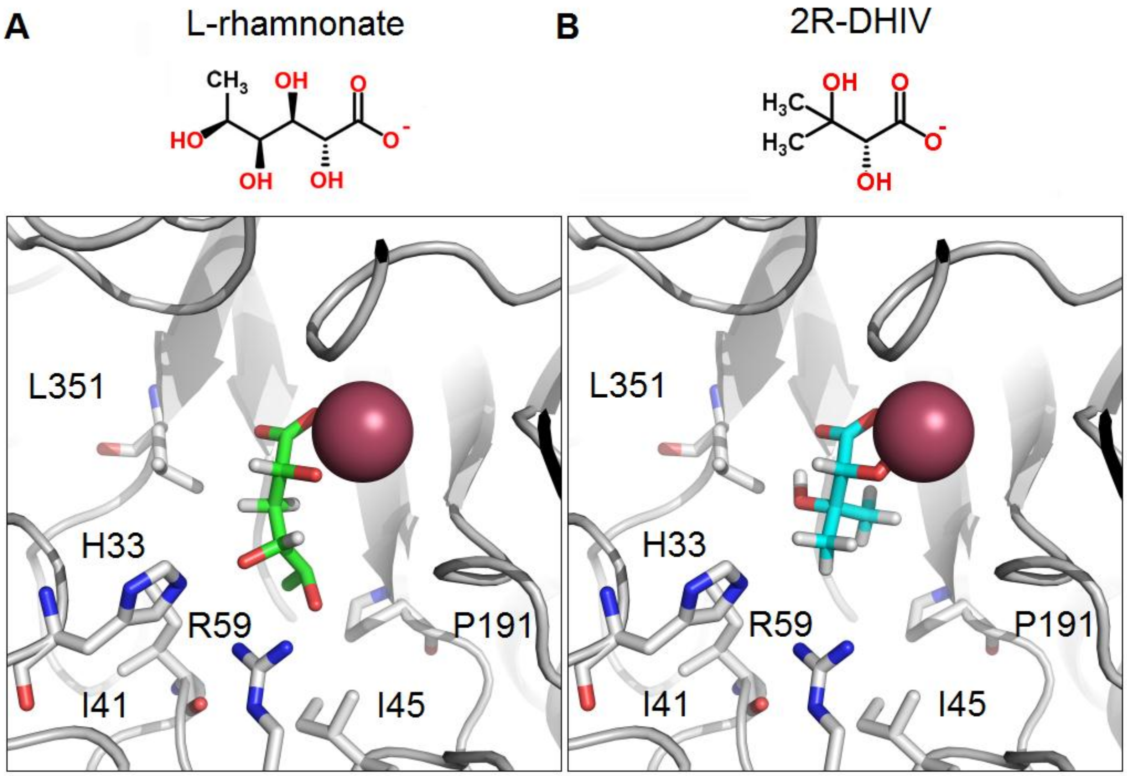


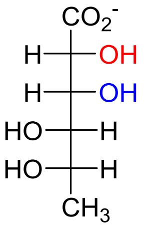
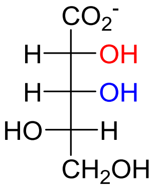
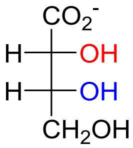
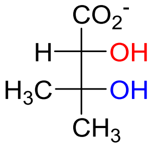
Figure 4-1: Proposed substrate engineering for EcYfaW. A) The natural substrate of EcYfaw, L-rhamnonate. A substrate analogue 3-deoxy-L-rhamnonate is shown in the YfaW binding pocket (from *Salmonella typhimurium* YfaW x-ray crystal structure, PDB 3CXO[102]). **B)** The target substrate for EcYfaW engineering, 2R-DHIV. The molecule is modeled in the YfaW binding pocket by comparison with 3-deoxy-L-rhamnonate. Amino acid residues that were targets for mutagenesis are shown as sticks and labeled, with non-carbon heavy chain atoms colored. The essential magnesium ion is colored red.

4.3 Assessment of enolase candidates for engineering

An initial low level of activity is often desired when choosing starting proteins for directed evolution[103], so we first assessed whether there existed Mg^{2+}-dependent enolases that could catalyze the dehydration of 2R-DHIV. We cloned four enolases from *E. coli*, YfaW[102], GlucD[104], GalD[101], and ManD[105], which dehydrate L-rhamnonate, D-glucarate, D-galaconate, and D-mannonate, respectively. These enzymes were expressed as C-terminal His_6-tagged enzymes in *E. coli* BL21(DE3) ΔIlvD, a strain with the native *E. coli* DHAD gene, IlvD, deleted in order to eliminate background 2R-DHIV dehydratase activity. The enzymes were purified by affinity chromatography and their activities on 2R-DHIV were evaluated by semicarbazide derivatization followed by HPLC. None of these variants displayed detectable activity on 2R-DHIV.

E. coli YfaW (EcYfaW) is reported to have activity on the two substrates most similar to 2R-DHIV, L-rhamnonate and L-lyxonate[102]. In particular, these two substrates have similar stereocenters at carbons 2 and 3 (**Table 4-1**). This substrate similarity suggests that EcYfaW may be a viable candidate for engineering 2R-DHIV dehydration activity. We first verified that EcYfaW had activity on L-rhamnonate and ~5 fold less activity on L-lyxonate, as reported by Rakus *et al*[102]. It also had no activity on D-erythronate, a substrate intermediate between L-lyxonate and 2R-DHIV. We also tested the thermostability of EcYfaW, and found that the enzyme is stable up to 54 ºC (as measured by finding the residual activity after 10 min incubations).

Table 4-1: EcYfaW biochemical parameters. Sugar acids are displayed as Fischer projections, and the two stereochemically similar hydroxyl groups between L-rhamnonate and 2R-DHIV are colored red (C2) and blue (C3), as described in text.

	k_{cat} (s^{-1})	K_m (mM)	k_{cat}/K_m ($M^{-1}s^{-1}$)	Structure	Reference
L-rhamnonate	3.2+/-0.2	0.15+/-0.07	2.1x10^4	 CO_2^- H OH H OH HO H HO H CH_3	Rakus *et al.*[102]
L-lyxonate	0.6+/-0.03	2.0+/-0.3	3*10^2		Rakus *et al.*[102]
D-erythronate	no activity	no activity	no activity		This study
2R-DHIV	no activity	no activity	no activity		This study

4.4 Directed evolution of an enolase- Targeted mutagenesis

2R-DHIV is identical to L-rhamnonate at the dihydroxy acid moiety encompassing carbons 1 through 3, but differs at the distal end of the molecule; it does not have a carbon 5 or 6, does not have hydroxyl group at carbon 4, and has an extra methyl group at carbon 3. The amino acid residues of EcYfaW that form the substrate binding pocket and are adjacent to these distal changes are labeled in **Figure 4-1.** We targeted H33, I41, I45, R59, P191, and L351 sites for saturation mutagenesis.

We constructed individual NNK libraries for each of these sites, and expressed mutants in *E. coli* BL21(DE3) ΔIlvD. We screened 90 variants from each library, using a medium throughput assay involving colorimetric derivatization with 2,4-dinitrophenylhydrazine (DNPH). This DNPH assay has a limit of detection of ~30 µM based on comparison with standard curves. Based on EcYfaW protein purification yields (~10 mg/L), we estimate that the cell lysate used in this screen had ~ 0.2 µM enzyme. Over 2 h, the YfaW variants needed to perform ~150 turnovers of 2R-DHIV to be detected by the DNPH screen. Wild-type EcYfaW catalyzes ~3 turnovers/sec on L-rhamnonate, so if any of the YfaW variants had a fraction of wild-type activity on 2R-DHIV then they should be detected. No variants were found with improved activity in these single-site libraries.

Since the position 33 histidine and the position 59 arginine are predicted to make hydrogen bonds with the C4 and C5 hydroxyls of L-rhamnonate that are not present in 2R-DHIV (**Figure 4-1**), we speculated that they may be particularly important sites for substrate binding. We made two-site combinatorial libraries with NNK codons at both H33 and R59, and screened 2000 variants. No variants were found with improved activity in this library, however.

We next decided to mutate the entire substrate binding pocket, and constructed two large combinatorial libraries, one five-site library with NNK codons at H33, L41, L45, R59, and P191, and one six-site library with NNK codons at H33, L41, L45, R59, P191, and L351. 2800 variants of

Table 4-2: Libraries screened.

Method	Sites	Mutation rate	Variants assessed	Target substrate	Verified hits
Targeted mutagenesis	H33	NNK	90	2R-DHIV	0
	I41	NNK	90	2R-DHIV	0
	I45	NNK	90	2R-DHIV	0
	R59	NNK	90	2R-DHIV	0
	P191	NNK	90	2R-DHIV	0
	L351	NNK	90	2R-DHIV	0
	H33-R59	2xNNK	2000	2R-DHIV	0
	H33-L41-L45-R59-P191	5xNNK	2800	2R-DHIV	0
	H33-L41-L45-R59-P191-L351	6xNNK	2300	2R-DHIV	0
Error-prone PCR	EcYfaw	1-2 AA/gene	2000	2R-DHIV	0
	GzYfaW	1-2 AA/gene	2000	2R-DHIV	0
Growth selections	EcYfaw	3-4 AA/gene	~10^6	2R-DHIV	0
	GzYfaW	3-4 AA/gene	~10^6	2R-DHIV	0
	PpYfaW	3-4 AA/gene	~10^6	2R-DHIV	0
	H33-L41-L45-R59-P191-L351	6xNNK	~10^5	2R-DHIV	0
Substrate walking	EcYfaw	1-2 AA/gene	2000	L-lyxonate	0
	I41-P191	2xNNK	1200	L-lyxonate	0
	EcYfaw	1-2 AA/gene	2000	D-erythronate	0
	H33-L41-L45-R59-P191	5xNNK	2800	D-erythronate	0

the first library and 2300 variants of the second library were screened, and no improved variants were found.

In previous substrate specificity engineering work from this lab, mutations that changed activity could be found outside of the enzyme active site[106]. We applied error-prone PCR to EcYfaW, creating a library with an average of ~1-2 amino acid mutations distributed across the gene. We also created a library with a similar mutation rate on a more thermostable YfaW variant from *Gibberella zeae*, which is stable up to ~85 ºC, since more stable proteins have been found to be more evolvable[107]. We screened 2000 variants from both libraries but no improved variants were found.

4.4 Directed evolution of an enolase- Substrate walking

When an enzyme displays no activity on a substrate, it may be possible to engineer activity by successively evolving the enzyme towards substrates that are progressively more similar to the one of interest[106]. This is known as "substrate walking". L-lyxonate and D-erythronate are have structures intermediate between L-rhamnonate and 2R-DHIV, and we hypothesized that we could use substrate walking to shift the specificity of EcYfaW towards 2R-DHIV. We screened 2000 members of an EcYfaW error-prone library with an average of ~1-2 amino acid mutations per gene on both L-lyxonate and D-erythronate. Although wild-type EcYfaW has activity on L-lyxonate, we were unable to find any mutants with improved activity. We were also unable to find mutants with any activity at all on D-erythronate.

We also applied two-site NNK saturation mutagenesis to the I41 and P191 positions, targeting the two residues that were closest to the extra C6 group that distinguishes L-rhamnonate from L-lyxonate. We screened 1200 variants on L-lyxonate, but no variant had increased activity. Lastly, 2800 variants of the H33-L41-L45-R59-P191 five-site NNK library discussed above were screened on D-erythronate, and no variants were found.

4.6 Directed evolution of an enolase- Growth selections

The cellular dehydration of 2R-DHIV is an essential reaction for biosynthesis of the branched chain amino acids valine and isoleucine. The dehydration product ketoisovalerate is transaminated to form valine and transacetylated to form a precursor to leucine[108]. Therefore, strains without a DHAD cannot grow in minimal media lacking these branched chain amino acids. Since the ultimate goal of this engineering work is to replace DHAD in the cell with an engineered enolase, growth selections are a potentially useful approach to finding variants with activity

We confirmed that the *E. coli* BL21(DE3) ΔIlvD strain could not grow on M9 minimal media plates (with glucose as a carbon source). Moreover, supplementation with valine and leucine (each at 35 µg/mL) or expression of IlvD from the pET28a vector could both restore growth. We also found that M9 minimal media with both leucine at 35 µg/mL and valine at 2 µg/mL was insufficient for growth of *E. coli* BL21(DE3) ΔIlvD. Increasing the concentration of valine to 5 µg/mL or more allowed progressively more growth. We therefore used M9 minimal media with 35 µg/mL isoleucine and leucine and 2 µg/mL valine as the selection conditions, to allow small 2R-DHIV dehydration activities to restore growth.

We constructed large (~10^6 variants) error-prone libraries of EcYfaw, GzYfaW, and PpYfaW with ~3-4 amino acid mutations per gene, and expressed the variants in *E. coli* BL21(DE3) ΔIlvD. We also constructed a six-site combinatorial NNK libraries mutated at H33, L41, L45, R59, P191, and L351. We plated these variants on the selection plates, but after incubating the cells for over five days at 37 °C we were unable to detect activity.

4.7 Directed evolution of an enolase- Rational design

We were only able to screen a small fraction of our combinatorial NNK libraries, which means there may be a combination of mutants that allow 2R-DHIV binding but were not evaluated in our screen. Computational protein design could potentially find these rare combination mutants that have the desired activity on 2R-DHIV. We applied the Rosetta design algorithm[109] to YfaW structure to optimize the substrate binding pocket for 2R-DHIV. The top six designs (as well as the tenth, which had a mutation at L351 that we wished to test), were cloned and expressed (**Table 4-3)**.

We conducted activity assays on L-rhamnonate, L-lyxonate, D-erythronate, and 2R-DHIV, and found that each of the designed variants was active on L-rhamnonate, some were active on L-lyxonate, and none were active on D-erythronate or 2R-DHIV.

Table 4-3: Top Rosetta-based designs for 2R-DHIV. Mutants are scored by predicted ligand binding.

Mutant	Rosetta rank	Active on L-rhamnonate	Active on L-lyxonate	Active on D-erythronate	Active on 2R-DHIV
H33T, I45V, R59T, I64S, P191V	1	Yes	No	No	No
H33T, I45L, R59T, I64S, P191T	2	Yes	No	No	No
H33T, R59T, I64S, P191A	3	Yes	Yes	No	No
H33T, I45L, R59T, I64S, P191A	4	Yes	Yes	No	No
H33T, I45D, R59T, I64S, P191T	5	Yes	No	No	No
H33T, I45L, R59T, I64S, P191H	6	Yes	Yes	No	No
H33T, I45V, R59T, I64S, P191V, L351I	10	Yes	Yes	No	No

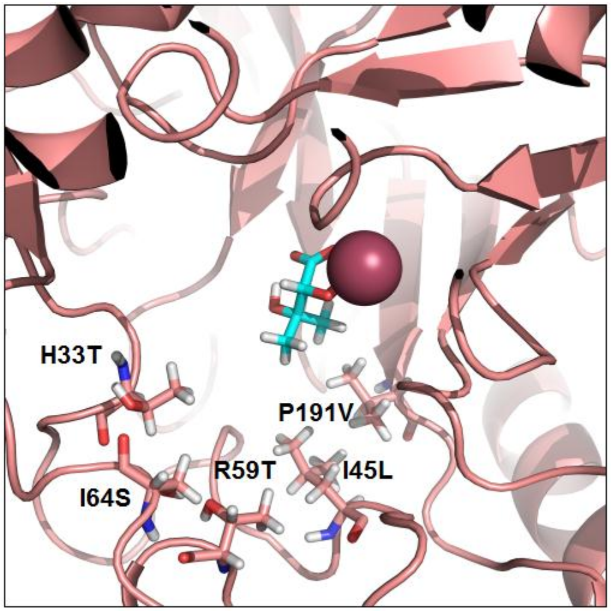


Figure 4-2. Predicted structure of top Rosetta design for EcYfaW. The positively charged arginine at position 59 is replaced by threonine, and histidine 33 is replaced by a less polar threonine as well. Proline at position 191 is replaced with a valine, helping to fill the substrate binding pocket. Isoleucines at positions 45 and 64 are replaced by leucine and serine, respectively.

4.8 Discussion

We report here a comprehensive protein engineering study that was unable to engineer an Mg^{2+} dependent enolase to replace an Fe-S dependent dehydration reaction in an essential metabolic pathway in the cell. There are at least three possible reasons that directed evolution was unable to change the specificity of YfaW from L-rhamnonate to 2R-DHIV: the specificity changes may have been too small to have been detected in the screen, the mutations that could have changed YfaW specificity may not have been accessed, or this enzyme may in fact never be able to be engineered to catalyze this reaction.

As discussed above, the DNPH assay we used to screen YfaW mutants had a sensitivity of ~30 µM, which should allow detection of 2R-DHIV turnovers as low as 1/min. Moreover, the site-saturation and error-prone libraries we constructed cover a very large mutation space around wild-type EcYfaW. These include near complete diversification of every substrate binding pocket residue contacting the terminal end of the substrate, as well as a sizeable fraction of single mutants across the protein (**Table 4-2**). We conclude that even if there are YfaW variants that can dehydrate 2R-DHIV, they are likely rare, and possibly inaccessible to current protein engineering approaches.

Why is it so difficult to engineer 2R-DHIV dehydration activity in YfaW? One key difference between L-rhamnonate and 2R-DHIV is that there are two less hydroxyl groups in 2R-DHIV, and two less hydrogen bonds are formed with H33 and R59 (**Figure 4-1**). These interactions may be important for binding the substrate for catalysis. However, previous protein engineering work by this lab has created P450 monooxygenases that have high activities on molecules as small as propane[106]. KARIs have also been created that have high activities on NADH, which makes four to five fewer hydrogen bonds to the enzyme than the native substrate NADPH[98].

An essential question in protein engineering is what makes particular protein scaffolds evolvable[110]. The enolase super family of enzymes displays a diversity of substrate specificities and reaction types[101], which have been altered in the laboratory through single mutations[111]. The dehydratase subgroup in particular is reported to catalyze dehydration of many different six carbon sugars (and one five carbon sugar, lyxonate). In this study it may be the case that diversity of substrate specificities within a class of molecules (*e.g.* six carbon sugars) does not imply ease of evolvability of reactivity for a related class of molecules (*e.g.* the smaller 2R-DHIV).

In a broader sense, it may also be the case that Nature has optimized cofactor choice for metabolic reactions in the cell, which may explain why the Fe-S cofactor is used in preference to an Mg^{2+} cofactor. "Biological optimality", however, is a complex and multi-dimensional property[112; 113]. Indeed, this lab has shown that changing cofactors can improve cellular biosynthesis of isobutanol under anaerobic conditions[99]. Optimality, then, may be highly context and environment dependent. The results of this study are unable to show that an Fe-S dependent DHAD is not the best enzyme for dehydrating 2R-DHIV in amino acid metabolism.

4.9 Methods

General methods are described in Appendix 1.

Cloning enolases

E. coli DH5α (Novagen) was used for cloning and BL21(DE3) (Novagen) was used for protein overexpression. *E. coli* BL21(DE3) ΔilvD strain was generated as previously described[114]. Standard methods for DNA isolation and manipulation were performed as described[115]. GlucD, GalD, ManD, EcYfaW, GzYfaW, and PpYfaW were synthesized by DNA 2.0 (Menlo Park, CA) and cloned into pET22b. See Appendix 2 for sequence information.

Expressing and purifying enolases

Constructs were transformed into *E. coli* BL21(DE3) ΔilvD strain for expression and selected on LB agar with 100 µg/ml ampicillin. Isolated transformants were grown in 400 mL LB +

ampicillin at 37 °C, 250 rpm, until an OD_{600} of 0.6. IPTG (0.1 mM) was used to induce gene expression at 25 °C, 250 rpm for 20 hours. The cells were harvested by centrifugation at 6,000 *g* for 12 min at 4 °C, and the pellets were stored at -80 °C or directly used for protein purification. Cell pellets were resuspended in 10 ml solvent A (50 mM Tris-Cl, pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, and 16 mM imidazole) and sonicated using a Sonicator 3000 (Misonix, Farmingdale, NY). Supernatant was collected after centrifugation at 36,000 g at 4 °C for 30 min. His-tagged proteins were purified with 1 mL His-trap column (GE Healthcare) with solvent A and solvent B (50 mM Tris-Cl, pH 8.0, 100 mM NaCl, 300 mM imidazole, and 10 mM $MgCl_2$) on an Akta FPLC system (GE Healthcare, Uppsala, Sweden). The purified proteins were desalted using Vivaspin 20 ultrafiltration spin tubes (GE Healthcare, Pittsburgh, PA), and their concentrations were calculated by their extinction coefficients at 280 nm (determined using ProtParam on the ExPASy server[68]). Protein yields from 1 L cultures were ~10mg.

Chemical synthesis of L-rhamnonate, L-lyxonate, and D-erythronate

Acid sugars were synthesized as previously described[116]. Briefly, 5 g sugar (Carbosynth, UK) and 10 g barium carbonate were combined in 42mL ddH2O on ice. 2 mL bromine (Sigma-Aldrich, St Louis, MO) was added in .5mL aliquots, with stirring, and the reaction mix was incubated at room temperature for 6 h with stirring. The mixture was aerated by a stream of air to remove excess bromine, and filtered with a Büchner funnel and cellite (Sigma-Aldrich) to remove barium carbonate. The solution was concentrated to < 50 mL by rotor-evaporator, and re-concentrated after addition of 50 mL H_2O. H_2O was added to make ~ 150 mL and the pH was adjusted to 10 with NH_4OH. The sample was loaded onto an ion exchange column (150 mL bed-volume Dowex AG1X8 column, pre-washed with 3N formic acid and H_2O). The product was eluted using a linear gradient of 0 to 1.5 M formic acid in 2000 mL and collected in 25 mL fractions. Fractions with product were found by spotting on Merck silica 60 TLC plates

(EMDMillipore, Billerica, MA) and developing with p-anisaldehyde reagent (1 mL p-anisaldehyde, 2 mL of conc. sulfuric acid, and 100 mL glacial acetic acid (Sigma-Alrich)). After removing solvent by rotary evaporation, the lactones were hydrolyzed by raising pH to 10 with NaOH. Products were analyzed by 1H NMR (600 MHz, MeOD) and identified by comparison to published data[116].

Biochemical analysis of purified YfaWs

Enzyme reaction mixtures contained 2 µM EcYfaW and 5 mM substrates (L-rhamnonate, L-lyxonate, and D-erythronate) in 100 µl of reaction buffer (50 mM Tris-Cl, pH 8.0, and 10 mM MgCl2). Reactions were incubated at 37 °C for 30 min and terminated by adding 12.5 µl of 20% trichloroacetic acid (TCA). A 1200 series HPLC from Agilent Technologies (Santa Clara, CA) was used to quantitatively detect the product after derivatization with SCA. The Agilent Eclipse XDB-C18 column (5µ, 4.6 x 100 mm) was used with solvent A (0.2% formic acid in water) and solvent B (acetonitrile). The products were eluted by increasing solvent B percentage from 1% to 15% over 6 min at 1 ml/min. The eluted products were detected at wavelength of 250 nm. All experiments were repeated at least twice.

Semi-rational design of libraries, random mutagenesis, and library screening

Mutation sites were selected following an inspection of 3-deoxy-L-rhamnonate binding in StYfaW (3CXO)[102]. The PyMOL Molecular Graphics System (Version 1.3, Schrodinger, LLC)[117] was used to identify H33, I41, I45, R59, P191, and L351 as sites potentially influencing substrate binding. Overlap-extension PCR reactions was used to introduce NNK degenerate codons as described[65], either as single or combinatorial mutations. Error-prone PCR was carried out as described[118], using $MnCl_2$ concentrations ranging from 150 µM to 250µM to generate desired mutation rates (which were determined by sequencing 10 variants per library). The resulting constructs were transformed into *E. coli* BL21 (DE3) ΔilvD and grown as described[119]. For lysis, cell pellets were resuspended in 300 µL Tris-Cl (50 mM, pH 8, 0.6 mg/mL lysozyme, 2-4 U/mL DNase I, and 10 mM $MgCl_2$) and were incubated at 37 °C for 1 h. After centrifugation at 5,000 x g, 4 °C for 15 min, 100 µL crude lysates were transferred to 96-well PCR plates containing 10 mM substrates. The reactions were incubated at 37 °C for 1 hour for L-rhamnonate and 3 hours for L-lyxonate and D-erythronate. The reactions were then terminated with 2.2% TCA. Protein was precipitated at room temperature for 5 min and removed by centrifuging at 5000 *g* for 10 min. DNPH saturated in 2N HCl (125 µL) was used to derivatize 2-keto acids in 100 µL supernatant at 37 °C for 30 min. The pH of derivatization mixture was adjusted to >7 with 33 µL 10 N NaOH. After thorough mixing the solution was further incubated at 37 °C for 10 min and all precipitates were removed by centrifuging at 5000 *g* for 5 min. The absorbance of 160 µL supernatant was record in 96-well plates at 550 nm.

Electrocompetent BL21(DE3)ΔilvD

Very highly competent BL21(DE3)ΔilvD were created according to a method described by Sidhu and Fellouse, allowing ~10^8 cfu/µg of DNA, which exceeded other competent cell protocols by at least an order of magnitude[120]. Briefly, cells were grown in 2mL 2YT for 8 h at 37 ºC 250 rpm. Cells were inoculated into 25 mL 2YT culture overnight at 37 ºC 25 rpm. 5 mL overnight culture was transferred into each of six 2.8 L baffled flasks, containing 1 L superbroth (12 g tryptone, 24 g yeast extract, 5 mL glycerol, 17 mM KH_2PO_4, 72 mM K_2HPO_4 in 1 L ddH_2O), and grown until an OD_{550} of 0.8. Culture was centrifuged at 5000 *g* for 10 min at 4 ºC, and resuspended in 1 mM Hepes pH 7.0 (300 mL total volume for combined cells). Culture was centrifuged again, and additional Hepes washes were performed twice. Culture was washed a final time with 10% glycerol. Cells were resuspended in 3 mL 10% glycerol, and frozen as 350 µL aliquots in liquid nitrogen, and stored at -80 ºC.

For transformation, electrocompetent cells were mixed with 5-20 µg of plasmid in 50 µL, and transferred to a chilled 0.2 cm gap electroporation cuvette (USA Scientific, Orlando, FL). Electroporation was done in a Gene Pulsar II Electroporation system (Bio-Rad Laboratories, Hercules, CA) at the following settings: 2.50 kV field strength, 200 Ω resistance, 25 µF capacitance. SOC medium (0.5 % w/v yeast extract, 2% w/v tryptone, 10 mM NaCl, 2.5 mM KCl, 20 mM $MgSO_4$, 20 mM glucose) was immediately added and used to transfer the cells to a final volume of 25mL SOC medium in a 250mL baffled flask. Culture was incubated for 30 min at 37 ºC with shaking, and then transferred to selective media.

Growth selections

Libraries were constructed using error-prone PCR and transformed into highly electrocompetent *E. coli* BL21 (DE3) ΔilvD cells. Growth selections were carried out on M9 minimal media agar plates (64 g $Na_2HPO_4.7H_2O$, 15 g KH_2PO_4, 2.5 g NaCl, 5 g NH_4Cl, 2mM $MgSO_4$, 0.1mM $CaCl_2$ in 1L H_2O)[115], with glucose (0.4%) as the carbon source and supplemented with 35 µg/mL leucine and 2 µg/mL valine. ~10^5 transformants were plated on 1 ft^2 plates, which were incubated at 37 °C for up to five days. Growth was assessed every 12 h.

Computational design

The Rosetta design program[109] was run on the StYfaW structure with modeled 2R-DHIV to optimize the substrate binding pocket. 208 rounds of Rosetta design were carried out, and the top 66 designs by number of counts were ranked by ligand binding. The top ten designs were visually inspected, and designs one through six and ten were chosen for evaluation. These designs were constructed by overlap PCR, and expressed and purified as described above. Reactions were run overnight at 37 °C, with 10 µM enzyme, 5mM substrate (L-rhamnonate, L-lyxonate, D-erythronate, 2R-DHIV) in a 200 µL reaction volume.

Chapter 5

DIRECTED EVOLUTION OF AN ALCOHOL DEHYDROGENASE FOR IMPROVED BIOFUELS PRODUCTION FROM LIGNOCELLULOSE

5.1 Abstract

The conversion of lignocellulose into soluble sugars for fermentation is a major obstacle to the development of next-generation biofuels. Enzymatic degradation of lignocellulose is expensive and inefficient, while chemical degradation generates toxic byproducts that inhibit subsequent fermentation. One possible solution to this problem is the engineering of fermenting microorganisms that have high tolerance to the inhibitors present in lignocellulose hydrolysate. Alcohol dehydrogenases are attractive candidates for increasing aldehyde resistance, since they can reduce aldehydes, a major class of inhibitory compounds, into less toxic alcohols.

We applied directed evolution to engineer an alcohol dehydrogenase that confers increased resistance to a broad range of toxic aldehydes. We chose *Saccharomyces cerevisiae* ADH6 for our studies, since it is known to reduce many of the toxic aldehydes present in lignocellulose hydrolysate and could increase resistance to the aldehydes cinnamaldehyde and 5-hydroxymethylfurfural (5-HMF). We created error-prone libraries of ADH6, and selected for resistance to combinations of different aldehydes, including cinnamaldehyde, 5-HMF, syringealdehyde, coniferaldehyde, and vanillin. However, we were unable to find any variants that could improve resistance above the wild-type enzyme.

This work highlights the difficulty of engineering cellular resistance to aldehydes by enzymatic methods, and in particular the difficulty of engineering a single enzyme for broad resistance. We discuss possible reasons why directed evolution of ADH6 was not able to increase

resistance, which include 1) enzymatic activity increases may have been too small to have been detected in the selection, 2) ADH6 may already have been optimized by natural evolution for activity on aldehyde inhibitors, and 3) NADPH pools may be limiting under the conditions we used. Alternate strategies may be required to address the problem of toxic lignocellulose degradation byproducts.

5.2 Toxic byproducts restrict chemical methods for hydrolyzing lignocellulosic biomass

Next-generation biofuels are a potentially sustainable, affordable, and environmentally friendly replacement to fossil fuels. Biofuel production involves degradation of lignocellulose from plant material, followed by fermentation of the resulting soluble sugars into fuels such as ethanol and isobutanol[121]. One of the major constraints to the development of biofuels is the difficulty in degrading lignocellulose. Lignocellulose is composed of cellulose, a polymer of glucose, and lignin, a heterogeneous polymer of aromatic aldehydes and acids. Both polymers are tightly bound, forming a paracrystalline material that is recalcitrant to degradation[79].

Two methods have been explored to degrade lignocellulose: enzymatically with cellulases and chemically using acid and high temperatures and pressure. Enzymatic degradation is the most commonly used method, but is constrained by the low activity of cellulases, as well as their low stabilities in the industrial conditions of high temperature, long incubation times, and extreme pH values[72]. This makes the process costly and time-consuming.

On the other hand, chemical hydrolysis of lignocellulose biomass generates toxic byproducts that inhibit growth of the fermenting microorganisms (such as the yeast *Saccharomyces cerevisiae*), which compromises biofuels yields[122]. In particular, degradation of sugars from the cellulose fraction yields furans and carboxylic acids, while degradation of the lignin fraction yields phenolic acids and aldehydes[123]. Physical and chemical means have been employed to remove these inhibitors prior to fermentation; however, they are expensive and decrease sugar yields. Improving cellular resistance to lignocellulose hydrolysate inhibitors would therefore be one way to improve the overall cost-effectiveness of biofuels production.

5.3 Lignocellulose Hydrolysate Toxicity

Table 5-1 lists the major growth inhibitors present in lignocellulose hydrolysate and their minimal inhibitory concentrations for *S. cerevisiae*. The observed concentration of these chemicals (~g/L) often exceeds their minimal inhibitory concentrations, which suggests that improving resistance to these chemicals can improve microbial growth and biofuel yields. The mechanisms of toxicity for these inhibitors is complex, and depends on strain characteristics such as cell membrane composition and metabolism[123]. In some cases biofuel production can actually be increased by adding sub-inhibitory concentrations of toxins like phenolics, acetic acid, and furfural, because reducing cell mass allows more of the biomass to be converted into fuel. In general, inhibitor toxicity is determined both by the functional group (aldehyde > acid > alcohol) and the hydrophobic nature of the compound (hydrophobic > hydrophilic)[124].

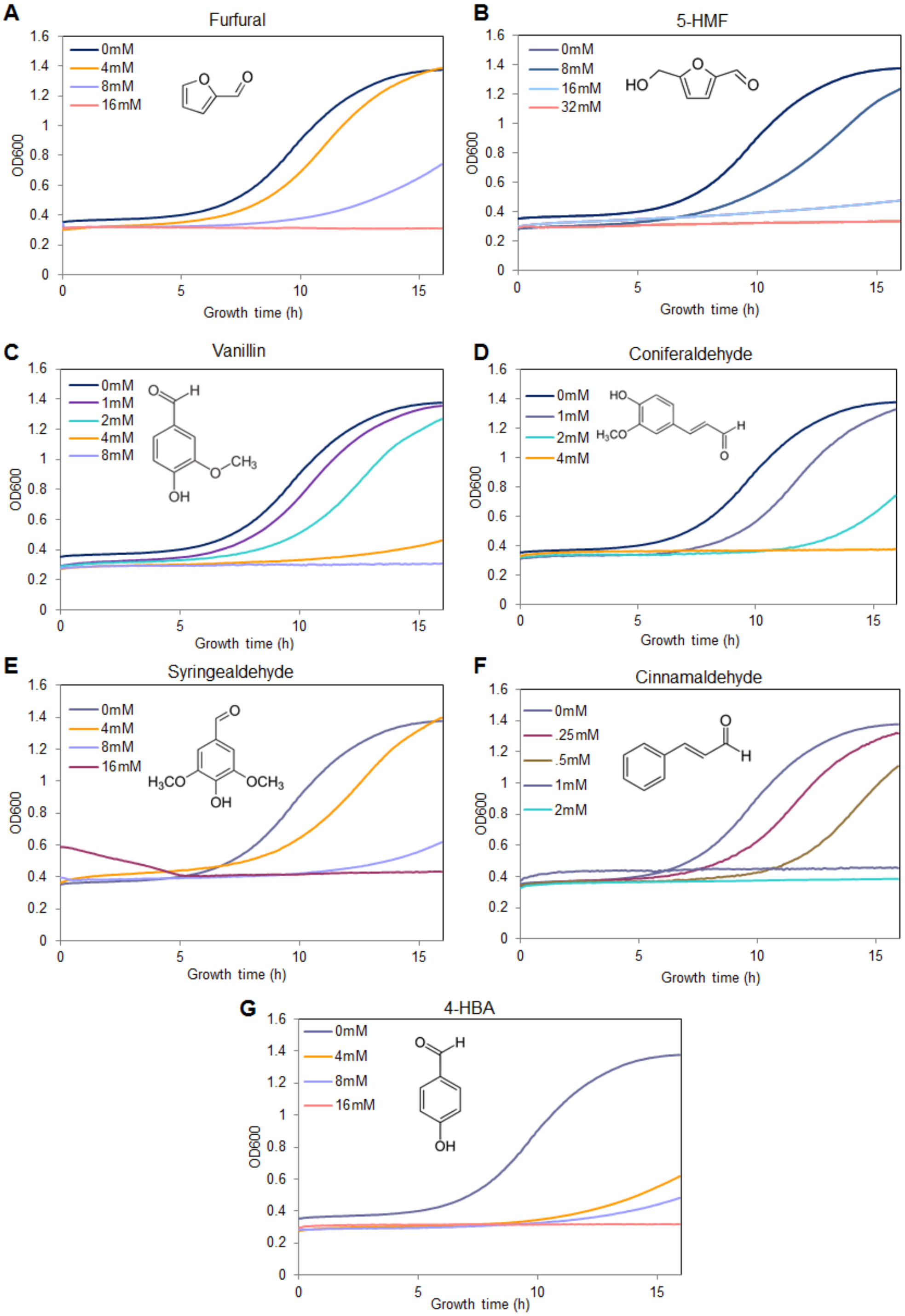
The most important class of inhibitors is the aldehyde inhibitors. This class consist of furans (furfural and 5-hydroxymethyl furfural (5-HMF)) and phenols (4- hydroxybenzaldehyde (4-HBA), vanillin, coniferaldehyde, cinnamaldehyde, and syringaldehyde). Aldehyde inhibitors are the most abundant class of inhibitors by diversity, and the most toxic. **Figure 5-1** shows growth curves of the wild-type CEN-PK2 strain of yeast in six important aldehyde inhibitors; minimal inhibitory concentrations are in the low millimolar range which is typical in lignocellulose hydrolysate[123].

The inhibitory effects of 5-HMF and furfural may arise from inhibition of enzymes involved in glycolysis. It has been reported that furfural competitively inhibits alcohol dehydrogenase, aldehyde dehydrogenase, and pyruvate dehydrogenase in yeast, thus preventing pyruvate from entering the citric acid cycle, and compromising metabolism[125]. 5-HMF and furfural have also been found to be nonspecifically reduced by various NADPH-dependent alcohol

Table 5-1: **A few key *S. cerevisiae* growth inhibitors present in lignocellulose degradation products.** From Klinke *et al.*[123]

Inhibitor	Molecular weight (g/mol)	Typical hydrolysate concentrations (mM)	Minimal inhibitory concentration (mM)
Furans			
Furfural	96.09	10	10
5-hydroxymethyl furfural	126.11	15	8
Phenols		~5	
4-hydroxybenzaldehyde	122.12		5
Vanillin	152.15		3
Orthovanillin	152.15		1
Coniferaldehyde	178.18		1
Syringaldehyde	182.17		4
Carboxylic Acids			
Acetate	60.05	40	150
Ferulic Acid	194.18	?	1
4-hydroxycinnamic acid	164.16	?	6
Vanillic acid	168.15	?	6

Figure 5-1: Growth inhibition of *S. cerevisiae* by aldehydes present in lignocellulose hydrolysate. Growth curves of the CEN-PK2 yeast strain in furfural (**A**), 5-HMF (**B**), vanillin (**C**), coniferaldehyde (**D**), syringealdehyde (**E**), cinnamaldehyde (**F**), and 4-HBA (**G**) are shown.



dehydrogenases in E. coli, which depletes cellular NADPH pools, thereby inhibiting cell growth[126]. The NADPH pool does not appear to be limiting in yeast, since overexpression of NADPH-dependent alcohol dehydrogenases can increase resistance, as described below.

5.4 Modes of Resistance

Resistance to aldehydes inhibitors can in principle be acquired in three ways, and each offers attractive possibilities for bioengineering efforts. One way is to alter intracellular inhibitor targets, such as metabolic pathways or cell membranes[123; 127], by engineering transcription factors that govern metabolism or enzymes that control cell membrane composition. A second way would be to have efflux pumps actively transport the inhibitor out of the cell. This has been shown to work for increasing resistance to the inhibitory effects of biofuels in *E. coli*[128]. Lastly, enzymes could convert the inhibitors into less toxic compounds.

This last mode of resistance would in principle be the preferred of the three, because a detoxifying enzyme would be able to be used in a variety of fermenting microorganisms. Some enzymes that have been studied for this purpose include aldehyde dehydrogenases that oxidize the aldehydes into acids[129], alcohol dehydrogenases that reduce aldehydes into alcohols[127; 130], and laccases that oxidize phenol groups so that the inhibitors precipitate[131].

In Nature there exists a family of enzymes known as the Cinnamyl Alcohol Dehydrogenases (CAD), which can catalyze the interconversion of the alcohol and aldehyde forms of many aldehyde inhibitors[132]. In plants, these enzymes catalyze the final steps in the biosynthesis of monolignols, which are precursors for lignin[133]. In bacteria and yeast, it has been hypothesized that these enzymes are responsible for assisting in the degradation of lignin. *Saccharomyces cerevisiae*, for example, has two members of the CAD family: ADH6 and ADH7. ADH6 (YMR318C) in particular has been studied in the context of aldehyde resistance[133]. This enzyme is an NADPH-dependent member of the cinnamyl alcohol dehydrogenase family, and functions as a homodimer with 39.6 kDa subunits. ADH6 reduces a broad spectrum of aldehyde inhibitors with high activity, including cinnamaldehyde, coniferaldehyde, vanillin, and furfural[134]. The broad substrate range of ADH6 is likely related to the way the substrate fits into the catalytic cleft of the enzyme: the substrate is predicted to be flanked by hydrophobic residues around the phenol ring, and the end of the phenol ring that can have hydrophilic modifications (*e.g.* in syringaldehyde and coniferaldehyde) is solvent-exposed[133].

5.5 ADH6 promotes aldehyde tolerance

In the BY4741 strain of yeast (a standard wild-type strain that has deletion mutant libraries available[135]), ADH6 deletion mutants are reported to have decreased resistance to 5-HMF and syringaldehyde (personal communication, Joseph T. Meyerowitz). We overexpressed ADH6 from the yeast expression vector pJTM031, resulting in approximately 10-fold increased ADH6 activity in cell lysate (as assessed by activity on cinnamaldehyde and 5-HMF, **Figure 5-2**). When tested on 5-HMF and cinnamaldehyde overexpression of ADH6 increased resistance slightly (~20% increase in minimal inhibitory concentration) (**Figure 5-3**). Overexpression of ADH6 did not, however, increase resistance to furfural, vanillin, coniferaldehyde, or syringealdehyde, although purified protein was active on these substrates (**Figure S5-1** and **S5-2**). These results confirm and extend the literature reports that ADH6 may play a role in aldehyde resistance.

5.6 Directed evolution of ADH6 for broadly increased aldehyde resistance

We applied directed evolution to ADH6 to explore two major questions. First, to what degree can aldehyde resistance be improved in *Saccharomyces cerevisiae* by intracellular aldehyde reduction activity? Second, is it possible to simultaneously engineer for *both* high activity and broad substrate range in an enzyme[136]? To answer these questions, we created error-prone libraries of ADH6 in *S. cerevisiae* and selected for variants that increased growth on media containing mixtures

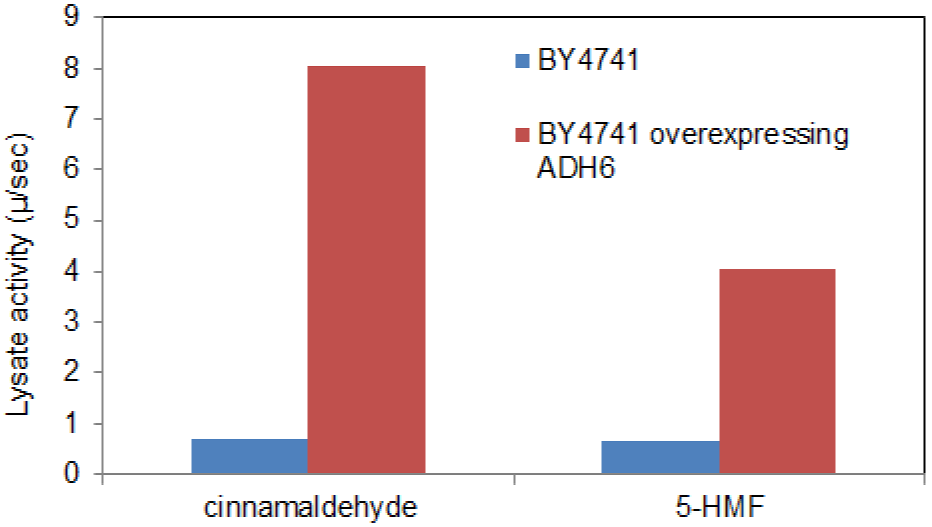


Figure 5-2: Overexpression of ADH6 in the yeast strain BY4741 increases lysate activity on cinnamaldehyde and 5-HMF. Lysate activity on both substrates is approximately 10 times greater in a BY4741 strain overexpressing ADH6 compared to a control strain BY4741 strain.

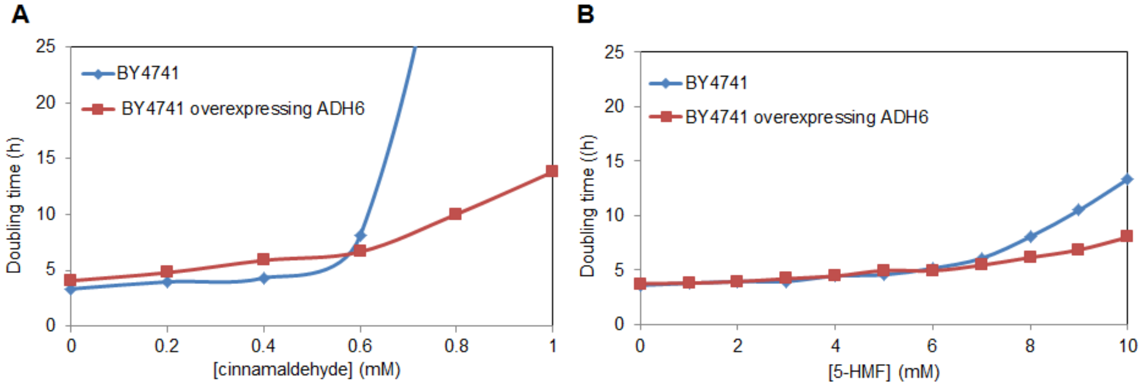


Figure 5-3: **Overexpression of ADH6 in the yeast strain BY4741 can improve aldehyde resistance** on (**A**) cinnamaldehyde and (**B**) 5-HMF. Doubling times were evaluated over an 8 h period in log phase growth. The minimal inhibitory concentration is shifted by 0.2 mM for cinnamaldehyde and by 2 mM for 5-HMF.

of different aldehydes. The yeast strain we used in our studies was CEN.PK2, a strain with high innate tolerance to aldehydes[137] and that is preferred for physiological characterization[138] and aldehyde tolerance studies[139]. Unlike in BY4742, overexpression of ADH6 in CEN.PK2 was not able to increase resistance to cinnamaldehyde or 5-HMF, due to higher tolerance in the control strain expressing ADH6 at normal levels. CEN.PK2 therefore represents a more industrially relevant but more difficult target for strain engineering efforts.

Selections were carried out on YPD agar plates containing combinations of aldehydes which were slightly higher (~25%) than a concentration that gave barely detectable growth. Since combinations of aldehydes were more toxic than each aldehyde alone, selection concentrations were optimized for each condition. Plates were incubated at 30 ºC for up to 5 days. Colonies that grew were rescreened on selective plates, and those that passed this rescreen had their plasmids isolated and retransformed into wild-type CEN.PK2 yeast. These re-transformants were again screened to ensure increased aldehyde tolerance was not due to change in genetic background.

Two ADH6 error-prone libraries containing 10^7 variants with mutation levels of 2 and 3 amino acids per gene were selected in conditions of: 2 mM cinnamaldehyde + 12 mM vanillin, 2 mM cinnamaldehyde + 12 mM coniferaldehyde, 2 mM cinnamaldehyde + 15 mM syringealdehyde, and 15 mM furfural + 15 mM 5-HMF (**Table 5-2**). If variants were found that had higher resistance to these aldehydes, selections with three or more aldehydes could then be carried out. Unfortunately, no variants with improved resistance were found.

5.7 Directed evolution of alcohol dehydrogenases for increased resistance on single aldehydes

It may be the case that resistance cannot be increased to two aldehydes simultaneously by directed evolution of ADH6, but that resistance to single aldehydes could be improved one at a time. We investigated this possibility by selecting for ADH6 variants that conferred increased resistance to 5-HMF, cinnamaldehyde, coniferaldehyde, vanillin, and 4-HBA. We also tested three

Table 5-2: Alcohol dehydrogenase libraries created and screened. Mutations were introduced by error-prone PCR.

Gene	Mutation level	Variants assessed	Target substrate	Verified hits
ADH6	2AA/gene	10^7	2 mM cinnamaldehyde + 12 mM vanillin	0
	3AA/gene	10^7	2 mM cinnamaldehyde + 12 mM vanillin	0
	2AA/gene	10^7	2 mM cinnamaldehyde + 12 mM coniferaldehyde	0
	3AA/gene	10^7	2 mM cinnamaldehyde + 12 mM coniferaldehyde	0
	2AA/gene	10^7	2 mM cinnamaldehyde + 15 mM syringealdehyde	0
	3AA/gene	10^7	2 mM cinnamaldehyde + 15 mM syringealdehyde	0
	2AA/gene	10^7	15 mM furfural + 15 mM 5-HMF	0
	3AA/gene	10^7	15 mM furfural + 15 mM 5-HMF	0
	2AA/gene	10^6	12 mM 5-HMF	0
	2AA/gene	10^6	2 mM cinnamaldehyde	0
	2AA/gene	10^6	5 mM coniferaldehyde	0
	2AA/gene	10^6	10mM vanillin	0
	2AA/gene	10^6	16mM syringealdehyde	0
	2AA/gene	10^6	15mM 4-HBA	0
ADH1	2AA/gene	$5*10^6$	12 mM 5-HMF	0
	2AA/gene	$5*10^6$	2 mM cinnamaldehyde	0
	2AA/gene	$5*10^6$	5 mM coniferaldehyde	0
	2AA/gene	$5*10^6$	10 mM vanillin	0
	2AA/gene	$5*10^6$	16 mM syringealdehyde	0
	2AA/gene	$5*10^6$	15mM 4-HBA	0
ARI1	2AA/gene	$5*10^5$	12 mM 5-HMF	0
	2AA/gene	$5*10^5$	2 mM cinnamaldehyde	0
	2AA/gene	$5*10^5$	5 mM coniferaldehyde	0
	2AA/gene	$5*10^5$	10mM vanillin	0
	2AA/gene	$5*10^5$	16mM syringealdehyde	0
	2AA/gene	$5*10^5$	15mM 4-HBA	0
GRE2	2AA/gene	10^5	12 mM 5-HMF	0
	2AA/gene	10^5	2 mM cinnamaldehyde	0
	2AA/gene	10^5	5 mM coniferaldehyde	0
	2AA/gene	10^5	10mM vanillin	0
	2AA/gene	10^5	16mM syringealdehyde	0
	2AA/gene	10^5	15mM 4-HBA	0

additional alcohol dehydrogenases from *S. cerevisiae* which are known to have activity on aldehyde inhibitors, ADH1[139], ARI1[140], and GRE2[141].

We constructed error-prone libraries with mutation levels of two amino acids per gene for ADH6 (10^7 variants), ADH1 ($5*10^6$ variants), ARI1 ($5*10^5$ variants), and GRE2 (10^5 variants). We selected these libraries on YPD agar plates containing 12 mM 5-HMF, 2 mM cinnamaldehyde, 5 mM coniferaldehyde, 10 mM vanillin, 16 mM syringealdehyde, and 15mM 4-HBA. No variants with resistance improved over wild-type were found, however.

5.8 Discussion

We were unable to increase the resistance of the CEN.PK2 yeast strain to aldehyde inhibitors by directed evolution of ADH6. There are many possible reasons for this, including that the selection may have not been sensitive enough to detect improvements in the enzyme, ADH6 may not have been able to be improved for these enzymes, and other cellular factors may limit resistance.

Recently, the successful directed evolution of the alcohol dehydrogenase GRE2 for increased resistance to 5-HMF was reported, increasing activity by 13 to 15-fold and allowing faster growth on 30 mM 5-HMF[141]. Unlike the agar plate-based selection used in this study, the authors used a 96-well plate based selection, and assayed for increased levels of 5-HMF reduction and cell growth using absorbance spectroscopy. This approach may have allowed smaller improvements to aldehyde reduction activity to be observed. The authors also used the INVSC1 strain of yeast, which is less tolerant to aldehydes than the CEN.PK2 strain used here[137]. It is not clear whether their improved GRE2 variant would have increased resistance in the CEN.PK2 background. 5-HMF is also not the preferred substrate of GRE2, which catalyzes furfural reduction with approximately 10-fold higher activity and heptanal reduction with approximately 40-fold higher activity. Promiscuous, low-activity functions may be evolved without compromising activity on primary substrates[142], but it remains an open question whether the natural activity can be improved, in particular for the alcohol dehydrogenases described here.

We chose to use agar plate-based selection, since we believed that increasing ADH6 activity on multiple aldehyde inhibitors may have required rare or multiple amino acid mutations. Selections allow orders of magnitude more variants to be assessed than 96-well plate based screens, but lack sensitivity. However, since we were able to detect a difference in resistance to a variety of aldehydes on agar plates for the BY4742 yeast strain with deletions in ADH6 and overexpressing ADH6, we reasoned that this method should have been sufficiently sensitive.

The BY4742 strain (as well as the INVSC1 strain) is less tolerant to aldehyde inhibitors than CEN.PK2. Using this strain would allow smaller improvements in aldehyde resistance to be detected. However, as shown in **Figures 5-2** and **5-3**, a ten-fold increase in ADH6 activity for the BY4742 strain only corresponds to a ~20% increase in resistance to cinnamaldehyde and 5-HMF. In order to translate to industrially relevant levels of resistance, at least a 50% increase in aldehyde resistance would be required, necessitating a further 100% increase in ADH6 activity. Such increases may not be possible with an enzyme which is already very active on these substrates-though it is still an open question whether enzymes outside of primary metabolism are optimized for their substrates[143].

One final concern with increasing aldehyde reductase activity to increase aldehyde resistance is that activity would consume reducing equivalents in the cell. NADPH depletion is a reported source of toxicity in *E. coli* grown in furfural, which could be relieved by the NADH dependent reductase FucO[126]. Although this mechanism of toxicity has not been reported in yeast, it may become a limiting factor once aldehyde reductase activity is sufficiently high. Interestingly, the GRE2 variant engineered to have increased activity on 5-HMF was reported to use NADPH in addition to its native cofactor NADH. Further work will be needed to understand the limits of cellular resistance to toxic chemicals, and to what degree it can be increased by protein engineering.

5.9 Methods

General methods are described in Appendix 1.

ADH6 Plasmid Construction

The gene encoding *S. cerevisiae* ADH6 was amplified using PCR from *S. cerevisiae* genomic DNA using standard techniques[115]. The gene was cloned into the yeast expression vector pJTM031, a custom made vector received as a gift from Joseph T. Meyerowitz. The sequence of this gene and the plasmid map of pJTM031 are in Appendix 2.

Expression and purification of ADH6

Yeast colonies expressing ADH6 with C-terminal His_6 tags were grown at 30 °C, first overnight in 10mL YPD + 50 µg/mL hygromycin medium and then expanded into 1 L YPD +50 µg/mL hygromycin medium for an additional 48 h. Cultures were centrifuged at 4500x*g* for 20 min, and then cell pellets were resuspended in 10 ml solvent A (50 mM Tris-Cl, pH 8.0, 100 mM NaCl, 10 mM $MgCl_2$, and 10 mM imidazole) and lysed using 20 mL Y-PER yeast protein extraction reagent (Thermo Scientific) by incubation at 25 ºC for 30 min with 100 rpm shaking. Lysate was centrifuged at 36,000 g at 4 °C for 30 min, and the supernatant was collected and filtered. His-tagged proteins were purified with 1 mL His-trap column (GE Healthcare) with solvent A and solvent B (50 mM Tris-Cl, pH 8.0, 100 mM NaCl, 300 mM imidazole, and 10 mM $MgCl_2$) on an Akta protein purifier (GE Healthcare, Uppsala, Sweden). The purified proteins were desalted using Vivaspin 20 ultrafiltration spin tubes (GE Healthcare, Pittsburgh, PA), and their concentrations were calculated by their extinction coefficients at 280 nm (determined using ProtParam on the ExPASy server[68]).

ADH6 activity assays

Reactions were carried out on purified protein as described[133]. Briefly, 0.5 µM purified enzyme was combined with 0.5 mM NADPH, 1 mM substrate, in 33 mM sodium phosphate buffer at pH 7.0 in a total volume of 300 µL, and assayed continuously at 365 nm for 3 min at 25 ºC. Reactions on yeast cell lysate were carried out by lysing 5 mL of cells in 500 µL Y-PER extraction reagent as above. 30 µL of lysate was added to substrate, NADPH, and phosphate buffer, and mixture activity was assessed as above.

Circular dichroism and thermal denaturation

Circular dichroism scans were performed with protein in phosphate buffer at a concentration of 5 µM using a 1 mm cuvette. Wavelength scans were performed at 25°C, scanning through the 190-250 nm range, with an averaging time of 5 sec and a wavelength step of 1.0 nm. Circular dichroism signal at 220 nm was used to monitor thermal denaturation. Protein at 5 µM was monitored from 25-95 °C in steps of 1 °C. The sample was subjected to an equilibration period of 1 min per each step before collecting measurements.

Plate-reader based growth assay

Overnight cultures of yeast strains grown in 5 mL YPD + 50 µg/mL hygromycin were diluted to an OD600 of 0.1 in fresh YPD + 50 µg/mL hygromycin, and 100 µL was added to each well of a 96-well microtiter plate. Aldehyde inhibitors were diluted in YPD + 50 µg/mL hygromycin, and 200 µL was added to the microtiter plate. Gas permeable transparent seals (Thermo Fisher Scientific, Waltham, MA) were affixed to the plates. Cells were grown for 18 h at room temperature in a Tecan Infinite M200, with 3 sec of 0.1 mm orbital shaking every 5 min. OD600 readings for each well were taken every 5 min.

Error-prone ADH6 library creation

Error-prone PCR was carried out as described, with $MnCl_2$ ranging from 150 to 250 µM, resulting in an average mutation rate of 2-3 amino acids per gene[118]. Large libraries (up to 10^7) of CEN-PK2 *S. cerevisiae* were created according to a method developed by Gietz and Schiestl[144]. Briefly, a single yeast colony was inoculated into 50 mL YPAD (YPD with an additional 40 mg/L adenine sulfate (Sigma Aldrich) in a 250 mL flask, and grown 14 h at 30 ºC at 250 rpm. After assessing OD600 and using the conversion that one OD600 unit corresponds to 10^7 yeast cells, $1.25*10^9$ yeast cells were transferred into a 50 mL centrifuge tube and spun for 3000 g for 5 min at 4 ºC. Cells were resuspended in 20 mL YPAD, and transferred into a 2 L culture flask with 280 mL 2xYPAD (YPD in half the volume of H_2O). Cells were grown at 30 ºC for 4 h at 200 rpm, until an OD600 of 2.0. Cells were centrifuged at 3000 g for 5 min at 20 ºC in 50 mL falcon tubes, and resuspended in 150 mL sterile H_2O. Cells were centrifuged again, and resuspended in 60 mL sterile H_2O. Cells were centrifuged a third time, and resuspended in 21.6 mL transformation mix (14.4 mL 50% w/v PEG, 2.16 mL 1.0M LiAc, 3.0mL 2mg/mL ssDNA, 2.04 mL 5 µg plasmid). After vigorous vortexing, cells were incubated for 50 min at 42 ºC, mixing at 5 min intervals. Cells were then centrifuged and the pellet was resuspended in 300 mL YPD. Cells were then grown for 4 h at 30 ºC with 250 rpm shaking. Cells were then harvested by centrifugation, either for selections or for freezing in 10 % glycerol for storage at -80 ºC.

Aldehyde growth selections

~10^6 transformants were plated on YPD agar + 50 µg/mL hygromycin + aldehyde inhibitors, in 245 mm x 245 mm square plates. Plates were incubated at 30 ºC for 5 days.

5.10 Supplementary Information

Figure S5-1 shows SDS-PAGE gels of the four alcohol dehydrogenases described in this study. **Figure S5-2** shows an example activity assay of ADH6 on cinnamaldehyde. **Figure S5-3** shows the circular dichroism spectra of ADH6, and its melting curve from change in ellipticity at 220nm, giving a T_m of ~50 ºC.

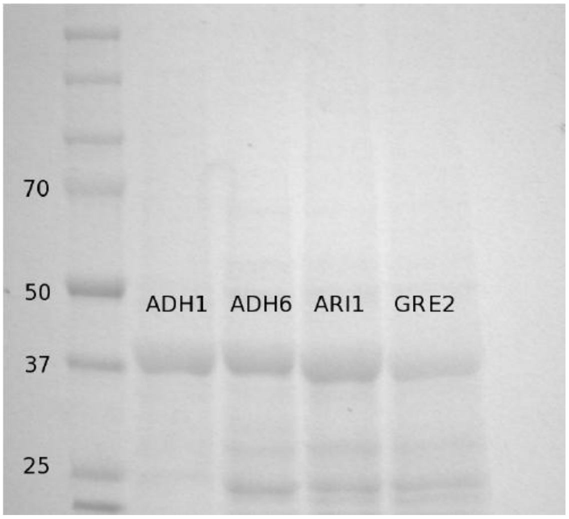


Figure S5-1. SDS-PAGE of alcohol dehydrogenases expressed in *S. cerevisiae.* Molecular weight markers are shown in kDa.

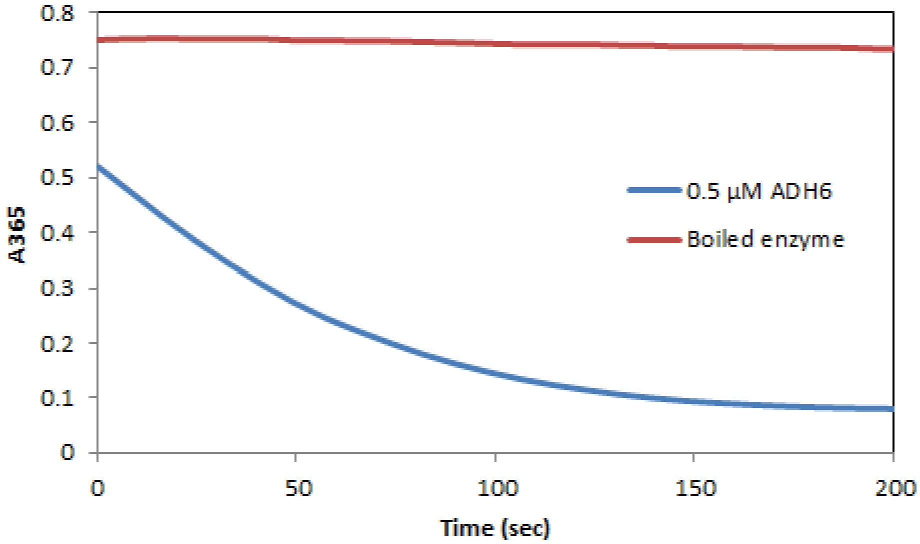


Figure S5-2. Example activity assay of ADH6. 0.5 µM ADH6 was incubated with 1 mM cinnamaldehyde and 0.5 mM NADPH in 33 mM sodium phosphate buffer pH 7.0. Absorbance was measured at 365 nm.

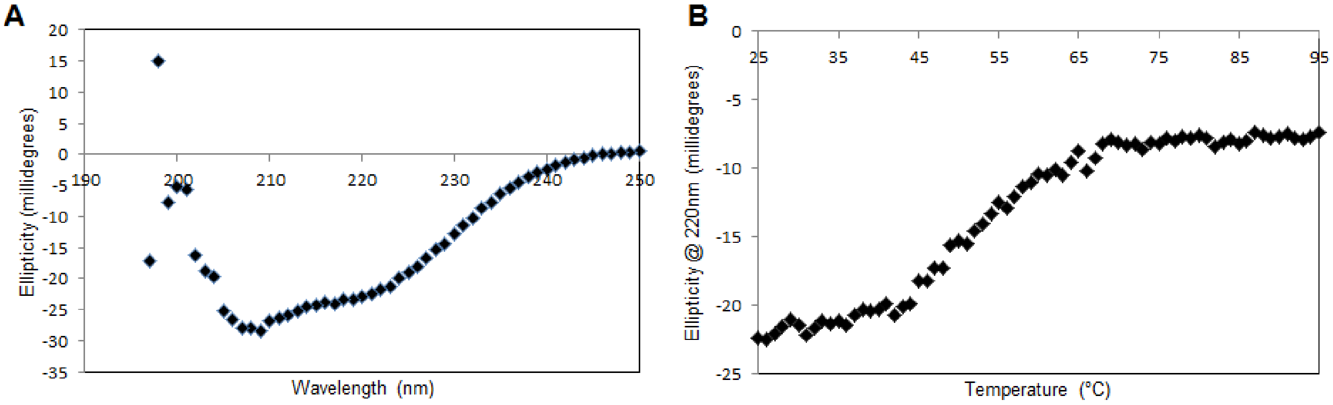


Figure S5-3. ADH6 circular dichroism spectra and thermal denaturation curve. A) ADH6 Ellipticity was measured from 190 nm to 250 nm. B) Change in ADH6 ellipticity was measured at 220nm as temperature was increased from 25 ºC to 95 ºC. T_m was found to be ~50 ºC.

Chapter 6

TOWARDS SYNTHETIC GENE-TARGETED HYPERMUTAGENESIS IN *E. COLI*

6.1 Abstract

The development of *in vivo* systems for mutating genes of interest with high rates and specificity would accelerate selection processes for directed evolution. Somatic hypermutagenesis is a process used in nature to diversify immunoglobulins in B lymphocytes, which is dependent on the enzyme activation induced cytidine deaminase (AID). We aimed to emulate this system in *E. coli*.

Although the molecular mechanisms for targeting AID to the immunoglobulin locus are poorly understood, we postulated that gene-targeted mutagenesis by AID could be engineered by coupling its activity to a gene-specific RNA polymerase (T7 RNA polymerase, T7RNAP). To test this hypothesis, we created a fusion protein between T7RNAP and AID, and compared its transcription and mutagenic activity to that of co-expressed T7RNAP and AID. We found that this T7RNAP-AID fusion protein could both transcribe and mutate a gene of interest, though at lower levels than co-expressed T7RNAP and AID. Using an *in vivo* mutation assay that could evaluate the specificity of mutagenesis, we found that the fusion protein mutated non-specific sites in the genome at a similar level as the co-expressed proteins, indicating that gene-targeted mutagenesis was not achieved.

These results suggest that locus-specific mutagenesis, such as somatic hypermutagenesis in the mammalian immune system, requires more than simply fusing transcription and mutation activities. Allosteric activation may be necessary to localize mutagenic activity to genes of interest.

6.2 Gene-targeted mutagenesis

A key limitation to the use of selection methods for engineering proteins *in vivo* is the necessity of an *in vitro* mutagenesis step to selectively incorporate mutations in the gene of interest[145]. This restricts the number of rounds of mutation to less than one round per day, which ultimately limits the time scale of directed evolution. Various strategies for *in vivo* mutagenesis have been developed using error-prone polymerases, viruses, and cell lines[145; 146; 147; 148; 149; 150; 151]; however, these methods have low mutation rates and selectivity.

In B lymphocyte diversification in the adaptive immune system, Nature has found a solution to the problem of targeted mutagenesis by having a process, known as somatic hypermutagenesis, that specifically introduces mutations into the immunoglobulin locus[152]. This process uses an enzyme known as activation induced cytidine deaminase to convert cytosine bases into uracils[153; 154]; subsequent DNA repair converts the resulting base mismatch into a spectrum of mutations. The molecular mechanisms behind AID-directed mutagenesis are still unclear[155; 156]. It has been reported that AID has increased activity at highly transcribed sites[157] and epigenetically accessible sites[158], is activated by phosphorylation[159], and is associated with proteins involved in transcription[160], suggesting possible mechanisms for gene-targeting.

E. coli has been used as a system to study AID biochemistry[161; 162]. Mammalian AID can be expressed in an active form in *E. coli* and broadly increases mutation rate throughout the genome by one to two orders of magnitude[162]. AID could be useful for *in vivo* selection applications in *E. coli* if its activity could be localized to a gene of interest. Increasing specificity may also have the effect of increasing its activity at sites of interest, which would also be desired for selections.

As a first step towards creating effective *in vivo* gene-targeted mutagenesis in *E. coli* and other organisms, we investigated the hypothesis that linking transcriptional and mutagenic activities could create targeted mutations at sites of interest. We created a fusion protein between AID and the RNA polymerase from T7 phage. T7RNAP specifically transcribes from genes under the T7 promoter, and the T7RNAP-AID fusion was tested for increased mutation rate at a gene under this promoter.

6.3 Coupling targeted transcription with mutation activity

We created five plasmids to test our hypothesis (**Table 6-1**). The plasmid 007a is a derivative of pET22b, and encodes the kanamycin-resistance gene Npt2 under the T7 promoter. The plasmid 007b contains the Npt2 gene with a L94P mutation, which can be reverted by C to T transversions commonly introduced by AID-directed mutagenesis. Selections on kanamycin can therefore test for both transcription and mutation of Npt2.

The plasmid 008b is a derivative of pLysS and contains T7 RNAP under the tet promoter. This vector served as the negative control, possessing only transcriptional activity at the T7 promoter (*i.e.* Npt2 from 007b). The plasmid 012a is derived from 008b, and contains the mouse AID gene (reported to express in active form in *E. coli*[162]) fused to the N-terminus of T7 RNAP, with a 10-amino acid linker between the two genes. The plasmid 012c is also derived from 008b, and contains AID and T7RNAP in an operon.

We first verified that each T7RNAP construct with and without AID could productively transcribe from the T7 promoter by co-transforming plasmids 008b, 012a, and 012c with 007a into *E. coli* XL1-blue. Each of these pairs of genes was able to allow XL1-blue to grow on kanamycin.

Next, mutation rate assays were carried out by co-transforming 008a, 012a, and 012c with 007b into *E. coli* CJ236. This strain has defective uracil-*N*-glycosylase and deoxyuridine triphophatase activities, which results in defective removal of uracil bases, and consequently higher DNA mutation rates when uracil is incorporated into the genome[163]. Gene-targeted mutagenesis was measured by finding the number of revertants to the Npt2 L94P mutation, which confer kanamycin resistance. Non-specific mutagenesis was measured by finding the number of mutations at various sites in the distal RpoD gene, which are known to confer rifampicin resistance[164]. To account for the

Table 6-1. **Constructs used in this study**.

Vector	Expressed gene	Derived from	Purpose
007a	Npt2	pET22	Assay for transcription
007b	Npt2 L94P	pET22b	Assay for selective mutagenesis
008b	T7RNAP	pLysS	Negative control for mutation rate
012a	T7RNAP-AID fusion	pLysS	Test for gene-targeted mutator
012c	T7RNAP and AID	pLysS	Positive control for mutation rate

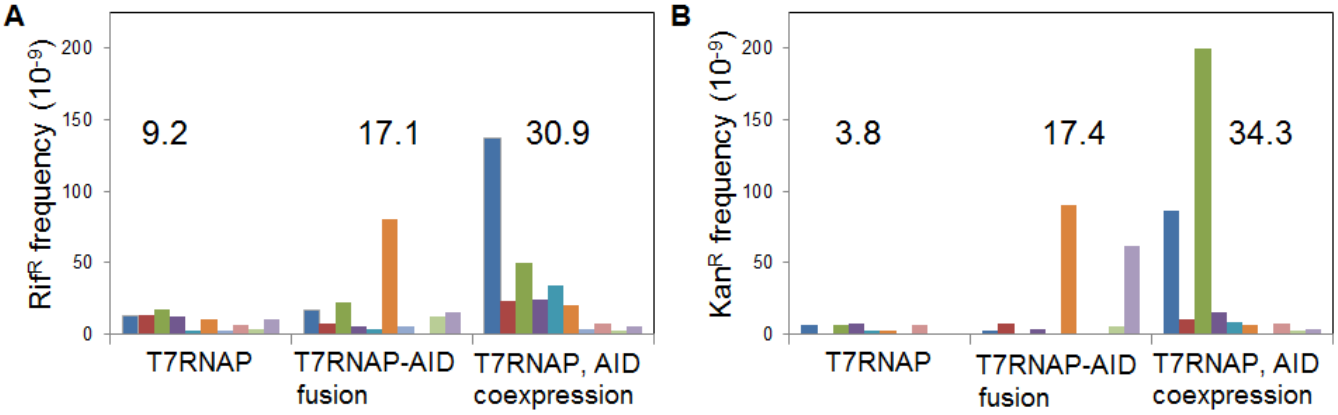


Figure 6-1. **General and targeted mutation rates of fused and co-expressed T7RNAP and AID.** A) General mutation rate assessed by acquisition of rifampicin resistance by mutation at the RpoD gene. B) Targeted mutation rate assessed by acquisition of kanamycin resistance by mutation at the Npt2 L94P gene. T7RNAP was from the 008A construct, T7RNAP-AID fusion was from 012a, and T7RNAP, AID coexpression was from 012c. Average mutation rates are noted.

stochastic and highly variable nature of mutation rate, 10 independent replicates were carried out for each construct.

Results for this mutation rate assay are shown in **Figure 6-1**. We observed that the T7RNAP-AID fusion expressed from 012a can increase mutation rate, both at the target Npt2 gene expressed from the T7 promoter and at distal sites (RpoD). This mutation rate is approximately half of that observed when AID is expressed separately from T7RNAP, indicating either decreased expression or decreased activity in the fusion protein. Notably, the ratio of general to specific mutation rate was unchanged between the fusion and co-expression constructs, indicating that fusing AID to T7RNAP could not restrict mutagenesis activity to genes under the T7 promoter.

6.4 Discussion

These results indicate that simple co-localization of AID with transcription complexes may be insufficient to direct mutations to genes of interest. These results are not completely unexpected, since studies on mammalian somatic hypermutagenesis have found a variety of factors that may contribute to AID specificity[155]. Creating locus-specific mutagenesis in *E. coli* will likely require a more complex system than the one implemented here.

The ratio of general to specific mutation in the T7RNAP-AID fusion gene was virtually identical to that of the co-expressed genes. One possible reason is that although T7RNAP only transcribes at the T7 promoter, the polymerase may be found throughout the genome as it searches for its promoter[165]. The RpoD gene that served as control for non-specific mutagenesis encodes the major sigma factor in *E. coli*[164] and has high constitutive expression. This likely allowed AID to access the DNA and cause the observed mutations when the fused T7RNAP was in the vicinity of the RpoD gene.

One way to increase mutation specificity would be to engineer allosteric regulation into AID, such that it only activates when it is bound to the gene of interest. Reports suggest that AID is regulated in this manner in somatic hypermutagenesis, possibly through interactions with replication protein A or the protein kinase A alpha regulatory subunit[155]. Creating allosteric control *de novo* is a challenging problem in protein engineering, but progress has been made to link conformational changes in one protein to another[166].

AID activity is highest on single-stranded DNA[162; 167], which occurs not only at transcriptional bubbles but also at resection events following DNA double-strand breaks. Indeed, engineering an artificial endonuclease I-*Sce*I cut site at a locus of interest was reported to increase AID activity at the locus by approximately four-fold [168]. While this improvement on its own is insufficient for the gene-targeted mutagenesis desired here, combining multiple approaches with small improvements may ultimately result in a viable method.

6.5 Methods

General methods are described in Appendix 1.

Strains and cloning

E. coli XL1-blue (Novagen) was used for cloning and *E. coli* CJ236 FΔ*(HindIII)::cat* (Tra$^+$ Pil$^+$ Cam^R)/ *ung-1 relA1 dut-1 thi-1 spoT1 mcrA* (New England Biolabs) was used for mutation assays. Standard methods for DNA isolation and manipulation were performed as described by Sambrook et al[115]. pDev vectors were created using Gibson assembly[67]. Plasmid encoding the AID gene and the Npt2 L94P gene were received as gifts from Ramiro Almudena.

Mutagenesis assays

An assay for determining selective *in vivo* mutagenesis was adapted from a protocol by Coker *et al.*[164] Briefly, 10 colonies per genotype per condition were grown to an OD600 of ~0.5 at 37 ºC with 250 rpm shaking in LB + 50 µg/mL carbenicillin (for pDev007b selection) + 35 µg/mL chloramphenicol (for pDev008b, 012a, and 012c selection). Cultures were diluted 1:1 in LB+ 50 µg/mL carbenicillin + 35 µg/mL chloramphenicol + 1 mM IPTG + 200 ng/mL anhydrotetracycline, and grown for 4 h at 37 ºC with 250 rpm shaking. Each culture was plated on three different agar plates: LB + 100 µg/mL rifampicin + 50 µg/mL carbenicillin + 35 µg/mL chloramphenicol (250 µL of culture), LB + 50 µg/mL carbenicillin + 35 µg/mL chloramphenicol + 50 µg/mL kanamycin + 1 mM IPTG + 200 ng/mL anhydrotetracycline (250 µL of culture), and LB + 50 µg/mL carbenicillin + 35 µg/mL chloramphenicol (20 µL of 1:100 diluted culture). Plates were grown overnight at 37 ºC, and colonies were counted the next morning.

APPENDIX I: GENERAL MATERIALS AND METHODS

A1.1 Chemicals and commercial kits

Media

Yeast extract, peptone, tryptone, casamino acids, and yeast nitrogen base were purchased from BD Biosciences (Franklin Lakes, NJ). SD-Ura powder was purchased from MP Biomedicals (Santa Ana, CA). LB powder and TB powder was purchased from RPI Corp (Mount Prospect, IL). Sodium chloride, D-glucose, ampicillin, tetracycline, chloramphenicol, and kanamycin were purchased from Sigma Aldrich (St. Louis, MO).

Buffers

Except where otherwise noted, chemicals were purchased from Sigma-Aldrich (St Louis, MO). Sodium phosphate monobasic and sodium phosphate dibasic heptahydrate were purchased from Mallinckrodt Chemicals (St. Louis, MO). Distilled deionized water was obtained from a Mega-Pure water distillation system (Corning, NY).

Cloning

Taq DNA polymerase was purchased from Roche Applied Science (Penzberg, Germany). Phusion high-fidelity DNA polymerase, *Taq* DNA ligase, T4 DNA ligase, and endonucleases NheI, BamHI, XhoI, and NdeI were purchased from New England Biolabs (Ipswich, MA). Sybr gold was purchased from Invitrogen (Carlsbad, CA). Oligonucleotides were purchased from Integrated DNA Technologies (San Diego, CA, USA). DNA sequencing was performed by Retrogen (San Diego, CA, USA). T5 exonuclease was purchased from Epicentre Biotechnologies (Madison, WI). Molecular biology grade H_2O was purchased from Sigma-Aldrich.

QIAprep Miniprep Kit and QIAquick Gel Extraction Kit were purchased from Qiagen (Venlo, Limburg). Frozen-EZ Yeast Transformation II Kit and Zymoprep Yeast Plasmid Miniprep II Kit were purchased from Zymo Research (Irvine, CA).

A1.2 Laboratory equipment

96-well absorbance measurements were taken in a Tecan Infinite M200 (Mannedorf, Switzerland). Centrifugation was carried out in an Allegra 25RCentrifuge (Beckman Coulter, Pasadena, CA). Protein purification was carried out in an AKTAxpress protein purifier (GE Life Sciences, Pittsburgh, PA). PCR and thermostability assays was carried out in an Eppendorf Mastercycler (Hamburg, Germany). Protein and DNA absorbance readings were read on a NanoVue Plus spectrophotometer (GE Life Sciences). Cultures were grown in a Multitron II incubated shaker (Infors HT, Basel, Switzerland). 96-well manipulations were carried out with a Multimek 96 liquid handler (Beckman Coulter, Pasadena, CA).

A1.3 Molecular cloning

Primer design

Simple heuristics were used to design primers for gene amplification, sequencing, and overlap PCR mutagenesis. Primers ranged in length from 20 base pairs (typical for sequencing), to up to 60 base pairs (for overlap PCR). The 5’ and 3’ ends of the primers were either G or C, and GC content was between 40-60%. When mutations were to be introduced, at least 12bp of nucleotides were placed at either side to facilitate annealing. When restrictions sites were desired, an extra random six nucleotides were added 5’ or 3’ of the restriction site (*e.g.*ATGCTA).

PCR was optimized by following the rule: start with an annealing temperature of 57 °C, and if no product is found, reduce the annealing temperature in steps of 2 °C; if multiple products are found, increase the annealing temperature in steps of 2 °C.

Error-prone PCR

PCR mix contained 1 µL plasmid (2ng/uL), 2 µL forward primer (50 µM), 2 µL reverse primer (50µL), 4 µL dNPT (10mM), 10 µL Taq buffer (10x), 28 µL $MgCl_2$, 1.6 µL Taq polymerase (5 u/µL), X µL $MnCl_2$ (1 mM), in a total volume of 100 µL molecular biology grade H_2O. X ranged from 10 to 35 µL, to give a final $MnCl_2$ concentration of 100 µM to 350 µM.

PCR was run in an Eppendorf Mastercycler, first at a temperature of 95 ºC for 5 min, followed by 30 cycles of 57 ºC for 30 sec, 72 ºC for 1 min/kb, and 95 ºC for 30 sec. This was followed by 72 ºC for 10 min and then 4 ºC indefinitely. PCR product was gel purified, eluting in 50 µL elution buffer, and then digested and ligated as described below.

gBlock design and assembly

500 bp "gBlocks" were synthesized by Integrated DNA Technologies (Coralville, IA). After designing genes with Gene Designer[66], they were divided into two 500 bp blocks, with 30-40 bp overlap between each block. Primer design heuristics noted above were used for these overlap regions. Primers were also designed to amplify the entire assembly using overlap PCR, and to add homology sites for Gibson cloning into the vector of interest. Primers were often made longer if the gene length was slightly over 1000 to keep number of blocks at a minimum.

Overlap PCR

Primers were designed as described above, with a pair for gene amplification, and a pair for each mutated site (sometimes one pair of primers could suffice for two or more very close sites). This results in n + 1 fragments, where n is the number of mutated sites.

Fragments were amplified in a PCR mixture of 4 µL Phusion buffer, 0.4 µL dNTP (5mM of each), 0.2 µ Pfu Turbo polymerase, 1 µL plasmid (30ng/µL), 1 µL forward primer (50 µM), 1 µL reverse primer (50 µM), in 12.4 µL molecular biology grade H_2O. PCR was carried out as follows:

98 ºC for 30 sec; 32 cycles of 98 ºC for 30 sec, 57 ºC for 30 sec, and 72 ºC for 30 sec/kb; 72 ºC for 4 min; and 4 ºC indefinitely.

Fragments were gel purified and eluted in 25 uL EB. Fragments were annealed by PCR, in a mixture of 4 µL Phusion buffer, 0.4 µL dNTP (5mM), 0.2 µL Pfu Turbo polymerase, 1 µL each fragment, in 13.4 µL H_2O. PCR reaction was carried out as before, but only 12 cycles. The annealed fragments were then amplified, in a PCR mixture of 4 µL Phusion buffer, 0.4 µL dNTP, 0.2 µL Pfu Turbo polymerase, 0.4 µL PCR product, 0.4 µL forward primer, and 0.4 µL reverse primer, in 14.2 µL H_2O. The PCR reaction was carried out as before, with 32 cycles. Product was gel purified, eluting in 25 µL EB.

PCR product was digested and restriction digested, or Gibson assembled into vector (depending on available homology or restriction sties), as described below. 5 µL of final product was transformed into chemically competent XL1-blue *E. coli*.

Restriction digestion

Digestions were performed with restriction enzymes (NheI, NdeI, XhoI, BamHI, depending on vector) at a dilution of 1/40, NEB Buffer 4 at a dilution of 1/10, and BSA at a dilution of 1/100. Mixtures were incubated at 37 ºC for 4 h. Total volumes ranged from 10 µL to 100 µL, depending on amount and concentration of DNA.

Ligation

Vector and insert were combined in a 1:5 molar ratio. T4 ligase buffer was added at a 1/10 dilution and T4 ligase was added at a 1/20 dilution. Reaction was incubated at either 16 ºC overnight (for libraries), or 25 ºC for 1 h (for general cloning).

Gibson assembly

DNA fragments with homology lengths of 30-40 bp were joined using one-step isothermal Gibson assembly[67]. Where possible, linear plasmid fragments were created by restriction enzyme digestion; otherwise plasmid was amplified by PCR.

Gibson master mix consisted of 320 µL 5x isothermal reaction buffer (1.5 mL 1 M Tris buffer pH 7.5, 75 µL 2M $MgCl_2$, 120 µL 400 mM dNTP, 150 µL 1M DTT, 0.75 g PEG8000, 150 µL 100mM NAD, in 3 mL sterile ddH_2O) with 6.4 µL of 1 U/µL T5 exonuclease, 20 µL of 2 U/µL Phusion polymerase, 160 µL of 40,000 U/µL *Taq* DNA ligase, and 694 µL sterile ddH_2O. To assemble DNA fragments, equimolar concentrations of DNA fragments in 5 µL total volume were added to 15 µL Gibson master mix and incubated at 50 ºC for 45 min. 5 µL of the mixture was transformed directly into 50 µL of chemical competent *E. coli* and plated on selective medium.

Media preparation

All media was sterilized by autoclaving at 120 ºC for 20 min. For *E. coli* growth, LB medium was prepared by combining 5 g of yeast extract, 10 g of NaCl, and 10 g of tryptone in 1 L distilled deionized water. 2xYT medium was prepared by combining 10 g of yeast extract, 5 g of NaCl, and 20 g of tryptone in 1 L distilled deionized water. SOB media was made by combining 20 g tryptone, 5 g yeast extract, 0.5 g of salt, and 10 mL of 250 mM KCl in a total volume of 1 L distilled deionized water, and adjusted to pH 7.0. Before use, 5 mL of sterile 2 M $MgCl_2$ is added.

For *S. cerevisiae* growth, SD-Ura medium was prepared by dissolving SD-Ura powder in 1 L distilled deionized water. YPD medium was prepared by combining 10 g yeast extract and 20 g peptone in 900 mL distilled deionized water. After autoclaving, 100 mL 20% glucose was added.

Agar plates were made by adding 1.5 % w/v agar for *E. coli* plates, and 2.0% w/v agar for *S. cerevisiae* plates.

Chemically competent E. coli

Chemically competent DH5α and XL1-blue *E. coli* were prepared using the Inoue method[169]. A single bacterial colony was inoculated into 25 ml of LB broth in a 250 mL flask. The culture was incubated for for 6-8 h at 37 ºC at 250 rpm. This culture was then inoculated into three 1 L flasks, each containing 250 ml of SOB media; the first flask receives 10 mL of starter culture, the second 4 mL, and the third 2 mL. The flasks were incubated overnight at 20 ºC at 250 rpm.

The next morning, the OD600 of all three cultures was taken, and the culture closest to but no greater than 0.55 was grown to that value. The culture was transferred to an ice water bath for 10 min, and then cells were harvested by centrifugation at 2500 g for 10 min at 4 ºC. The cells were resuspended in 80 mL of ice-cold Inoue transformation buffer (55 mM $MnCl_2$, 15 mM $CaCl_2$, 250 mM KCl, 10 mM PIPES (0.5M, pH 6.7)). Cells were centrifuged again, and resuspended in 20 mL transformation buffer. 1.5 mL DMSO was added, and cells were incubated on ice for 10 min.
50 µL of cells were dispensed as aliquots into 1.7 mL eppendorf tubes and snap-frozen in liquid nitrogen. Aliquots were stored at -80 ºC.

For transformation, 5 µL of DNA was added to competent cells on ice and mixed gently. Tubes were stored on ice for 30 min, and then transferred to a 42 ºC water bath for 90 sec. Cells were then transferred to ice for 2 min. Cells were transferred directly to pre-warmed LB agar + antibiotic.

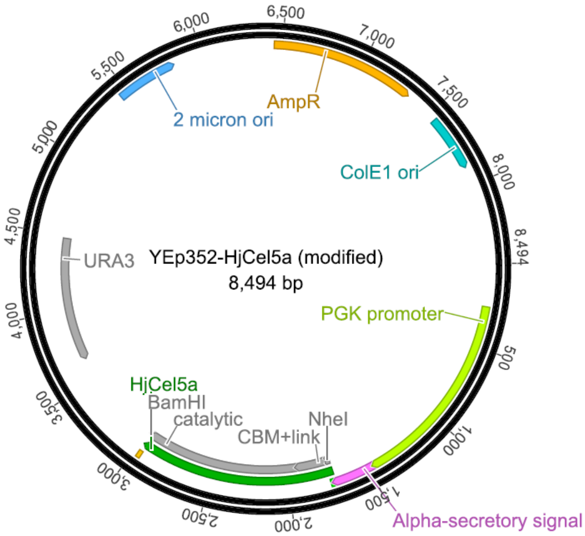
A1.4 Protein purification

His-tagged proteins were purified with HisTrap HP columns (GE Healthcare, Little Chalfont, UK). Except where otherwise noted, binding buffer was 20 mM Tris pH 8.0, 100 mM sodium chloride, and 10 mM imidazole, and elution buffer was 20 mM Tris pH 8.0, 100 mM sodium chloride, and 300 mM imidazole. After sample loading, columns were washed with 5 column volumes of binding buffer. A linear gradient of 0-80% elution was used to elute protein.

APPENDIX 2: SEQUENCES AND ALIGNMENTS

CHAPTER 2:

Yep352 vector used in this study (containing HjCel5a):



Gene sequences- Wild-type Cel5a

Hypocrea jecorina Cel5a Cellulose Binding Module (CBM) + linker (appended to N-terminus):

```
GCTAGCCAACAAACAGTATGGGGTCAATGTGGTGGTATTGGATGGTCTGGTCCGACAAACTGTGCT
CCAGGCTCGGCATGTTCGACACTAAATCCATATTACGCTCAATGTATCCCTGGCGCTACCACTATA
ACAACTTCTACTAGACCACCTTCTGGTCCGACGACAACTACAAGGGCTACCTCAACCTCTTCCTCT
ACACCCCCTACTTCCAGC
```

Linker+6xHis-tag+Stop codon (appended to C-terminus):

```
GGAGGTAGCGGAAGCGGACACCACCACCACCACCACTAA
```

H. jecorina Cel5a:

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACTTGT
GTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGAAGACGGTATGACAATATTTCGTTTGCCTGTTGGATGGCAA
TATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
AATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCCGCT
```

```
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGGTGCCTTCAGC
CCGTTGGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGTAAT
GTACAAAGTTGTATTCAGGATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TCCAGTGGTAATAGTTGGACCGATACTTCTTTGGTATCTTCTTGCTTAGCTAGAAAG
```

Phialophora G5 Cel5a:

```
GGAAGGACACGCTTTGCTGGTGTTAACATAGCTGGATTTGATTTTGGTTGTGCTACCGATGGTACC
TGTAACACCACGGCTGTTTATCCGCCGGTTAAAGATATGCCCCCATACTATAATAACCCTGATGGT
GCAGGACAAATGGATCATTTTAGTAAGGATGATAACTTAAATATTTTTCGTTTGCCAGTTGGTTGG
CAATATCTGGTGAACTCTAACTTAGGTGGTACCCTTGACTCAACGAACTTAGGCTATTACGATCAA
CTTGTTCAATCATGTCTGTCAACCGGAGCTTATTGTATTGTAGATATCCATAATTACGCTCGTTGG
AATGGCGCCATAATAGGCCAAGGTGGACCAACAAACGAACAATTTGTTTCTGTTTGGACACAACTC
GCGACTAAGTATGCTTCACAAGCCAGGGTGTGGTTTGGTATTATGAACGAGCCACATGATGTTCCA
TCTATCACCACATGGGCTGCAACAGTTCAAGCTGTGGTGACAGCCATTAGAAATGCCGGCGCCACG
AGTCAATTCATCTCTCTCCCTGGCAACGACTGGCAATCAGCAGCCGCGGTCATCTCCGATGGTTCT
GCCGCCGCTCTTAGCACGGTCACAAATCCAGATGGCACTACGACAAACTTGATATTTGATGTTCAC
AAATATCTGGACTCAGATAACTCAGGTACTCACACTGAGTGTGTCACTAACAATATTGATGATGCA
TTTGCTCCTTTAGCGACGTGGTTGAGGCAGAATGGAAGACAGGCTATATTGACGGAAACAGGAGGA
GGGAACACTGCATCTTGTGAGACATACTTATGTCAACAAATTGCGTATCTGAATGCTAACGCCGAT
GTGTACTTAGGTTACGTTGGTTGGGGGGCTGGTTCTTTTGACAGCACGTATGCATTAGACGAAACA
CCTACAGGTTCAGGTTCAAGTTGGACCGACACCCCTCTGGTTAAGGCGTGCATTGCAAGGAGCTCT
```

Penicillium decumbens Cel5a:

```
GGTAAAGTGAGATTCGCTGGAGTAAATATCGCTGGTTTCGACTTCGGCGTTGTTATTTCAGGTACC
CAAGATATGACTCAAATAGTAGATGAAAGTGTTGATGGGGTCAACCAAATGCGTCATTTTGTAAAC
GATGATGGTTTTAACATCTTCAGATTACCTAGTGGCTGGCAATTCTTGGTCAACAACAATTTGGGT
GGTTCTCTGGACTCAAACAATTTCGCCAAGTATGATAAATTGGTCCAAGGTTGTCTGTCGCTAGGC
GCCTATTGCATTGTCGACGTCCATAATTACGCAAGATGGAATGGCGGTGTAATAGGTCAAGGCGGT
CCGACCGACGACCAGTTTACATCACTGTGGACTCAACTTGCCACTCATTATAAGAGTGAGTCAAAG
ATAATTTTTGGGGTGATGAATGAGCCTCATGATCTTGACATTAATCGTTGGGCAACTACTGTTCAA
AAGGCTGTGACAGCCATAAGAAAGGCTGGTGCAACCAGTCAAATGATCCTATTGCCTGGTACCGAT
TTCACTAGTGCGGCGAATTTTGTCGAGAATGGTTCTGGAGCTGCACTGAGCGCTGTCACTAATTTG
GATGGTAGCACGACTAACCTAATTTTCGACGTCCACAAATACTTGGATTCTGACAACTCTGGAACG
CATGCGGAGTGTGTTACCAACAATGCAGATGCTTTCAACAGCCTTGCTCAATGGCTTAGAACGAAC
AAAAGACAGGCAATGCTTACTGAGACAGGCGGTGGTAACGTTCAATCTTGCGGAACATATATGTGT
CAACAATTGGACGTCCTGAATCAAAACAGCGATGTTTATTTAGGTTGGACAAGTTGGAGTGCTGGC
GGGTTCCAAGTTTCGTGGAATTATGTTTTAGGCGAAGTGCCAACTAATAATGTAGATACTTATTTG
GTCAAACAATGTTTTGTTCCAAAATGGAAGAAT
```

Penicillium pinophilum Cel5a:

```
GGGAAAGTTCAGTTCGCAGGGGTAAATATTGCGGGTTTTGATTTTGGCATGGTAACATCGGGTACT
CAAGACCTAACTCAGATTGTTGATGAAAGTGTAGATGGTGTAACGCAGATTAAACATTTCGTTAAT
GATGACACTTTCAACATGTTTAGACTCCCTACCGGGTGGCAATATTTAGTAAACAATAACCTTGGT
GGTCAACTCGATGCAACTAACTTTGGCCAATACGATAAACTTGTTCAAGGATGTCTAAGTACAGGT
```

```
GCACATTGTATAGTGGACATTCATAATTACGCCAGATGGAATGGAGCGATTATTGGCCAAGGTGGT
CCATCAGACGCTCAATTCGTTGATTTATGGACTCAATTGGCGACTAAGTACAAAGCCGATTCTAAA
GTTGTTTTCGGGGTTATGAATGAGCCTCATGATTTGACCATAAGTACATGGGCTGCCACTGTCCAA
AAAGTAGTCACTGCCATTCGCAACGCAGGAGCTACTTCCCAAATGATTCTACTCCCGGGTACGGAT
TACACATCTGCTGCTAACTTTGTTGAGAATGGCAGTGGCGCGGCACTAGCTGCCGTTGTTAATCCA
GATGGAAGTACACATAACCTGATATTCGATGTGCACAAGTACTTGGATAGCGACAACTCAGGTACG
CACGCTGAATGCGTAACGAATAATGTTGATGCATTTTCATCCTTAGCTACATGGTTGAGAAGCGTC
GGGAGACAAGCACTGCTTTCCGAAACTGGAGGTGGCAACGTTCAAAGTTGTGCAACCTACATGTGT
CAACAGTTAGATTTTTTAAACGCAAATTCTGATGTCTATTTGGGGTGGACATCGTGGTCCGCCGGG
GGTTTTCAGGCTTCTTGGAATTACATATTGACTGAAGTACCAAATGGAAACACTGATCAGTATCTA
GTTCAACAGTGTTTTGTACCAAAGTGGAAATCC
```

Gene alignment for recombination

```
>HjCel5a
G-VRFAGVNIAGFDFGCTTDGTCVTSKVYPPLKNFTGSNNYPDGIGQMQHFVNEDGMTIFRLPVGW
QYLVNNNLGGNLDSTSISKYDQLVQGCLSLGAYCIVDIHNYARWNGGIIGQGGPTNAQFTSLWSQL
ASKYASQSRVWFGIMNEPHDV-NINTWAATVQEVVTAIRNAGATSQFISLPGNDWQSAGAFISDGS
AAALSQVTNPDGSTTNLIFDVHKYLDSDNSGTHAECTTNNIDGAFSPLATWLRQNNRQAILTETGG
GNVQSCIQDMCQQIQYLNQNSDVYLGYVGWGAGSFDST—YVLTETPTSSGNSWTDTSLVSSCLARK
---
>PgCel5a
GRTRFAGVNIAGFDFGCATDGTCNTTAVYPPVKDMPPYYNNPDGAGQMDHFSKDDNLNIFRLPVGW
QYLVNSNLGGTLDSTNLGYYDQLVQSCLSTGAYCIVDIHNYARWNGAIIGQGGPTNEQFVSVWTQL
ATKYASQARVWFGIMNEPHDVPSITTWAATVQAVVTAIRNAGATSQFISLPGNDWQSAAAVISDGS
AAALSTVTNPDGTTTNLIFDVHKYLDSDNSGTHTECVTNNIDDAFAPLATWLRQNGRQAILTETGG
GNTASCETYLCQQIAYLNANADVYLGYVGWGAGSFDST—YALDETPTGSGSSWTDTPLVKACIARS
--S
>PdCel5a
GKVRFAGVNIAGFDFGVVISGTQDMTQI---------VDESVDGVNQMRHFVNDDGFNIFRLPSGW
QFLVNNNLGGSLDSNNFAKYDKLVQGCLSLGAYCIVDVHNYARWNGGVIGQGGPTDDQFTSLWTQL
ATHYKSESKIIFGVMNEPHDL-DINRWATTVQKAVTAIRKAGATSQMILLPGTDFTSAANFVENGS
GAALSAVTNLDGSTTNLIFDVHKYLDSDNSGTHAECVTNNAD-AFNSLAQWLRTNKRQAMLTETGG
GNVQSCGTYMCQQLDVLNQNSDVYLGWTSWSAGGFQVSWNYVLGEVPTNN----VDTYLVKQCFVP
KWKN
>PpCel5a
GKVQFAGVNIAGFDFGMVTSGTQDLTQI---------VDESVDGVTQIKHFVNDDTFNMFRLPTGW
QYLVNNNLGGQLDATNFGQYDKLVQGCLSTGAHCIVDIHNYARWNGAIIGQGGPSDAQFVDLWTQL
ATKYKADSKVVFGVMNEPHDL-TISTWAATVQKVVTAIRNAGATSQMILLPGTDYTSAANFVENGS
GAALAAVVNPDGSTHNLIFDVHKYLDSDNSGTHAECVTNNVD-AFSSLATWLRSVGRQALLSETGG
GNVQSCATYMCQQLDFLNANSDVYLGWTSWSAGGFQASWNYILTEVPNGN----TDQYLVQQCF
VPKWKS
```

Gene sequences- Initial chimera test set

00012032:

```
GGTAAGGTACGGTTTGCAGGTGTTAACATCGCAGGCTTTGATTTCGGTTGCACAACCGATGGTACT
TGTGTCACTTCCAAAGTTTATCCCCCATTAAAAAATTTCACAGGTTCAAACAACTATCCAGACGGC
```

```
ATAGGTCAAATGGACCATTTCTCGAAAGATGACGGTTTTAACATTTTCAGACTTCCTGTTGGATGG
CAATACTTAGTGAATAACAACCTGGGTGGCAATCTGGACAGCACAAGTATTTCAAAGTACGATCAA
CTAGTTCAGGGTTGTCTTTCTACTGGAGCTTACTGCATTGTTGATATTCATAACTACGCCAGATGG
AATGGTGGTGTTATTGGCCAAGGTGGTCCAACCAATGCTCAATTTACCTCATTATGGTCGCAATTG
GCATCCAAGTATAAATCTGAGTCGAAAATTATTTTTGGCGTGATGAACGAACCCCATGATGTAAAC
ATTAACACTTGGGCTGCAACCGTTCAAGAAGTCGTTACAGCTATAAGAAACGCAGGTGCCACATCT
CAAATGATCCTGCTCCCAGGGAACGATTGGCAATCGGCCGGTGCTTTTATTTCCGATGGTTCGGCT
GCTGCTTTATCGCAAGTAACGAATCCGGACGGGTCTACAACAAACTTAATCTTCGATGTTCATAAA
TACCTGGACAGCGATAATTCAGGAACCCATGCTGAATGTGTTACAAATAATATCGACGGAGCATTC
TCACCTTTAGCCACTTGGTTGAGAACAAACAAAAGACAAGCAATGCTAACAGAAACCGGTGGAGGA
AACGTGCAGTCCTGTGCCACCTATATGTGTCAGCAATTAGACGTTTTAAATCAGAATAGTGATGTC
TATCTGGGTTGGACTTCATGGTCTGCTGGTTCTTTCCAAGCTTCGTACATACTAACAGAGGTACCT
ACCGGCTCCGGTAGTAGTTGGACGGATCAATATTTGGTTCAGCAATGTTTTGTACCAAAATGGAAG
AAC
```

00031021:

```
GGTCGCACTAGATTTGCCGGTGTTAACATAGCAGGGTTTGATTTTGGCTGTACCACAGATGGAACT
TGCGTCACTTCGAAAGTTTACCCACCCCTAAAGAATTTTACTGGAAGCAATAATTATCCTGATGGT
ATCACTCAGATTAAGCACTTTGTTAATGACGATAATCTAAATATTTTCAGATTACCAGTAGGGTGG
CAATACTTGGTCAATAATAACTTAGGCGGCAATCTGGATTCTACAAGTATTTCTAAGTATGACCAG
CTTGTCCAGGGTTGTTTGTCAACTGGCGCTTACTGTATTGTGGATATACACAACTATGCAAGATGG
AATGGTGCTATCATAGGCCAAGGTGGCCCAACAAATGCGCAATTTACTTCATTGTGGTCTCAGCTT
GCTTCCAAATACGCTTCCCAAGCTCGGGTATGGTTTGGTATTATGAATGAACCACACGATGTCAAT
ATTAACACCTGGGCTGCGACCGTGCAGGAAGTTGTTACAGCTATCAGAAACGCAGGGGCTACATCA
CAATTCATTTCACTTCCAGGTAATGATTGGCAATCAGCTGGCGCTTTCATTTCTGACGGTAGCGCC
GCCGCGTTAAGTCAAGTGACTAACCCTGATGGTTCAACTACAAACTTAATATTCGATGTGCATAAG
TACCTGGATTCAGATAACTCCGGAACTCACGCTGAATGCGTGACTAATAATATAGATGGGGCCTTT
TCGCCTCTAGCTACATGGCTGAGACAAAACGGAAGGCAAGCTATTTTAACTGAAACTGGTGGCGGG
AACGTACAGAGTTGTGGAACTTACATGTGTCAACAAATTGCATATTTAAACGCGAATGCCGATGTT
TATTTGGGGTACGTTGGTTGGAGCGCTGGCTCTTTTCAAGTCTCTTGGAACTATGTTTTAGGTGAA
GTCCCTAACGGCAATACAGATACTTATCTCGTAAAACAATGTATCGCCCGTTCCTCT
```

01200030:

```
GGCGTCCGATTTGCTGGGGTAAACATAGCCGGATTTGACTTTGGATGCGCTACTGATGGCACTTGT
AACACAACTGCTGTCTATCCCCCTCTGAAAAATTTTACCGGGAGCAACAATTACCCTGATGGTATC
GGCCAAATGCAACACTTCGTTAATGAAGATGGTATGACTATTTTTAGGTTGCCAGTAGGTTGGCAA
TATTTGGTGAATAGTAACCTTGGTGGCAATCTAGACTCCACAAACCTCGGAAAGTATGATCAACTG
GTGCAGGGATGCTTGTCCCTGGGTGCTTACTGCATTGTTGACATACACAACTATGCCAGGTGGAAT
GGCGGGATTATTGGCCAAGGAGGTCCTACTGATGATCAGTTCACCTCACTCTGGACCCAGCTTGCA
ACCAAATACGCATCACAATCGAGAGTTTGGTTCGGCATTATGAATGAGCCGCACGATGTCAATATA
AATACATGGGCAGCCACAGTACAAAAGGCTGTTACAGCGATAAGAAAGGCTGGAGCAACGTCGCAA
TTTATTTCGTTACCCGGTAATGATTGGCAATCAGCGGGCGCTTTCATATCAGATGGTAGTGCTGCG
GCTCTGAGTGCCGTGACTAATTTAGATGGCTCAACTACAAATTTAATTTTTGATGTGCACAAGTAC
TTAGATTCCGATAACAGCGGTACCCACGCTGAATGTGTCACAAATAACATCGACGGGGCCTTTTCT
CCATTGGCCACCTGGTTAAGACAGAATAATCGCCAGGCTATCCTAACTGAAACAGGTGGTGGTAAT
GTGCAAAGTTGTGCCACATACATGTGTCAGCAAATTCAATACCTAAACCAAAATTCAGACGTTTAC
```

```
TTGGGTTATGTAGGATGGAGTGCCGGGTCATTTCAGGCTTCCTATATACTAACTGAAGTTCCAACG
TCCTCCGGCAATTCCTGGACAGACCAATACTTGGTCCAACAATGTTTGGCTCGCAAA
```

03011110:

```
GGTGTTAGATTCGCCGGAGTCAATATCGCTGGATTTGATTTTGGTATGGTAACCAGTGGTACCCAA
GATCTGACTCAGATTTACCCTCCCTTAAAGAATTTCACTGGCTCAAATAATTACCCAGACGGTATC
GGACAAATGGATCATTTTTCAAAAGATGACGGCATGACTATCTTTCGGTTACCAACAGGTTGGCAA
TATTTAGTTAATAATAATTTGGGTGGTAATTTAGACGCTACGAATTTCGGTAAGTATGATCAATTA
GTTCAAGGATGTTTGTCGACCGGTGCATATTGCATTGTTGACATACATAATTACGCGCGCTGGAAT
GGTGCCATCATTGGTCAGGGAGGACCTACCAATGCTCAATTTACATCGTTATGGTCCCAGTTAGCC
TCAAAATATGCTTCGCAGGCCAGGGTATGGTTTGGTATTATGAACGAGCCTCATGATGTCTCGATC
ACTACCTGGGCAGCTACAGTTCAAGAAGTGGTTACTGCCATACGTAATGCCGGGGCGACTTCACAG
TTCATATCTTTACCTGGTAATGACTGGCAATCAGCCGCAGCCGTTATATCTGACGGGTCAGCTGCT
GCGTTGTCCCAAGTTACAAATCCTGATGGTTCAACGACAAATTTGATATTTGACGTGCATAAATAC
TTGGATTCAGATAATTCCGGCACTCACACAGAATGCGTCACGAACAATATTGACGATGCATTTGCC
CCTTTGGCAACTTGGTTGAGGCAAAATGGCCGTCAAGCGATCTTGACCGAAACAGGTGGTGGTAAC
ACGGCCAGCTGTGAGACGTATCTTTGTCAACAGATCCAGTACCTAAATCAAAATTCTGACGTTTAT
TTAGGATACGTTGGATGGGGTGCTGGTTCCTTTGATTCAACATACGCATTAGACGAAACGCCGACG
GGGTCGGGGAGCTCTTGGACCGACACCCCATTAGTTAAGGCTTGT
```

03110301:

```
GGGAGAACGAGATTTGCCGGTGTTAATATCGCAGGCTTTGACTTTGGAATGGTTACGTCCGGTACA
CAAGATCTAACACAAATTTATCCTCCATTGAAGAATTTTACCGGTTCAAATAACTATCCAGATGGT
ATCGGCCAAATGGATCATTTTTCTAAAGATGATAACTTAAACATATTTAGACTACCTACTGGTTGG
CAATATTTAGTCAACAATAATCTTGGTGGTAATTTAGACGCGACTAATTTCGGTAAGTATGATCAG
TTAGTTCAGGGTTGTTTGAGTACAGGTGCGTATTGTATTGTCGATATCCATAACTATGCCCGCTGG
AACGGGGGTATCATCGGTCAAGGCGGTCCTACCAATGAACAATTCGTTTCAGTTTGGACACAGTTG
GCAACTAAGTATGCATCACAATCACGAGTATGGTTCGGTATCATGAATGAACCTCATGATCTTACC
ATCTCAACATGGGCCGCTACAGTTCAAGCAGTTGTAACTGCTATTAGGAACGCTGGAGCTACTTCT
CAGTTCATTTCCTTACCAGGTACAGATTACACTTCAGCTGCAAATTTTGTTGAAAACGGGTCTGGT
GCCGCTTTGAGCACGGTCACTAACCCGGATGGTACAACGACAAATCTTATCTTCGACGTTCACAAA
TACCTAGATTCAGATAACTCCGGAACACACGCGGAGTGCACAACTAACAATGTTGATGCTTTTTCT
TCGCTCGCTACATGGCTTAGGCAAAATAATAGACAAGCTATTTTAACTGAGACAGGTGGCGGTAAC
GTTCAATCATGTATACAAGACATGTGTCAACAGATTGCGTATCTAAATGCAAATGCAGATGTTTAC
CTCGGATATGTTGGTTGGGGGGCCGGATCTTTTGATTCAACTTATGTCCTAACAGAGACTCCCACT
GGCTCTGGTAGTAGCTGGACTGACACTTCATTAGTTTCGTCCTGTATTGCTCGTTCTAGC
```

10310232:

```
GGTAAAGTTCGTTTCGCTGGTGTGAACATCGCAGGTTTCGACTTTGGATGCACGATTGATGGTACG
TGCGTAACATCTAAGGTTTATCCACCAGTCAAAGACATGCCTCCATACTATAATAATCCTGACGGA
GCAGGTCAGATGGATCATTTTTCTAAAGATGACGGTTTCAATATATTTAGATTGCCTGTTGGATGG
CAATACTTAGTTAATAATAATTTGGGCGGTACACTTGACTCTACCTCCATTTCATATTACGACCAA
TTAGTTCAATCGTGTTTGTCCACAGGTGCTTATTGTATTGTCGACATACATAACTATGCAAGATGG
AATGGAGGCATTATAGGTCAGGGTGGGCCTTCTGATGCGCAGTTTGTCGACCTCTGGACACAATTA
GCCACCAAATACGCATCACAATCAAGAGTATGGTTTGGAATCATGAATGAACCGCATGACTTGCCT
GATATCAATAGATGGGCAACGACTGTTCAAAAGGTCGTGACAGCTATCAGAAATGCTGGTGCCACA
```

```
TCGCAATTCATCAGTTTGCCAGGGACGGACTTTACTAGTGCTGCCAATTTCGTTGAAAACGGTAGT
GGGGCCGCATTAGCCGCGGTCGTAAATCCTGACGGCTCAACTCACAACCTCATTTTTGATGTACAT
AAATACTTGGACTCTGACAATTCTGGTACGCATGCCGAGTGTGTAACCAATAATGCCGATGCATTT
AATTCTTTAGCTCAATGGCTCAGACAAAATAACCGGCAAGCAATCCTAACTGAAACGGGAGGTGGT
AATGTCCAATCTTGCGCTACCTACATGTGTCAACAACTTGATGTGTTAAACCAAAATTCTGATGTG
TATTTGGGGTGGACGTCATGGTCTGCCGGATCATTCCAAGCTTCCTACATACTGACTGAGGTCCCA
ACCGGTTCTGGATCTTCCTGGACAGATCAGTATTTAGTTCAACAATGCTTTGTACCCAAATGGAAG
AAT
```

11010323:

```
GGAAAAGTAAGGTTTGCAGGAGTTAACATAGCCGGCTTCGATTTTGGGTGTGCAACTGATGGGACG
TGCAATACGACAGCAGTATATCCACCGGTAAAAGATATGCCACCTTATTATAACAATCCTGACGGT
GCAGGACAAATGGATCATTTCTCAAAAGACGACACATTCAATATCTTTAGATTGCCAGTAGGTTGG
CAATATCTGGTAAATTCAAATTTAGGGGGTACACTAGATTCTACTAACTTAGGTTACTACGACCAG
CTGGTTCAGTCGTGTTTGTCTACCGGAGCATATTGTATCGTTGATATTCATAATTATGCCAGATGG
AATGGTGGTATTATTGGGCAAGGTGGACCAACTAACGCGCAGTTCACTAGTTTATGGAGCCAATTA
GCATCAAAGTACGCCTCGCAGTCTAGAGTATGGTTCGGTATTATGAATGAACCTCACGATTTACCA
ACTATTTCAACCTGGGCTGCGACCGTGCAAGAAGTGGTTACAGCGATTAGAAATGCTGGTGCTACT
TCTCAATTTATTAGTTTGCCTGGAACAGATTATACATCCGCCGCGAACTTTGTTGAAAATGGCTCA
GGGGCAGCGTTATCTCAAGTTACAAATCCAGACGGAAGCACTACCAATCTTATATTTGACGTCCAT
AAATACCTTGATTCAGATAACTCCGGGACCCATGCCGAATGTGTTACTAACAACGTGGATGCCTTT
AGCTCACTAGCCACTTGGTTAAGACAAAATAATAGACAGGCTATCTTGACGGAAACTGGTGGGGGT
AACGTACAGAGTTGTGGTACCTACATGTGCCAACAATTAGATTTTCTCAACGCAAACTCTGATGTA
TACTTGGGTTGGACAAGCTGGTCCGCAGGCAGTTTTCAGGTTTCATATGTACTAGGCGAAGTGCCA
ACTGGTTCTGGAAGCAGCTGGACCGATACTTACTTAGTGAAGCAATGCTTCGTGCCGAAATGGAAA
TCT
```

22030130:

```
GGCGTGCGTTTTGCAGGTGTTAACATCGCTGGATTTGATTTCGGTGTTGTTACCTCCGGAACACAA
GACATGACACAAATTGTGGATGAGAGTGTCGATGGTGTGACCCAAATTAAACATTTTGTTAATGAT
GATGGAATGACTATCTTCAGACTTCCCAGTGGCTGGCAGTTTTTGGTTAATAATAACCTGGGCGGT
TCGTTAGACAGCAATAATTTCGCCAAATATGATAAGCTAGTGCAAGGCTGTTTGAGCACTGGTGCC
TATTGTATTGTTGACGTCCATAATTACGCTCGATGGAATGGAGGTATTATAGGCCAAGGTGGTCCC
ACGAACGCTCAATTTACCTCATTATGGTCACAATTGGCATCCCATTACGCTAGTCAGAGTCGTGTT
TGGTTCGGTATAATGAATGAGCCTCACGATGTATCTATTACTACTTGGGCTGCTACTGTACAAGAA
GTTGTTACTGCTATTAGAAATGCAGGAGCTACCTCCCAGTTTATTTCTTTACCTGGTAATGACTGG
CAATCCGCCGCCGCCGTTATTAGTGATGGTAGTGCTGCCGCATTGTCCCAAGTTACCAATCCTGAT
GGTTCTACCACAAATCTTATTTTTGATGTCCATAAATATTTGGATAGTGACAATAGTGGTACCCAT
GCAGAATGCGTCACTAACAACATAGATGACGCCTTTGCGCCCTTAGCTACATGGCTGCGACAGAAC
AACAGACAGGCAATCTTGACAGAGACCGGTGGAGGTAACGTGCAGTCTTGCGCCACGTACATGTGT
CAGCAAATTCAATATTTGAATCAAAATTCAGATGTGTATTTAGGTTATGTAGGTTGGAGTGCTGGC
GGATTTCAAGCTTCCTGGAATTATATACTCACCGAAGTGCCTAATGGTAATACAGACCAGTATTTA
GTACAGCAGTGTCTGGCCAGAAAG
```

20320310:

```
GCTAGACAACAAACAGTATGGGGTCAATGTGGTGGTATTGGATGGTCTGGTCCGACAAACTGTGCT
CCAGGCTCGGCATGTTCGACACTAAATCCATATTACGCTCAATGTATCCCTGGCGCTACCACTATA
ACAACTTCTACTAGACCACCTTCTGGTCCGACGACAACTACAAGGGCTACCTCAACCTCTTCCTCT
ACACCCCCTACTTCCAGCGGCGTAAGATTTGCAGGCGTTAACATTGCAGGTTTCGACTTCGGCTGC
ACGACAGACGGAACTTGTGTGACCAGTAAAGTTGTTGATGAGTCTGTAGACGGTGTAAACCAAATG
AGGCATTTTGTCAATGATGATGGCATGACCATATTCAGACTTCCGGTAGGTTGGCAATATTTGGTC
AATAACAATCTCGGCGGTTCGTTGGATTCTACTAGCATATCAAAATACGATAAACTCGTTCAAGGG
TGTCTATCGTTAGGTGCATACTGCATAGTGGATATACACAATTACGCACGTTGGAATGGCGGTATC
ATTGGTCAAGGAGGCCCAAGTGACGCCCAGTTTGTGGACCTGTGGACTCAATTGGCTACGAAGTAT
GCCAGCCAAAGCAGAGTTTGGTTCGGTATTATGAACGAGCCACATGACCTGACTATTAGCACATGG
GCAGCTACCGTACAGAAAGTCGTTACCGCTATAAGAAATGCTGGTGCGACTTCACAATTTATCTCA
TTACCGGGTACTGATTATACATCAGCAGCCAATTTCGTAGAAAATGGCTCAGGTGCTGCATTAGCA
GCCGTAGTCAATCCAGACGGGTCTACACACAACTTGATCTTCGACGTTCATAAATACCTTGACAGT
GATAATTCTGGAACTCATACAGAGTGTGTTACTAATAATGTTGATGCATTTAGCTCTCTTGCGACT
TGGTTAAGGCAGAATAATCGTCAAGCCATATTGACTGAAACAGGGGGTGGAAATACCGCATCCTGT
GAAACATATCTCTGTCAACAGATTCAATACCTTAATCAAAACTCAGACGTTTATTTAGGTTATGTG
GGTTGGGGTGCCGGCGGATTTGACTCTACATGGAACTATGCATTGGACGAAACTCCAACTAACAAT
GTTGATACACCTTTGGTGAAAGCGTGTTTAGCTAGAAAA
```

23111331:

```
GGTAGGACACGTTTTGCTGGCGTAAACATCGCTGGCTTCGATTTTGGCATGGTAACATCAGGTACA
CAAGACTTAACTCAAATAGTTGATGAAAGTGTTGATGGTGTTGGACAGATGGACCACTTTTCCAAG
GATGATAATTTAAATATCTTTAGATTGCCGACAGGATGGCAATACCTTGTTAATAACAATTTGGGT
GGTTCATTGGATGCTACGAACTTTGGTAAGTACGATAAACTTGTCCAAGGTTGTTTGAGCACTGGC
GCTTATTGTATTGTTGATATACATAATTACGCTAGATGGAATGGTGCAATAATTGGTCAAGGTGGA
CCAACTAACGAACAATTCGTGAGCGTTTGGACACAATTAGCCACCAAGTATGCTTCGCAAGCGAGG
GTATGGTTCGGTATTATGAACGAACCGCATGATCTGACTATCTCAACATGGGCCGCAACTGTCCAA
GCCGTGGTCACTGCCATCAGAAATGCAGGGGCGACGTCTCAATTTATATCCTTGCCGGGAACAGAC
TACACATCAGCGGCTAATTTTGTGGAAAACGGTTCAGGTGCGGCTCTGTCCACCGTAACCAATCCC
GATGGAACAACAACCAATTTAATTTTCGATGTACATAAATATCTGGATTCTGACAATAGCGGTACA
CATGCAGAATGTGTGACGAACAATGTCGATGCTTTTAGCAGTTTAGCTACTTGGCTAAGACAAAAT
GGTCGGCAAGCAATATTGACCGAAACTGGTGGAGGCAATGTTCAGAGCTGTGCAACGTACATGTGT
CAGCAGATCGCATACTTAAATGCCAATGCAGATGTCTACCTGGGTTACGTTGGATGGTCGGCTGGC
GGTTTCCAAGCTTCATATATATTAACTGAGGTTCCAACTGGATCGGGCAGTAGCTGGACCGACCAG
TATCTTGTTCAACAATGTATTGCTCGGAGCTCT
```

23121233:

```
GGTAAGGTACGCTTTGCCGGTGTGAACATCGCCGGTTTTGACTTTGGTATGGTCATATCAGGTACT
CAAGATTTAACGCAAATCGTTGATGAATCAGTGGATGGTGTTAATCAGATGCGTCATTTCGTTAAT
GATGACACATTCAATATTTTCAGGCTACCCACAGGATGGCAATACTTGGTTAACAATAATTTAGGA
GGTTCCTTGGATGCCACTAATTTTGGTAAATATGACAAGTTGGTACAAGGCTGTCTAAGCCTAGGA
GCTTATTGTATCGTTGATATTCATAATTACGCTAGATGGAACGGTGCGATTATAGGTCAAGGTGGC
CCAACAAACGAGCAGTTCGTATCTGTATGGACTCAATTAGCGACGAAATATGCTTCCCAAGCAAGG
GTCTGGTTCGGCATCATGAATGAACCACACGACCTAGATATCAATAGATGGGCGACAACAGTTCAG
GCCGTTGTTACAGCAATACGTAACGCTGGAGCAACTTCTCAGTTCATATCTTTGCCAGGGACTGAT
TTCACTAGCGCTGCAAATTTCGTAGAAAATGGCTCTGGTGCAGCCTTGTCCACAGTTACCAATCCG
GATGGTACAACAACTAACCTAATATTTGACGTTCATAAGTATTTGGACAGCGATAATAGTGGCACC
```

```
CACGCCGAGTGTGTTACCAATAACGCCGACGCTTTCAATAGTTTAGCTCAATGGCTACGGCAAAAT
GGTAGACAAGCCATACTGACCGAAACTGGAGGTGGTAACGTCCAATCATGCGCCACCTATATGTGT
CAGCAGTTAGATTTTCTAAACGCCAATTCCGATGTCTACCTTGGATGGACATCGTGGTCAGCGGGT
GGTTTTCAAGCGAGTTGGAACTATATCCTGACCGAAGTTCCCACTAACAATGTTGACCAATATTTG
GTGCAGCAATGCTTTGTCCCTAAATGGAAAAGT
```

32321133:

```
GGAAAGGTGAGATTTGCAGGGGTCAATATAGCTGGTTTTGATTTCGGCGTCGTTACTAGTGGTACT
CAAGATATGACACAGATCGTTGATGAATCTGTTGATGGAGTAAACCAAATGAGACATTTCGTTAAT
GATGACACATTCAATATCTTCAGACTACCATCCGGTTGGCAATTTCTGGTAAATAACAACCTTGGA
GGGCAGTTAGACTCAAACAATTTCGCCCAATATGACAAGTTGGTGCAGGGTTGTCTTAGCTTAGGA
GCTTACTGCATAGTTGATGTCCATAACTACGCAAGGTGGAACGGTGCGATTATTGGCCAGGGCGGT
CCCAGTGATGCTCAATTTGTTGACTTATGGACACAGTTGGCCACCCATTACGCCTCCCAAGCCAGG
GTATGGTTTGGCATAATGAACGAACCTCACGACGTTTCTATAACCACTTGGGCTGCTACTGTGCAA
AAGGTAGTTACCGCTATAAGAAATGCTGGTGCTACCTCTCAGTTCATATCATTGCCAGGTAACGAT
TGGCAATCCGCTGCTGCCGTTATAAGCGACGGCTCGGCTGCCGCCTTGGCTGCGGTGGTTAATCCT
GATGGTAGTACCCATAATCTGATCTTCGACGTACATAAGTACCTGGATTCCGATAATTCCGGTACC
CACGCCGAATGTGTTACGAATAACATAGATGACGCATTCGCTCCTCTAGCTACATGGTTGAGACAA
AATGGTCGTCAAGCCATTCTCACTGAAACCGGTGGTGGAAATGTACAAAGCTGTGCTACTTACATG
TGCCAACAATTGGATTTCTTAAATGCAAACAGTGACGTCTATCTAGGCTGGACATCTTGGAGCGCA
GGGGGTTTCCAAGCAAGCTGGAACTACATTTTAACTGAAGTTCCGACCAACAACGTTGACCAGTAC
TTGGTTCAACAATGTTTTGTACCTAAGTGGAAGTCA
```

33103312:

```
GGTAAGGTACAATTCGCAGGGGTAAATATAGCGGGATTCGATTTTGGAATGGTCACCTCCGGCACC
CAAGATCTAACTCAAATAGTTGACGAATCGGTGGATGGCGTTGGTCAAATGCAACACTTTGTTAAC
GAGGATGGTTTTAACATGTTCCGTCTGCCTACGGGTTGGCAATATTTAGTCAATAACAATCTCGGT
GGTCAATTGGATGCAACCAATTTTGGTCAATATGATAAGTTGGTGCAAGGTTGCCTGTCCCTGGGC
GCACATTGCATTGTTGATATTCATAATTACGCTAGGTGGAATGGTGCAATCATCGGACAGGGTGGC
CCTACTAATGAACAATTTGTTTCCGTCTGGACTCAATTGGCAACTAAGTATAAAGCTGATTCAAAA
GTAGTATTTGGTGTAATGAACGAGCCACACGACTTGACTATCTCCACGTGGGCCGCCACCGTACAA
GCAGTTGTTACTGCAATACGAAACGCAGGAGCTACTTCACAAATGATTTTGCTTCCTGGGACGGAC
TACACTTCTGCTGCAAATTTCGTCGAAAATGGTTCTGGTGCCGCATTGTCAACTGTTACTAACCCA
GATGGCACTACTACCAATCTTATTTTTGATGTACATAAATATCTTGATAGCGATAATTCCGGTACC
CACACCGAATGTGTTACGAATAATGTGGACGCCTTTTCTTCTTTAGCTACATGGCTAAGGTCAGTT
GGTAGGCAAGCCCTACTGTCGGAGACTGGTGGTGGGAATACTGCCTCTTGTGAAACATACCTGTGT
CAACAGCTTGACGTACTAAACCAAAACTCGGATGTGTATTTAGGTTGGACCAGCTGGGGCGCTGGT
GGTTTCGACTCAACATATGCTTTAGATGAAACTCCCACTAGCTCTGGTAACAGCTGGACAGATACG
CCTTTAGTTAAGGCTTGTTTTGTACCTAAATGGAAG
```

33113333:

```
GGAAAGGTACAATTCGCTGGAGTAAATATCGCGGGTTTTGATTTTGGTATGGTCACCTCCGGTACG
CAGGATCTAACTCAAATAGTCGATGAATCCGTAGACGGTGTAGGCCAAATGGATCATTTTTCTAAA
GATGATACCTTCAATATGTTTCGATTGCCTACCGGATGGCAATACTTGGTTAATAATAATTTGGGT
GGTCAATTGGATGCTACCAATTTTGGTCAATATGATAAACTAGTGCAGGGTTGTCTTAGCACAGGT
GCACATTGCATTGTCGATATTCACAATTACGCGAGGTGGAACGGTGCTATTATTGGACAGGGAGGA
```

```
CCAACAAACGAGCAGTTCGTATCTGTGTGGACGCAATTAGCCACGAAATATAAGGCAGATTCGAAA
GTTGTATTCGGAGTAATGAATGAACCACATGATCTTACAATTTCTACCTGGGCAGCAACCGTTCAA
GCAGTGGTCACGGCCATTCGTAACGCTGGCGCAACCTCTCAAATGATTTTATTGCCAGGGACTGAC
TACACTTCCGCCGCCAACTTTGTGGAAAACGGTTCCGGCGCTGCGCTATCCACAGTTACAAATCCA
GATGGTACCACAACGAATCTAATCTTTGATGTTCACAAATACTTGGACTCCGACAACTCCGGCACG
CATGCAGAATGTGTCACCAATAATGTCGACGCATTTTCTTCTTTAGCAACATGGCTTAGATCTGTT
GGCAGACAAGCTTTGTTATCCGAAACAGGTGGTGGTAACGTCCAGTCATGTGCCACTTATATGTGT
CAACAGTTGGACTTCTTGAACGCTAATTCCGATGTGTACTTGGGGTGGACCTCATGGTCCGCTGGT
GGATTTCAGGCAAGTTATATCTTAACAGAAGTGCCTACTGGATCTGGTTCTTCATGGACCGACCAA
TATCTGGTCCAACAATGTTTTGTTCCTAAATGGAAATCT
```

33212131:

```
GGTCGCACAAGGTTTGCAGGAGTGAATATTGCTGGTTTTGATTTCGGTATGGTAACGTCTGGAACG
CAAGATCTCACACAGATTGTCGATGAATCAGTAGATGGTGTAGGACAGATGGACCATTTCAGTAAG
GACGACAATTTGAATATTTTTCGACTACCAACGGGATGGCAGTACTTAGTTAATAATAACCTCGGC
GGTCAGCTAGACGCTACTAATTTTGGTCAGTATGATAAGCTGGTACAAGGGTGCTTATCTACTGGT
GCTTACTGCATTGTAGACATCCACAATTATGCCCGCTGGAATGGTGGTGTCATCGGTCAAGGAGGT
CCTACTGACGATCAGTTCACCTCCTTGTGGACTCAATTAGCAACAAAATATAAAAGCGAGTCAAAA
ATTATTTTCGGAGTAATGAATGAACCACATGACGTGTCTATAACTACTTGGGCTGCCACTGTTCAA
AAAGCTGTTACAGCCATAAGAAAGGCGGGGGCAACTAGTCAAATGATTCTGTTGCCAGGTAACGAT
TGGCAATCCGCTGCTGCTGTCATATCGGATGGAAGTGCTGCAGCTTTGTCTGCAGTCACAAATTTA
GATGGTTCAACCACCAATTTGATCTTTGATGTACATAAATATCTTGATAGTGACAACTCCGGCACA
CACGCTGAATGTGTCACTAACAACATCGATGACGCTTTTGCGCCTTTAGCAACCTGGCTAAGAACC
AACAAAAGACAAGCCATGTTGACAGAGACGGGTGGTGGAAATGTTCAATCCTGTGCTACTTACATG
TGTCAGCAAATTGCCTACCTAAACGCTAATGCTGATGTTTATTTAGGTTATGTTGGATGGTCTGCT
GGAGGCTTTCAAGCGAGCTATATCCTGACTGAAGTCCCGACAGGCTCCGGGAGCTCCTGGACTGAC
CAATATTTGGTACAACAGTGTATTGCCAGATCAAGT
```

33231313:

```
GGAAAAGTCAGGTTTGCTGGAGTAAACATTGCTGGCTTTGATTTTGGAATGGTTACTTCAGGTACC
CAAGATCTGACCCAAATTGTAGATGAGAGTGTAGATGGTGTGACTCAGATTAAACATTTCGTCAAT
GACGATACCTTCAACATCTTTAGGTTGCCAACAGGTTGGCAATATCTAGTGAATAACAATCTTGGT
GGTCAACTGGATGCCACCAACTTCGGTCAATACGATAAGCTAGTACAAGGTTGTTTGTCTACTGGT
GCTTACTGTATTGTTGATATTCATAACTACGCTAGGTGGAACGGTGCCATTATTGGTCAGGGAGGT
CCTACAGACGATCAGTTTACTTCCTTGTGGACCCAGTTAGCTACTAAATATGCAAGTCAAGCTAGA
GTCTGGTTTGGCATTATGAATGAACCACATGATCTAACTATTTCAACATGGGCTGCCACAGTCCAA
AAAGCTGTTACCGCGATTAGAAAAGCTGGAGCTACTTCTCAATTTATTTCATTGCCTGGAACAGAC
TACACCTCTGCCGCTAACTTTGTTGAAAATGGTTCGGGCGCAGCTCTTAGCGCTGTAACTAACCTA
GACGGTAGTACAACTAACCTGATCTTCGATGTTCATAAATATCTGGACAGTGATAACTCTGGTACG
CACACTGAATGCGTTACTAACAATGTTGACGCCTTCAGTTCTCTTGCTACATGGTTAAGACAAAAT
GGCCGACAAGCAATTTTAACTGAAACAGGAGGGGGTAACACCGCAAGCTGCGAAACATATTTATGT
CAGCAGTTAGACTTCTTGAATGCTAACTCTGACGTCTACTTAGGATGGACTTCTTGGGGTGCAGGT
GGCTTCGACTCGACTTGGAATTATGCGTTAGACGAAACCCCCAATGGCAATACAGATACACCATTG
GTAAAAGCCTGCTTCGTCCCAAAGTGGAAATCA
```

11311330:

```
GGGGTTAGATTTGCTGGAGTTAACATAGCAGGATTCGACTTCGGTTGTGCGACGGACGGCACTTGC
AATACTACGGCAGTATATCCTCCAGTGAAAGACATGCCTCCCTATTACAATAATCCAGATGGAGCC
GGCCAAATGGACCATTTTTCTAAAGATGACGGTATGACAATTTTTCGCTTACCCGTTGGCTGGCAG
TACTTGGTAAATTCCAATTTGGGTGGTACATTAGACTCTACTAATCTAGGCTATTATGACCAACTG
GTGCAGAGCTGCTTGTCAACTGGCGCTTATTGCATCGTGGACATACATAACTATGCAAGATGGAAC
GGTGCTATAATTGGCCAGGGCGGACCTTCGGACGCACAGTTTGTTGACCTGTGGACACAATTAGCT
ACTAAATACGCATCCCAGGCAAGGGTTTGGTTTGGCATTATGAATGAGCCTCACGACTTGCCGACC
ATAAGCACATGGGCCGCCACGGTTCAGAAAGTAGTCACTGCTATTCGTAACGCGGGAGCAACTTCT
CAGTTTATTTCACTCCCTGGTACAGACTATACCTCTGCTGCAAATTTCGTAGAGAATGGTAGTGGT
GCTGCTCTGGCAGCTGTAGTAAATCCTGACGGATCGACTCATAACCTGATTTTTGATGTCCATAAG
TATTTGGATTCAGACAACTCTGGCACACACGCTGAATGTGTGACTAACAATGTTGATGCTTTCTCT
AGCCTTGCAACATGGCTGCGTCAAAATGGAAGACAAGCCATCTTGACGGAGACCGGCGGAGGTAAT
GTACAATCATGTGCAACTTACATGTGCCAGCAAATTCAATATTTGAACCAAAACTCCGATGTATAC
CTCGGTTATGTAGGTTGGAGCGCGGGCTCCTTCCAAGCTTCCTATATTTTGACTGAAGTCCCAACA
GGATCGGGTTCATCTTGGACGGATCAATACCTAGTTCAGCAATGCTTAGCCAGAAAA
```

20333123:

```
GGTAAAGTGCAATTTGCGGGGGTAAATATTGCGGGTTTTGATTTTGGTTGCACTACCGACGGTACT
TGTGTGACTAGTAAAGTAGTTGATGAATCAGTCGATGGCGTAACACAAATCAAACATTTTGTGAAT
GATGATACGTTCAATATGTTTAGATTACCTGTTGGTTGGCAGTACTTAGTCAACAATAATTTAGGA
GGGAGTTTAGACTCAACTTCTATTTCAAAATATGATAAATTAGTACAGGGTTGCTTATCAACAGGT
GCTCATTGTATTGTTGATATTCATAACTATGCTAGATGGAATGGTGCTATCATTGGCCAAGGTGGC
CCTAGTGATGCACAATTCGTGGATTTATGGACTCAATTGGCAACCAAATATAAAGCTGATTCTAAG
GTAGTATTCGGCGTCATGAACGAACCACATGATGTATCTATAACGACGTGGGCAGCAACAGTACAA
AAGGTCGTGACCGCTATTAGGAATGCTGGAGCGACATCTCAAATGATATTGTTACCTGGTAATGAC
TGGCAATCAGCTGCAGCCGTTATTTCAGATGGTTCGGCGGCTGCATTAGCAGCCGTCGTGAACCCT
GATGGGTCAACGCATAACCTAATTTTTGATGTACACAAATACCTAGATTCTGATAACTCAGGAACA
CATGCTGAGTGTGTCACTAATAACATTGATGATGCCTTCGCTCCCCTAGCTACCTGGCTTAGGAGT
GTGGGTCGGCAGGCCTTGCTTTCTGAAACGGGCGGAGGTAACGTCCAATCTTGTGGAACTTACATG
TGTCAACAACTGGATTTTCTAAACGCTAATTCTGATGTGTATCTCGGTTGGACATCTTGGTCAGCG
GGTGGGTTCCAAGTTTCTTGGAATTACGTCCTTGGAGAAGTGCCGAACGGTAACACAGACACGTAT
TTAGTAAAACAATGTTTTGTACCTAAATGGAAGAGC
```

10203103:

```
GGAAAGGTACAGTTCGCTGGAGTGAATATTGCAGGTTTTGACTTTGGCTGTACAACAGACGGCACT
TGTGTTACTTCCAAAGTATATCCCCCTGTCAAAGATATGCCGCCATACTACAATAATCCTGATGGA
GCAGGACAGATGCAACATTTTGTCAATGAAGATACCTTTAACATGTTCAGGCTTCCAGTCGGTTGG
CAATACTTAGTAAATAATAATTTGGGTGGAACTTTGGATTCCACGAGCATTTCTTATTACGACCAG
TTAGTTCAATCTTGCTTGTCATTGGGTGCTCATTGCATTGTTGACATCCATAACTATGCACGTTGG
AATGGTGCTATTATCGGGCAAGGTGGCCCTACCGATGATCAATTCACATCATTATGGACACAACTA
GCTACAAAGTATAAAGCCGACTCCAAAGTAGTCTTTGGTGTCATGAATGAACCTCATGACGTCCCC
AGCATAACAACATGGGCGGCTACGGTTCAGAAGGCTGTAACCGCTATCAGAAAAGCTGGTGCTACC
TCTCAAATGATTTTACTGCCTGGTAATGATTGGCAATCCGCAGCGGCTGTTATATCTGATGGAAGT
GCTGCGGCTTTATCTGCTGTAACCAACCTTGACGGCTCAACTACTAATCTGATCTTTGATGTTCAT
AAATACTTAGACTCTGACAACTCAGGTACGCATGCAGAATGTACTACCAATAACATTGATGATGCA
TTTGCACCACTGGCTACATGGTTGAGATCCGTAGGTCGTCAGGCCTTATTGTCTGAGACTGGCGGT
GGCAATGTCCAATCATGCATACAAGATATGTGCCAGCAACTAGATTTTCTTAACGCTAATTCAGAT
```

```
GTGTACCTTGGATGGACATCCTGGGGCGCAGGTAGTTTTGATTCGACATATGTTCTAACTGAAACC
CCCACGTCTTCCGGTAATTCATGGACTGATACGTCGCTAGTAAGCAGTTGTTTCGTACCTAAATGG
AAGTCT
```

00130002:

```
GGTAAAGTTAGGTTCGCTGGCGTAAACATTGCTGGTTTCGACTTTGGGTGTACAACCGACGGTACG
TGTGTTACTTCTAAAGTATATCCACCATTGAAGAATTTCACCGGTTCCAACAACTATCCTGATGGC
ATTACGCAGATTAAACACTTCGTCAACGATGATGGTTTCAATATCTTCAGATTACCAGTGGGTTGG
CAATATTTGGTCAACAATAATCTGGGTGGTAACCTAGATTCCACTTCAATCTCAAAGTATGATCAA
CTGGTCCAAGGCTGTTTATCTACCGGGGCCTACTGTATAGTTGACATTCATAACTACGCTAGATGG
AATGGAGGAATTATCGGTCAAGGGGGTCCGACTAATGAACAGTTCGTTAGCGTCTGGACCCAATTA
GCTACAAAGTACGCTTCACAGTCTAGGGTTTGGTTTGGGATTATGAACGAACCTCACGACGTTAAC
ATCAATACTTGGGCTGCAACAGTGCAAGCTGTAGTTACTGCGATTAGAAACGCAGGTGCCACCTCA
CAGTTTATTAGTCTTCCTGGCAACGATTGGCAATCAGCTGGTGCATTTATCTCTGATGGTTCAGCA
GCTGCCTTATCAACTGTCACAAACCCCGATGGTACAACCACAAATCTTATATTCGATGTCCATAAA
TATTTGGATTCTGATAATAGCGGGACACATGCTGAATGTACTACAAATAACATCGACGGGGCATTT
AGTCCTCTGGCAACATGGCTGAGACAAAATAATCGTCAGGCTATTTTAACTGAGACCGGTGGAGGG
AATGTACAATCTTGTATCCAAGACATGTGTCAACAATTAGACGTTCTGAACCAAAACTCAGACGTA
TATTTGGGCTGGACTAGCTGGGGCGCAGGTTCATTCGATAGTACCTGGAATTATGTTCTGACAGAA
ACGCCAAATGGTAACACAGACACTTCTCTAGTTTCGTCGTGCTTCGTTCCGAAATGGAAGAAT
```

13101033:

```
GGGAAAGTTAGATTTGCGGGTGTCAATATTGCTGGTTTTGATTTCGGTATGGTTACCAGCGGTACT
CAAGACTTGACACAAATCTATCCACCTGTAAAAGACATGCCGCCTTATTATAACAATCCGGACGGT
GCAGGTCAGATGCAACACTTTGTAAACGAAGATACCTTTAACATTTTTAGGCTTCCAACCGGATGG
CAATACTTAGTGAATAATAATTTAGGTGGTACCCTGGATGCCACGAACTTCGGTTATTACGATCAG
TTAGTACAATCTTGTTTAAGTTTGGGCGCTTATTGTATTGTCGACATACATAACTATGCTAGATGG
AATGGTGCAATCATAGGCCAAGGTGGTCCAACAAATGAACAATTTGTCTCAGTATGGACGCAGTTA
GCTACCAAATACGCTAGTCAGGCCCGTGTTTGGTTCGGTATAATGAACGAACCGCATGATGTCCCC
AACATTAACACATGGGCTGCAACAGTCCAGGCAGTTGTCACTGCTATCAGGAACGCCGGTGCTACA
TCTCAGTTTATTTCCCTACCGGGTAACGACTGGCAATCAGCTGGTGCTTTCATCTCAGACGGGAGC
GCCGCAGCATTGTCCACTGTGACCAACCCAGATGGTACAACTACTAACTTAATATTTGATGTGCAC
AAGTATCTAGATTCCGATAATTCTGGCACACATGCAGAATGCGTGACTAACAATATTGATGGCGCT
TTTTCTCCGTTAGCCACTTGGCTTAGGCAAAACGGGAGGCAGGCTATTCTCACCGAAACGGGTGGT
GGTAACGTACAGTCCTGTGCCACTTATATGTGTCAACAACTTGATTTCTTAAATGCCAACTCGGAT
GTTTACCTAGGTTGGACATCGTGGTCTGCAGGCTCGTTTCAAGCATCTTACATTTTGACAGAAGTC
CCTACATCATCGGGCAATTCTTGGACAGATCAATATTTGGTTCAACAATGTTTTGTACCGAAATGG
AAGTCC
```

31311011:

```
GGACGCACGAGATTCGCTGGAGTAAATATTGCTGGTTTTGATTTTGGCTGCGCCACAGATGGTACA
TGTAACACTACGGCAGTCGTAGACGAGTCAGTTGATGGGGTTGGTCAAATGGATCATTTCTCGAAA
GACGATAACCTAAACATATTCAGATTACCAGTGGGATGGCAGTATTTAGTGAACTCAAATCTTGGA
GGACAACTGGATTCCACTAACTTGGGCCAATACGACAAATTGGTTCAAGGTTGCTTATCCACTGGT
GCTTACTGCATAGTCGATATACATAACTACGCTCGATGGAACGGCGCCATTATCGGCCAAGGTGGT
CCCTCAGATGCTCAATTTGTTGACTTGTGGACACAACTCGCTACTAAATACGCATCACAAGCTAGG
```

GTTTGGTTCGGTATTATGAATGAACCTCATGACGTTAATATAAATACCTGGGCTGCTACAGTACAG
AAAGTGGTCACTGCTATTCGGAATGCCGGTGCAACCTCTCAGTTTATATCCTTGCCAGGAAACGAC
TGGCAGTCAGCTGGAGCATTTATCTCAGACGGTTCGGCCGCTGCTTTAGCTGCTGTCGTGAATCCC
GATGGGAGTACACACAACCTAATATTCGATGTGCATAAGTACTTGGATAGCGACAATAGTGGAACT
CATACGGAATGTGTCACTAATAATATTGATGGTGCCTTTAGTCCATTGGCAACCTGGCTTAGACAG
AATGGGAGACAAGCAATATTGACAGAAACCGGAGGGGGAAATACGGCTTCCTGCGAGACTTATTTG
TGCCAGCAAATCGCTTATCTTAACGCCAACGCTGATGTTTATTTAGGATATGTGGGTTGGGGCGCA
GGAGGTTTCGATTCAACGTACGCATTAGATGAAACTCCTACTGGTAGTGGTTCATCATGGACTGAT
ACACCCTTAGTTAAAGCTTGTATAGCACGGTCAAGT

Gene sequences- Optimized chimera test set

01003013:

GGTAAAGTTCAATTCGCTGGGGTTAACATTGCTGGTTTTGATTTCGGTTGCGCTACAGATGGAACA
TGCAATACCACCGCGGTCTACCCTCCCTTGAAGAACTTTACAGGATCGAATAACTACCCCGATGGT
ATTGGCCAGATGCAGCACTTTGTTAATGAAGATACATTTAATATGTTTAGGCTTCCCGTGGGCTGG
CAATATTTAGTCAACTCGAATCTGGGAGGTAACTTGGACTCTACAAATCTAGGTAAATATGATCAG
TTAGTTCAAGGATGTTTATCTTTGGGTGCTCATTGCATCGTCGACATTCATAACTACGCTAGATGG
AATGGTGCTATTATTGGCCAAGGTGGTCCTACTAACGCCCAGTTCACATCGCTATGGAGTCAATTG
GCGTCTAAGTACAAAGCTGATTCAAAGGTTGTGTTTGGTGTTATGAACGAACCACACGATGTTAAT
ATAAACACATGGGCTGCAACAGTCCAGGAAGTCGTTACTGCTATCCGTAACGCTGGTGCGACGAGT
CAAATGATCTTGTTACCAGGTAACGATTGGCAGTCAGCAGGCGCTTTCATCAGTGATGGTTCGGCT
GCTGCACTGAGCCAAGTAACAAATCCGGATGGTTCTACTACCAATTTAATCTTTGACGTACATAAA
TATTTGGACTCGGACAACTCCGGTACCCATACCGAATGTGTGACAAATAACATTGACGGGGCATTC
TCGCCCCTAGCAACTTGGCTAAGAAGTGTAGGGAGACAAGCTCTACTTTCGGAGACTGGTGGGGGC
AACACTGCCTCTTGTGAAACATACTTATGTCAACAGTTGGACTTTCTGAATGCAAACTCAGATGTT
TATTTGGGCTGGACCAGTTGGGGTGCCGGCTCTTTCGATAGTACCTACGCATTGGATGAAACTCCA
ACATCTTCTGGTAATAGCTGGACCGACACTCCACTGGTAAAAGCGTGTTTTGTCCCAAAATGGAAG
AGT

01003213:

GGTAAAGTTCAATTCGCTGGGGTTAACATCGCAGGATTTGATTTTGGATGTGCTATCGATGGAACT
TGCAACACTACAGCTGTGTATCCTCCACTTAAAAATTTTACAGGCTCCAATAATTACCCTGACGGG
ATTGGTCAGATGCAGCATTTTGTTAATGAAGACACTTTCAATATGTTTAGGCTACCTGTTGGCTGG
CAATACTTGGTCAATAGTAATTTAGGAGGCAACTTAGATTCTACGAATTTAGGCAAGTATGATCAA
CTTGTTCAGGGGTGCCTCTCATTGGGTGCTCACTGCATAGTAGATATTCATAATTACGCGCGATGG
AATGGCGCAATTATTGGACAAGGTGGTCCTACTAACGCCCAATTTACATCTCTGTGGTCACAATTG
GCGAGTAAGTACAAGGCCGACAGTAAAGTTGTTTTTGGTGTCATGAATGAACCACATGATTTGGAC
ATAAATAGATGGGCCACCACCGTGCAGGAAGTCGTTACCGCGATACGGAATGCTGGGGCAACCAGC
CAAATGATTCTGTTGCCTGGAACAGACTTCACATCAGCAGCAAATTTTGTCGAAAACGGAAGTGGT
GCGGCATTGTCCCAAGTGACAAACCCCGACGGTTCTACCACCAATCTAATTTTCGACGTACATAAG
TATTTAGATTCTGATAACTCCGGCACCCACACTGAGTGTGTTACCAATAATGCAGACGCATTTAAT
TCACTAGCCCAATGGTTAAGATCTGTCGGTCGACAAGCCCTTCTTTCTGAAACAGGGGGTGGCAAC
ACAGCCAGCTGTGAAACCTACCTATGCCAGCAACTTGATTTCTTAAACGCCAACTCTGACGTTTAC
TTGGGTTGGACTTCCTGGGGAGCCGGCTCGTTTGACTCAACTTACGCGTTGGACGAGACACCCACC
TCATCGGGAAATAGTTGGACAGACACCCCATTAGTCAAAGCATGTTTTGTTCCCAAGTGG

01013113:

```
GGTAAAGTTCAATTCGCTGGGGTTAACATTGCAGGTTTCGACTTTGGTTGTGCTACAGACGGTACC
TGTAATACTACCGCAGTTTACCCACCCTTGAAGAACTTCACTGGCTCGAACAATTACCCGGATGGA
ATTGGACAAATGGATCACTTTAGTAAAGACGACACATTTAACATGTTCAGGCTCCCAGTAGGATGG
CAATATCTGGTGAATTCCAACCTCGGCGGAAACCTAGATTCAACAAATTTAGGTAAGTACGATCAA
CTGGTTCAAGGCTGTCTTTCTACTGGTGCGCACTGTATTGTAGATATCCATAACTACGCACGGTGG
AATGGTGCAATAATTGGTCAGGGAGGCCCCACAAACGCGCAATTTACAAGTTTATGGTCTCAATTA
GCGAGCAAGTACAAGGCGGATTCTAAGGTCGTCTTTGGGGTGATGAACGAACCTCATGACGTGTCG
ATAACCACGTGGGCGGCAACTGTGCAGGAAGTAGTTACAGCCATACGAAATGCTGGCGCCACTTCA
CAAATGATATTGTTACCCGGTAACGATTGGCAATCTGCCGCAGCTGTAATCAGCGATGGTTCTGCT
GCAGCCCTCTCTCAAGTTACCAATCCAGATGGTAGCACTACAAATCTAATCTTTGATGTACACAAG
TATTTAGATTCTGATAACTCCGGTACACACACCGAATGCGTAACGAATAATATTGATGATGCTTTC
GCTCCATTGGCCACATGGCTACGCTCAGTCGGTCGTCAAGCCTTGCTGTCCGAAACTGGAGGAGGA
AACACAGCTTCATGTGAAACTTACCTGTGTCAGCAATTGGATTTCCTTAACGCCAATAGCGACGTC
TATCTCGGCTGGACATCATGGGGTGCAGGCTCCTTCGATTCAACATACGCTCTTGATGAAACGCCT
ACCGGTTCAGGATCAAGTTGGACTGACACACCCTTGGTCAAGGCTTGCTTCGTGCCTAAGTGGAAG
AGT
```

31013113:

```
GGTAAAGTTCAATTCGCTGGGGTTAACATCGCGGGGTTCGATTTCGGTTGCGCTACTGACGGAACC
TGTAATACAACTGCTGTAGTCGATGAATCGGTGGACGGTGTAGGTCAAATGGATCATTTCAGTAAA
GATGACACATTTAACATGTTCAGATTGCCTGTAGGTTGGCAGTATTTGGTTAACTCGAATTTGGGT
GGTCAGTTAGACAGCACCAACTTAGGACAATATGACAAATTAGTGCAAGGTTGCCTTTCCACAGGT
GCCCACTGTATCGTCGATATACACAATTACGCGAGATGGAACGGAGCTATTATAGGCCAGGGTGGT
CCCACCAATGCTCAATTCACATCCTTATGGTCCCAATTGGCTAGTAAGTATAAAGCAGACTCGAAA
GTCGTTTTTGGAGTGATGAACGAGCCTCACGACGTCAGTATTACTACATGGGCCGCTACAGTACAA
GAGGTCGTCACCGCTATCAGGAATGCTGGAGCAACCTCTCAAATGATTTTGCTGCCAGGTAACGAT
TGGCAAAGCGCAGCAGCAGTAATAAGCGATGGATCAGCAGCCGCACTGTCTCAAGTCACGAACCCC
GATGGATCTACGACTAATTTGATATTCGACGTTCATAAATATTTAGATAGTGATAACTCTGGGACG
CACACAGAATGCGTTACAAATAATATCGATGACGCATTTGCCCCGTTAGCTACGTGGTTAAGGTCT
GTAGGACGGCAAGCCCTACTCTCCGAAACAGGAGGTGGAAATACTGCTTCATGCGAAACCTACTTA
TGTCAACAGCTCGATTTCTTAAACGCAAACTCAGATGTTTATCTGGGGTGGACAAGCTGGGGAGCT
GGTGGATTTGATTCAACTTATGCGCTTGACGAAACTCCTACTGGTTCTGGCTCTAGCTGGACTGAT
ACACCTCTAGTTAAAGCTTGCTTCGTTCCGAAATGGAAGAGT
```

01003113:

```
GGTAAAGTTCAATTCGCTGGGGTTAACATTGCGGGCTTTGATTTCGGTTGTGCGACCGATGGGACC
TGCAATACAACAGCAGTTTATCCTCCACTAAAGAACTTCACGGGATCTAATAACTATCCGGACGGC
ATAGGACAAATGCAGCATTTTGTGAATGAAGATACCTTTAATATGTTTCGGCTTCCAGTTGGTTGG
CAATACCTGGTTAACTCTAATCTAGGTGGTAACCTCGATTCGACAAATTTGGGTAAATATGATCAA
TTGGTTCAAGGTTGTCTATCCTTAGGAGCTCATTGCATCGTGGATATACACAACTACGCAAGATGG
AATGGAGCTATTATCGGTCAGGGTGGACCAACTAACGCACAATTCACTTCACTGTGGAGCCAACTA
GCATCTAAGTACAAGGCAGACAGTAAAGTTGTCTTTGGAGTAATGAACGAACCGCACGATGTCTCC
ATAACGACCTGGGCGGCTACCGTACAAGAAGTGGTCACTGCTATTAGAAACGCTGGTGCGACATCA
CAGATGATCTTATTACCGGGCAATGATTGGCAGTCCGCTGCGGCAGTGATATCAGATGGGTCTGCA
GCTGCATTATCCCAAGTAACTAATCCAGACGGAAGCACCACAAATTTGATTTTCGACGTTCACAAG
```

```
TATCTAGATTCTGATAACTCAGGAACCCACACCGAATGTGTTACGAATAACATTGACGACGCCTTT
GCACCACTAGCTACCTGGTTGAGGAGTGTTGGGAGACAAGCATTACTAAGTGAAACTGGGGGTGGT
AATACAGCAAGTTGTGAGACATATCTCTGTCAGCAGTTGGACTTCTTGAATGCTAACAGCGATGTA
TATTTGGGATGGACATCTTGGGGCGCAGGTTCCTTTGACTCGACTTATGCTCTAGACGAAACACCA
ACTTCGAGTGGCAATTCCTGGACCGATACTCCCTTAGTCAAGGCATGCTTTGTTCCTAAATGGAAG
AGT
```

00000003:

```
GGCAAGGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACT
TGTGTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGA
ATAGGGCAGATGCAACATTTTGTTAATGAAGACACGTTTAACATATTTCGTTTGCCTGTTGGATGG
CAATATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAA
TTGGTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGG
AATGGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTG
GCTAGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAAC
ATTAATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCC
CAGTTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCC
GCTGCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAG
TATTTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGGTGCCTTC
AGCCCGTTGGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGT
AATGTACAAAGTTGTATTCAGGATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTT
TACTTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCT
ACATCCAGTGGTAATAGTTGGACCGATACTTCTTTGGTATCTTCTTGCTTAGCTAGAAAG
```

01013213:

```
GGTAAAGTTCAATTCGCTGGGGTTAACATCGCTGGTTTTGACTTCGGGTGCGCTATTGATGGCACG
TGTAATACTACAGCGGTATACCCACCCTTGAAGAATTTTACTGGTTCGAACAATTACCCGGATGGA
ATAGGTCAGATGGACCATTTTTCGAAAGATGACACCTTCAACATGTTTAGACTTCCCGTCGGTTGG
CAGTACCTAGTGAACTCAAACCTTGGTGGAAATTTGGATTCTACTAACTTAGGGAAATACGATCAA
TTAGTCCAAGGTTGTTTGTCCACTGGAGCACATTGTATAGTCGATATTCACAACTATGCTCGTTGG
AACGGCGCAATTATAGGTCAAGGTGGTCCTACAAACGCACAGTTCACATCTTTGTGGTCACAACTC
GCGTCCAAATACAAGGCGGACTCGAAGGTTGTTTTCGGTGTGATGAATGAGCCACACGACCTCGAC
ATTAACAGATGGGCTACAACAGTTCAGGAAGTGGTAACTGCAATTAGAAATGCCGGAGCTACATCA
CAGATGATTCTTTTGCCAGGTACTGACTTCACAAGTGCTGCCAACTTTGTGGAAAATGGCAGCGGT
GCGGCCTTGTCACAAGTCACAAATCCGGATGGTTCTACAACCAACCTAATATTTGACGTCCATAAG
TATCTTGACAGTGATAACAGTGGGACTCACACCGAGTGTGTCACGAATAATGCTGATGCGTTCAAC
TCTTTAGCGCAATGGCTCAGGAGTGTAGGTAGACAGGCTTTGCTGTCTGAAACGGGAGGGGGTAAC
ACTGCGTCTTGCGAGACCTACCTGTGCCAACAACTCGATTTTTTGAACGCCAATTCAGATGTCTAC
CTTGGCTGGACCTCTTGGGGTGCCGGGTCCTTTGATTCCACTTACGCTTTAGACGAAACCCCAACT
GGATCGGGGTCTTCTTGGACTGATACTCCTTTAGTAAAAGCCTGTTTTGTGCCAAAGTGGAAGAGT
```

01000000

00000010:

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACTTGT
GTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGAAGACGGTATGACAATATTTCGTTTGCCTGTTGGATGGCAA
```

```
TATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
AATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATACTGAATGTGTTACCAATAATATCGATGGTGCTTTTAGT
CCATTGGCAACCTGGCTGAGGCAGAACAATAGACAAGCTATTCTTACTGAGACTGGAGGAGGTAAT
ACCGCATCTTGCGAGACATATCTGTGCCAACAAATACAATACTTGAATCAGAACAGCGATGTTTAT
TTAGGTTACGTTGGCTGGGGTGCGGGATCATTTGATAGCACATACGCGCTTGATGAAACACCAACA
TCTTCCGGTAATTCATGGACTGACACTCCACTCGTAAAAGCTTGTCTTGCTAGGAAA
```

00003000:

```
GGTAAAGTTCAATTCGCTGGGGTTAACATTGCAGGTTTCGATTTTGGTTGTACTACCGATGGAACC
TGTGTTACCAGTAAAGTGTATCCCCCACTTAAAAATTTCACAGGCTCGAATAATTATCCTGATGGT
ATAGGTCAAATGCAGCATTTTGTGAATGAAGATGGCATGACTATGTTCCGTCTTCCTGTGGGCTGG
CAATACTTAGTTAATAACAATCTTGGTGGCAATCTAGACTCCACTTCTATATCAAAGTATGACCAA
CTAGTACAAGGCTGCCTTAGCCTTGGCGCACATTGTATAGTTGATATCCACAATTATGCAAGATGG
AACGGTGCCATTATCGGACAAGGCGGACCTACTAATGCCCAGTTTACATCCTTGTGGAGTCAACTG
GCAAGCAAATACAAAGCCGATTCAAAAGTTGTATTTGGTGTCATGAACGAGCCGCACGATGTCAAC
ATTAATACTTGGGCAGCTACCGTCCAGGAGGTCGTCACTGCCATCAGGAATGCAGGCGCTACTAGT
CAGATGATATTGCTTCCTGGAAACGACTGGCAATCCGCTGGTGCGTTTATTTCTGATGGATCAGCT
GCCGCTTTGTCACAAGTTACTAACCCCGATGGTAGTACCACTAATCTCATTTTTGATGTTCATAAG
TACCTTGATTCTGATAATTCGGGGACACACGCTGAGTGTACCACCAATAACATAGACGGAGCATTC
TCACCTCTAGCAACCTGGTTGAGGTCCGTGGGCAGACAAGCCTTGCTTTCGGAAACTGGTGGAGGT
AATGTTCAAAGCTGCATCCAAGATATGTGCCAACAAATTCAATACTTAAATCAAAACTCTGACGTG
TATTTAGGTTATGTTGGTTGGGGCGCTGGTTCTTTCGATTCAACATATGTCTTGACCGAAACCCCA
ACCTCGTCTGGCAATTCATGGACAGACACTTCACTAGTTTCAAGCTGTCTAGCCAGAAAA
```

00002000:

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACTTGT
GTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGAAGACGGTATGACAATATTTCGTTTGCCTGTTGGATGGCAA
TATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGCATTGTTGATATTCATAACTACGCCAGATGGAAT
GGTGGTGTTATTGGCCAAGGTGGTCCAACCAATGCTCAATTTACCTCATTATGGTCGCAATTGGCA
TCCAAGTATAAATCTGAGTCGAAAATTATTTTTGGCGTGATGAACGAACCCCATGATGTAAACATT
AACACTTGGGCTGCAACCGTTCAAGAAGTCGTTACAGCTATAAGAAACGCAGGTGCCACATCTCAA
ATGATCCTGCTCCCAGGGAACGATTGGCAATCGGCCGGTGCTTTCATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGGTGCCTTCAGC
CCGTTGGCAACCTGGTTACGTACAAACAAGAGACAAGCAATGTTGACGGAAACCGGTGGTGGTAAT
GTACAAAGTTGTATTCAGGATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TCCAGTGGTAATAGTTGGACCGATACTTCTTTGGTATCTTCTTGCTTAGCTAGAAAG
```

03000000:

```
GGTGTTAGATTCGCCGGAGTCAATATCGCTGGATTTGATTTTGGTATGGTAACCAGTGGTACCCAA
GATCTGACTCAGATTTACCCTCCCTTAAAGAATTTCACTGGCTCAAATAATTACCCAGACGGTATC
GGACAAATGCAGCATTTTGTAAATGAGGACGGCATGACTATCTTTCGGTTACCAACAGGTTGGCAA
TATTTAGTTAATAATAATTTGGGTGGTAATTTAGACGCTACGAATTTCGGTAAGTATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
AATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGGTGCCTTCAGC
CCGTTGGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGTAAT
GTACAAAGTTGTATTCAGGATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TCCAGTGGTAATAGTTGGACCGATACTTCTTTGGTATCTTCTTGCTTAGCTAGAAAG
```

10000000:

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTTGGCTGTACAACAGACGGCACTTGT
GTTACTTCCAAAGTATATCCCCCTGTCAAAGATATGCCGCCATACTACAATAATCCTGATGGAGCA
GGACAGATGCAACATTTTGTCAATGAAGATGGAATGACTATCTTCAGGCTTCCAGTCGGTTGGCAA
TACTTAGTAAATAATAATTTGGGTGGAACTTTGGATTCCACGAGCATTTCTTATTACGACCAGTTA
GTTCAATCTTGCTTGTCATTGGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
AATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGGTGCCTTCAGC
CCGTTGGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGTAAT
GTACAAAGTTGTATTCAGGATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TCCAGTGGTAATAGTTGGACCGATACTTCTTTGGTATCTTCTTGCTTAGCTAGAAAG
```

00000013:

```
GGCAAGGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACT
TGTGTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGA
ATAGGGCAGATGCAACATTTTGTTAATGAAGACACGTTTAACATATTTCGTTTGCCTGTTGGATGG
CAATATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAA
TTGGTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGG
AATGGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTG
GCTAGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAAC
ATTAATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCC
CAGTTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCC
GCTGCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAG
TATTTGGATTCGGATAATAGTGGTACTCATACTGAATGTGTTACCAATAATATCGATGGTGCTTTT
```

AGTCCATTGGCAACCTGGCTGAGGCAGAACAATAGACAAGCTATTCTTACTGAGACTGGAGGAGGT
AATACCGCATCTTGCGAGACATATCTGTGCCAACAACTAGATTTTCTTAACGCTAATTCAGATGTG
TACCTTGGATGGACATCCTGGGGCGCAGGTAGTTTTGATAGCACATACGCGCTTGATGAAACACCA
ACATCTTCCGGTAATTCATGGACTGACACTCCACTCGTAAAAGCTTGTTTCGTACCTAAATGGAAG
TCT

12002010:

GGCGTCAGATTTGCGGGTGTCAACATAGCCGGGTTTGATTTTGGTGTGGTTACATCCGGCACGCAA
GATATGACCCAGATCTACCCACCTGTTAAAGATATGCCACCATACTATAATAATCCCGATGGTGCT
GGTCAGATGCAACATTTTGTGAATGAAGACGGAATGACTATATTCCGTTTACCTTCCGGCTGGCAG
TTTCTAGTCAACAATAATTTGGGTGGCACATTAGATAGTAACAACTTTGCTTATTACGATCAACTG
GTTCAATCTTGTCTCAGCCTAGGCGCATATTGTATAGTTGATGTACATAACTACGCCCGCTGGAAT
GGCGGGGTCATTGGACAAGGTGGTCCAACCAATGCTCAGTTTACATCTCTGTGGTCCCAGCTTGCT
TCCCATTACAAGTCTGAGTCTAAAATTATTTTCGGAGTTATGAACGAACCTCACGATGTTCCTAAC
ATAAATACTTGGGCTGCTACCGTTCAAGAGGTCGTGACGGCTATCAGAAATGCTGGTGCAACTTCG
CAAATGATCCTGCTTCCAGGAAACGACTGGCAGTCAGCTGGGGCTTTTATAAGTGATGGATCGGCC
GCTGCATTATCGCAGGTCACAAACCCAGACGGGTCTACTACCAATCTAATTTTCGATGTTCATAAA
TATCTCGATTCTGATAACAGTGGTACACATACTGAGTGTGTCACTAATAACATTGATGGAGCATTC
TCACCGTTGGCTACCTGGCTCAGAACGAATAAAAGACAAGCCATGTTGACGGAAACAGGTGGTGGT
AATACTGCTAGTTGTGAAACATATCTGTGTCAGCAAATCCAGTACTTGAATCAGAATAGCGATGTG
TACCTGGGGTACGTTGGGTGGGGTGCCGGCTCATTTGACTCTACCTATGCACTAGACGAAACGCCA
ACTTCAAGTGGTAACTCATGGACCGATACACCATTAGTTAAAGCTTGCTTAGCTAGGAAG

12002013:

GGCAAGGTCAGATTTGCGGGTGTCAACATCGCTGGCTTCGACTTCGGTGTTGTTACATCAGGCACG
CAAGACATGACTCAAATATATCCCCCAGTAAAAGATATGCCCCCTTACTATAACAACCCAGACGGA
GCTGGGCAGATGCAACACTTTGTCAACGAAGATACATTCAATATCTTTCGACTTCCCTCTGGATGG
CAATTTTTGGTAAACAATAATTTGGGTGGTACTCTAGATAGCAATAATTTCGCATACTATGATCAA
CTGGTTCAATCCTGTCTCAGCCTAGGAGCATATTGCATTGTGGACGTACATAATTACGCGAGATGG
AACGGTGGCGTAATAGGGCAAGGCGGTCCAACAAATGCACAGTTCACTTCGCTATGGTCTCAATTA
GCGAGTCACTATAAGTCAGAATCGAAAATCATCTTTGGGGTTATGAATGAACCCCATGACGTTCCA
AATATCAACACTTGGGCTGCTACAGTTCAGGAAGTTGTGACTGCTATTAGGAATGCTGGTGCTACA
TCACAAATGATTCTGCTGCCGGGTAATGATTGGCAATCAGCTGGTGCTTTTATTAGCGACGGGTCA
GCTGCTGCTTTGTCACAGGTTACCAATCCCGACGGTAGCACTACAAATCTGATATTCGATGTTCAT
AAATATCTTGATTCTGACAACAGCGGTACACACACAGAATGTGTAACTAACAATATCGACGGTGCT
TTTTCACCTTTAGCTACCTGGTTGAGAACGAATAAAAGACAGGCTATGTTAACCGAAACAGGAGGA
GGTAACACTGCCAGTTGTGAAACCTATCTGTGCCAACAATTGGATTTTTTGAACGCTAACTCTGAT
GTCTATTTAGGCTGGACTTCTTGGGGTGCAGGGTCATTCGACTCGACATATGCCTTGGATGAAACC
CCTACTTCTTCCGGTAACAGTTGGACGGATACTCCTCTCGTTAAAGCATGTTTTGTTCCAAAGTGG
AAATCT

02000000:

GGCGTGCGTTTTGCAGGTGTTAACATCGCTGGATTTGATTTCGGTGTTGTTACCTCCGGAACACAA
GACATGACACAAATTTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGAAGACGGTATGACAATATTTCGTTTGCCTAGTGGATGGCAA
TTTCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTAACAATTTTGCTAAATATGATCAATTG

GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATGTCCATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTCATTACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
AATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGGTGCTTTCATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGGTGCCTTCAGC
CCGTTGGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGTAAT
GTACAAAGTTGTATTCAGGATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TCCAGTGGTAATAGTTGGACCGATACTTCTTTGGTATCTTCTTGCTTAGCTAGAAAG

13002010:

GGCGTCAGATTTGCGGGTGTCAACATTGCAGGTTTCGATTTTGGTATGGTCACGTCAGGCACCCAG
GACTTGACGCAAATATACCCTCCTGTCAAGGATATGCCCCCATACTATAATAATCCAGATGGGGCA
GGACAAATGCAGCATTTTGTTAATGAGGACGGTATGACTATATTCAGGTTACCAACTGGCTGGCAG
TACCTTGTCAATAACAACTTAGGTGGTACATTAGATGCCACAAATTTTGGTTACTATGACCAACTA
GTACAAAGTTGTCTAAGTTTAGGGGCATATTGCATCGTTGATATCCATAACTACGCAAGGTGGAAC
GGCGGTGTAATCGGACAGGGTGGACCAACGAATGCTCAATTCACGAGTCTGTGGTCTCAACTGGCG
TCTAAGTACAAGTCTGAAAGTAAAATAATTTTCGGGGTTATGAATGAACCCCACGACGTCCCAAAC
ATAAACACATGGGCTGCTACTGTTCAGGAAGTTGTTACAGCAATCAGAAATGCTGGTGCAACTTCG
CAAATGATCCTGCTTCCAGGAAACGACTGGCAGTCAGCTGGGGCTTTTATAAGTGATGGATCGGCC
GCTGCATTATCGCAGGTCACAAACCCAGACGGGTCTACTACCAATCTAATTTTCGATGTTCATAAA
TATCTCGATTCTGATAACAGTGGTACACATACTGAGTGTGTCACTAATAACATTGATGGAGCATTC
TCACCGTTGGCTACCTGGCTCAGAACGAATAAAAGACAAGCCATGTTGACGGAAACAGGTGGTGGT
AATACTGCTAGTTGTGAAACATATCTGTGTCAGCAAATCCAGTACTTGAATCAGAATAGCGATGTG
TACCTGGGGTACGTTGGGTGGGGTGCCGGCTCATTTGACTCTACCTATGCACTAGACGAAACGCCA
ACTTCAAGTGGTAACTCATGGACCGATACACCATTAGTTAAAGCTTGCTTAGCTAGGAAG

13002013:

GGCAAGGTCAGATTTGCGGGTGTCAACATAGCAGGTTTCGATTTTGGTATGGTTACCTCTGGAACT
CAAGATCTTACTCAGATCTATCCACCTGTCAAAGATATGCCACCATATTATAACAATCCTGATGGT
GCTGGTCAAATGCAACATTTTGTGAATGAGGACACCTTCAACATATTCCGTTTGCCTACTGGTTGG
CAGTATCTAGTCAATAATAACCTTGGAGGGACATTGGACGCTACTAATTTTGGTTACTATGATCAA
TTAGTCCAATCCTGCCTTTCCCTAGGAGCCTATTGTATAGTGGATATACACAATTATGCGAGATGG
AACGGTGGCGTGATCGGTCAAGGTGGCCCAACTAACGCTCAGTTCACCTCTCTATGGTCTCAATTG
GCATCCAAGTACAAGTCTGAGTCTAAAATTATTTTCGGTGTTATGAATGAACCCCATGATGTCCCC
AATATAAACACTTGGGCCGCGACCGTACAAGAAGTAGTCACTGCAATTAGAAACGCTGGTGCTACA
TCACAAATGATTCTGCTGCCGGGTAATGATTGGCAATCAGCTGGTGCTTTTATTAGCGACGGGTCA
GCTGCTGCTTTGTCACAGGTTACCAATCCCGACGGTAGCACTACAAATCTGATATTCGATGTTCAT
AAATATCTTGATTCTGACAACAGCGGTACACACACAGAATGTGTAACTAACAATATCGACGGTGCT
TTTTCACCTTTAGCTACCTGGTTGAGAACGAATAAAAGACAGGCTATGTTAACCGAAACAGGAGGA
GGTAACACTGCCAGTTGTGAAACCTATCTGTGCCAACAATTGGATTTTTTGAACGCTAACTCTGAT
GTCTATTTAGGCTGGACTTCTTGGGGTGCAGGGTCATTCGACTCGACATATGCCTTGGATGAAACC
CCTACTTCTTCCGGTAACAGTTGGACGGATACTCCTCTCGTTAAAGCATGTTTTGTTCCAAAGTGG
AAATCT

00000110: (best predicted chimera)

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACTTGT
GTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGAAGACGGTATGACAATATTTCGTTTGCCTGTTGGATGGCAA
TATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTTCTATT
ACTACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGCAGCTGTTATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATACTGAATGTGTTACCAATAATATCGATGACGCTTTTGCT
CCATTGGCAACCTGGCTGAGGCAGAACAATAGACAAGCTATTCTTACTGAGACTGGAGGAGGTAAT
ACCGCATCTTGCGAGACATATCTGTGCCAACAAATACAATACTTGAATCAGAACAGCGATGTTTAT
TTAGGTTACGTTGGCTGGGGTGCGGGATCATTTGATAGCACATACGCGCTTGATGAAACACCAACA
TCTTCCGGTAATTCATGGACTGACACTCCACTCGTAAAAGCTTGTCTTGCTAGGAAA
```

110F: (00000110 with deleterious mutations removed)

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACTTGT
GTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGTTCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGAAGACGGTATGACAATATTTCGTTTGCCTGTTGGATGGCAA
TATCTGGTCAACAATAACCTGGGAGGTAATTTAGATAGTACCTCTATCTCCAAATATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTGTCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGTAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
AATACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTGCAGCTGTTATTTCTGACGGCAGTGCCGCT
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTACTACTAACAATATCGATGACGCTTTTGCT
CCATTGGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGTAAT
ACTCAAAGTTGTATTCAGTACATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGGTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TCCAGTGGTAATAGTTGGACCGATACTCCATTGGTATCTTCTTGCTTAGCTAGAAAG
```

CHAPTER 3:

OptCel5a:

```
GGCGTTAGATTTGCCGGTGTTAATATTGCTGGTTTTGACTTCGGTTGCACTACCGATGGCACTTGT
GTTACTTCTAAGGTCTATCCTCCGCTTAAGAACTTTACGGGATCCAACAACTATCCTGATGGAATA
GGGCAGATGCAACATTTTGTTAATGATGACGGTATGAATATATTTCGTTTGCCTGTTGGATGGCAA
TATCTGGTCAACAATAACCTGGGAGGTAATTTAGATCCTGAGTCTATCTCCAAATATGATCAATTG
GTCCAAGGTTGTCTATCCTTAGGTGCATATTGTATTATCGATATACATAATTATGCTAGATGGAAT
GGCGGTATTATTGGTCAAGGCGGTCCAACAAACGCGCAATTTACTTCATTGTGGAGCCAGTTGGCT
AGGAAATACGCGTCACAGTCCAGGGTTTGGTTTGGAATTATGAATGAGCCACACGATGTTAACATT
GAGACCTGGGCTGCTACCGTTCAAGAAGTTGTCACAGCAATTAGAAATGCTGGCGCTACGTCCCAG
TTTATCAGTCTACCTGGTAATGATTGGCAATCTGCTAGTGCTGTTATTTCTGACGGCAGTGCCGCT
```

```
GCGTTGTCGCAAGTAACTAATCCAGATGGCTCCACAACTAATCTAATTTTCGACGTGCATAAGTAT
TTGGATTCGGATAATAGTGGTACTCATGCAGAGTGTGTTACTAACAATATCGATGAGGCCTTCAGC
CCGCTAGCAACCTGGTTACGTCAAAACAATAGACAAGCAATATTGACGGAAACCGGTGGTGGTAAT
ACTCAAAGTTGTATTCAGTATATGTGTCAACAAATACAGTACCTTAACCAAAACTCAGATGTTTAC
TTAGGCTACGTTGGCTGGGCTGCTGGTTCCTTCGACAGTACTTACGTTTTGACTGAGACACCTACA
TGGAGTGGTAATAGTTGGACCGATACTCCTTTGGTATCTTCTTGCTTAGCTAGAAAG
```

CHAPTER 4:

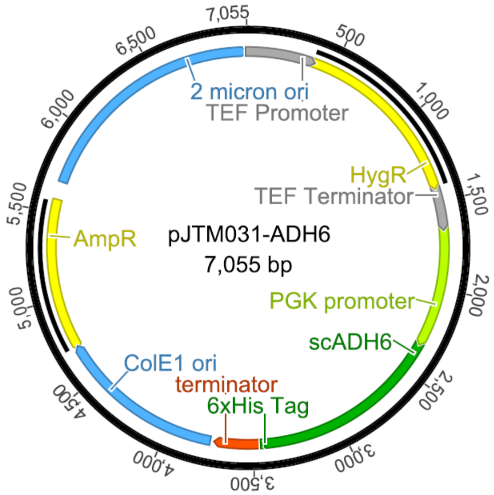
Gene	Source organism	Genbank ID
GlucD	*E. coli*	NC_000913.4
GalD	*E. coli*	NC_000913.5
ManD	*E. coli*	NC_000913.6
YfaW	*E. coli*	NC_000913.8
YfaW	*G. zeae*	XM_390059.1
YfaW	*P. pastoris*	XM_002490140.1

E. coli YfaW:

```
ATGACACTACCTAAGATCAAACAAGTTAGAGCATGGTTCACCGGAGGTGCAACAGCTGAGAAAGGC
GCTGGTGGAGGCGATTACCATGACCAAGGTGCCAATCATTGGATCGATGATCATATAGCTACACCA
ATGTCTAAGTATAGAGATTACGAACAATCTAGACAGTCTTTTGGTATCAATGTGCTTGGCACTTTA
GTAGTTGAAGTCGAAGCTGAAAATGGCCAAACTGGTTTTGCTGTCTCAACAGCAGGCGAAATGGGT
TGCTTTATCGTGGAAAAACACTTAAACAGGTTCATCGAGGGGAAATGTGTATCCGACATCAAATTG
ATACACGATCAAATGTTGAGTGCAACATTGTACTATAGTGGTTCTGGTGGTCTAGTGATGAATACT
ATCTCATGCGTCGATTTGGCCTTATGGGATCTGTTTGGCAAGGTAGTCGGACTTCCAGTATACAAG
CTACTTGGCGGAGCTGTCAGAGATGAAATCCAGTTTTACGCTACCGGTGCCAGACCAGACTTGGCA
AAAGAGATGGGCTTCATTGGTGGCAAAATGCCTACACATTGGGGTCCACATGACGGTGACGCTGGT
ATTAGAAAGGATGCAGCAATGGTTGCTGATATGAGAGAAAAGTGCGGGGAAGATTTCTGGCTGATG
CTTGACTGTTGGATGTCACAAGATGTGAACTACGCTACTAAGTTAGCACACGCCTGTGCTCCTTAC
AACTTAAAGTGGATTGAGGAATGCCTGCCACCTCAGCAATATGAGTCTTACAGAGAACTGAAGAGA
AACGCCCCAGTAGGTATGATGGTTACTTCCGGAGAGCACCACGGAACTCTACAATCTTTTAGAACC
TTATCTGAAACAGGGATTGACATAATGCAACCAGACGTTGGGTGGTGTGGAGGCTTAACAACTTTG
GTTGAAATTGCCGCAATCGCCAAATCAAGAGGTCAGTTAGTTGTTCCACATGGAAGTTCTGTGTAC
TCTCATCATGCTGTGATAACATTCACTAACACTCCATTCTCCGAATTTCTGATGACATCACCTGAT
TGTTCCACCATGCGTCCACAATTTGACCCAATTCTATTGAATGAGCCTGTCCCTGTTAATGGTAGA
ATACACAAATCCGTCTTGGATAAACCAGGGTTCGGGGTAGAGCTAAACAGAGATTGTAATCTTAAA
CGTCCTTATTCA
```

CHAPTER 5:

pJTM031-ADH6 vector map:



S. cerevisiae ADH6:

```
ATGTCTTATCCTGAGAAATTTGAAGGTATCGCTATTCAATCACACGAAGATTGGAAAAACCCAAAG
AAGACAAAGTATGACCCAAAACCATTTTACGATCATGACATTGACATTAAGATCGAAGCATGTGGT
GTCTGCGGTAGTGATATTCATTGTGCAGCTGGTCATTGGGGCAATATGAAGATGCCGCTAGTCGTT
GGTCATGAAATCGTTGGTAAAGTTGTCAAGCTAGGGCCCAAGTCAAACAGTGGGTTGAAAGTCGGT
CAACGTGTTGGTGTAGGTGCTCAAGTCTTTTCATGCTTGGAATGTGACCGTTGTAAGAATGATAAT
GAACCATACTGCACCAAGTTTGTTACCACATACAGTCAGCCTTATGAAGACGGCTATGTGTCGCAG
GGTGGCTATGCAAACTACGTCAGAGTTCATGAACATTTTGTGGTGCCTATCCCAGAGAATATTCCA
TCACATTTGGCTGCTCCACTATTATGTGGTGGTTTGACTGTGTACTCTCCATTGGTTCGTAACGGT
TGCGGTCCAGGTAAAAAAGTTGGTATAGTTGGTCTTGGTGGTATCGGCAGTATGGGTACATTGATT
TCCAAAGCCATGGGGGCAGAGACGTATGTTATTTCTCGTTCTTCGAGAAAAAGAGAAGATGCAATG
AAGATGGGCGCCGATCACTACATTGCTACATTAGAAGAAGGTGATTGGGGTGAAAAGTACTTTGAC
ACCTTCGACCTGATTGTAGTCTGTGCTTCCTCCCTTACCGACATTGACTTCAACATTATGCCAAAG
GCTATGAAGGTTGGTGGTAGAATTGTCTCAATCTCTATACCAGAACAACACGAAATGTTATCGCTA
AAGCCATATGGCTTAAAGGCTGTCTCCATTTCTTACAGTGCTTTAGGTTCCATCAAAGAATTGAAC
CAACTCTTGAAATTAGTCTCTGAAAAAGATATCAAAATTTGGGTGGAAACATTACCTGTTGGTGAA
GCCGGCGTCCATGAAGCCTTCGAAAGGATGGAAAAGGGTGACGTTAGATATAGATTTACCTTAGTC
GGCTACGACAAAGAATTTTCAGACTAG
```

S. cerevisiae ADH1:

```
ATGTCTATCCCAGAAACTCAAAAAGGTGTTATCTTCTACGAATCCCACGGTAAGTTGGAATACAAA
GATATTCCAGTTCCAAAGCCAAAGGCCAACGAATTGTTGATCAACGTTAAATACTCTGGTGTCTGT
CACACTGACTTGCACGCTTGGCACGGTGACTGGCCATTGCCAGTTAAGCTACCATTAGTCGGTGGT
CACGAAGGTGCCGGTGTCGTTGTCGGCATGGGTGAAAACGTTAAGGGCTGGAAGATCGGTGACTAC
GCCGGTATCAAATGGTTGAACGGTTCTTGTATGGCCTGTGAATACTGTGAATTGGGTAACGAATCC
AACTGTCCTCACGCTGACTTGTCTGGTTACACCCACGACGGTTCTTTCCAACAATACGCTACCGCT
GACGCTGTTCAAGCCGCTCACATTCCTCAAGGTACCGACTTGGCCCAAGTCGCCCCCATCTTGTGT
```

```
GCTGGTATCACCGTCTACAAGGCTTTGAAGTCTGCTAACTTGATGGCCGGTCACTGGGTTGCTATC
TCCGGTGCTGCTGGTGGTCTAGGTTCTTTGGCTGTTCAATACGCCAAGGCTATGGGTTACAGAGTC
TTGGGTATTGACGGTGGTGAAGGTAAGGAAGAATTATTCAGATCCATCGGTGGTGAAGTCTTCATT
GACTTCACTAAGGAAAAGGACATTGTCGGTGCTGTTCTAAAGGCCACTGACGGTGGTGCTCACGGT
GTCATCAACGTTTCCGTTTCCGAAGCCGCTATTGAAGCTTCTACCAGATACGTTAGAGCTAACGGT
ACCACCGTTTTGGTCGGTATGCCAGCTGGTGCCAAGTGTTGTTCTGATGTCTTCAACCAAGTCGTC
AAGTCCATCTCTATTGTTGGTTCTTACGTCGGTAACAGAGCTGACACCAGAGAAGCTTTGGACTTC
TTCGCCAGAGGTTTGGTCAAGTCTCCAATCAAGGTTGTCGGCTTGTCTACCTTGCCAGAAATTTAC
GAAAAGATGGAAAAGGGTCAAATCGTTGGTAGATACGTTGTTGACACTTCTAAACACCACCACCAC
CACCACTGA
```

S. cerevisiae ARI1:

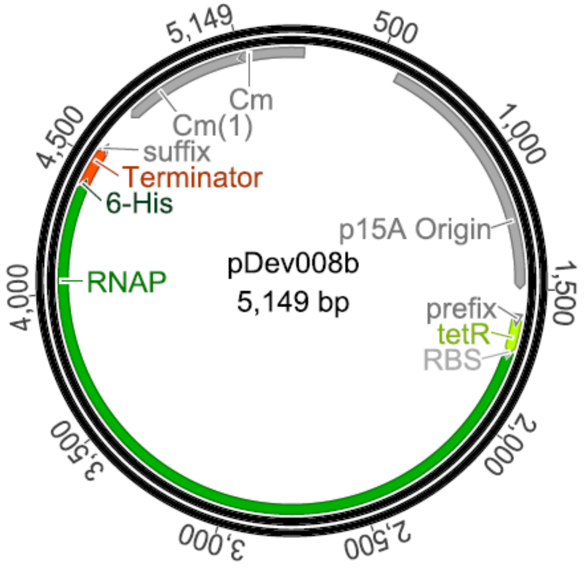
```
ATGACTACTGATACCACTGTTTTCGTTTCTGGCGCAACCGGTTTCATTGCTCTACACATTATGAAC
GATCTGTTGAAAGCTGGCTATACAGTCATCGGCTCAGGTAGATCTCAAGAAAAAAATGATGGCTTG
CTCAAAAAATTTAATAACAATCCCAAACTATCGATGGAAATTGTGGAAGATATTGCTGCTCCAAAC
GCCTTTGATGAAGTTTTCAAAAAACATGGTAAGGAAATTAAGATTGTGCTACACACTGCCTCCCCA
TTCCATTTTGAAACTACCAATTTTGAAAAGGATTTACTAACCCCTGCAGTGAACGGTACAAAATCT
ATCTTGGAAGCGATTAAAAAATATGCTGCAGACACTGTTGAAAAAGTTATTGTTACTTCGTCTACT
GCTGCTCTGGTGACACCTACAGACATGAACAAAGGAGATTTGGTGATCACGGAGGAGAGTTGGAAT
AAGGATACATGGGACAGTTGTCAAGCCAACGCCGTTGCCGCATATTGTGGCTCGAAAAAGTTTGCT
GAAAAAACTGCTTGGGAATTTCTTAAAGAAAACAAGTCTAGTGTCAAATTCACACTATCCACTATC
AATCCGGGATTCGTTTTTGGTCCTCAAATGTTTGCAGATTCGCTAAAACATGGCATAAATACCTCC
TCAGGGATCGTATCTGAGTTAATTCATTCCAAGGTAGGTGGAGAATTTTATAATTACTGTGGCCCA
TTTATTGACGTGCGTGACGTTTCTAAAGCCCACCTAGTTGCAATTGAAAAACCAGAATGTACCGGC
CAAAGATTAGTATTGAGTGAAGGTTTATTCTGCTGTCAAGAAATCGTTGACATCTTGAACGAGGAA
TTCCCTCAATTAAAGGGCAAGATAGCTACAGGTGAACCTGCGACCGGTCCAAGCTTTTTAGAAAAA
AACTCTTGCAAGTTTGACAATTCTAAGACAAAAAAACTACTGGGATTCCAGTTTTACAATTTAAAG
GATTGCATAGTTGACACCGCGGCGCAAATGTTAGAAGTTCAAAATGAAGCCCACCACCACCACCAC
CACTGA
```

S. cerevisiae GRE2:

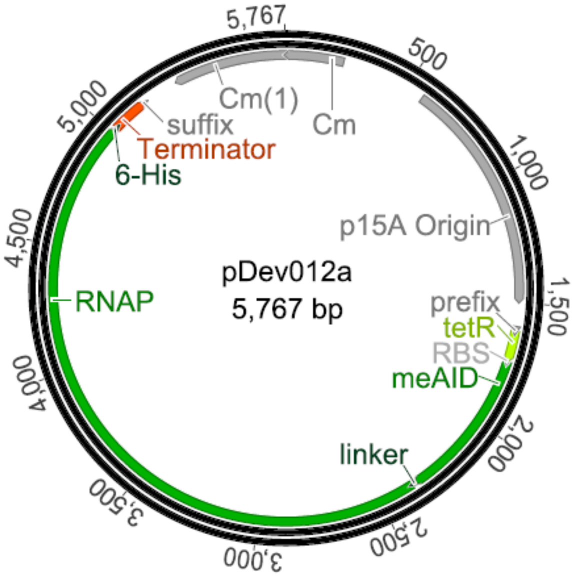
```
ATGTCAGTTTTCGTTTCAGGTGCTAACGGGTTCATTGCCCAACACATTGTCGATCTCCTGTTGAAG
GAAGACTATAAGGTCATCGGTTCTGCCAGAAGTCAAGAAAAGGCCGAGAATTTAACGGAGGCCTTT
GGTAACAACCCAAAATTCTCCATGGAAGTTGTCCCAGACATATCTAAGCTGGACGCATTTGACCAT
GTTTTCCAAAAGCACGGCAAGGATATCAAGATAGTTCTACATACGGCCTCTCCATTCTGCTTTGAT
ATCACTGACAGTGAACGCGATTTATTAATTCCTGCTGTGAACGGTGTTAAGGGAATTCTCCACTCA
ATTAAAAAATACGCCGCTGATTCTGTAGAACGTGTAGTTCTCACCTCTTCTTATGCAGCTGTGTTC
GATATGGCAAAAGAAAACGATAAGTCTTTAACATTTAACGAAGAATCCTGGAACCCAGCTACCTGG
GAGAGTTGCCAAAGTGACCCAGTTAACGCCTACTGTGGTTCTAAGAAGTTTGCTGAAAAAGCAGCT
TGGGAATTTCTAGAGGAGAATAGAGACTCTGTAAAATTCGAATTAACTGCCGTTAACCCAGTTTAC
GTTTTTGGTCCGCAAATGTTTGACAAAGATGTGAAAAAACACTTGAACACATCTTGCGAACTCGTC
AACAGCTTGATGCATTTATCACCAGAGGACAAGATACCGGAACTATTTGGTGGATACATTGATGTT
CGTGATGTTGCAAAGGCTCATTTAGTTGCCTTCCAAAAGAGGGAAACAATTGGTCAAAGACTAATC
GTATCGGAGGCCAGATTTACTATGCAGGATGTTCTCGATATCCTTAACGAAGACTTCCCTGTTCTA
AAAGGCAATATTCCAGTGGGGAAACCAGGTTCTGGTGCTACCCATAACACCCTTGGTGCTACTCTT
GATAATAAAAAGAGTAAGAAATTGTTAGGTTTCAAGTTCAGGAACTTGAAAGAGACCATTGACGAC
ACTGCCTCCCAAATTTTAAAATTTGAGGGCAGAATACACCACCACCACCACCACTGA
```

CHAPTER 6:

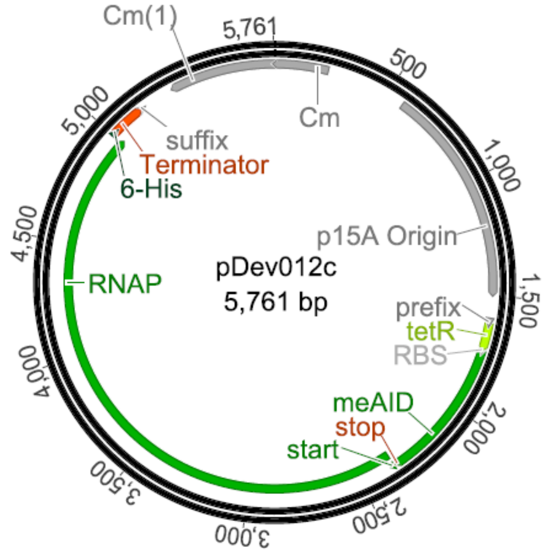
pDEV008b vector map:



pDev012a vector map:



pDev012c vector map:



Cm(1)
5,761
500
Cm
1,000
p15A Origin
1,500
prefix
tetR
RBS
2,000
meAID
stop
start
2,500
3,000
3,500
4,000
4,500
RNAP
5,000
suffix
Terminator
6-His
pDev012c
5,761 bp

APPENDIX 3: MATLAB CODE

CHAPTER 2:

Boltzmann 4 parameter sigmoidal curve:

```
clear
clc
clf
format compact
% the data, x is temperatures, y is measurements
     x=[63.3 65.6 68.2 70.9 73.6 76 77.9 79.3 79.9];
     y=[230.1372373 254.4959361 257.2802309 254.6480718 213.3706537
158.9781824 105.4681547 84.63925781 96.36978128];
% function
     fh=@(b,x) b(1)+ b(2)./(1 + exp(-(x-b(3))/(b(4))));
% guess values for parameters (beta0)
     b0=[0.3396 0.8871 80 -1.0399];
% third parameter is expected t50
% plot the raw data
     plot(x,y,'s','markersize',5,'color',[0,0,0]);
     hold on
% determine best fit values for coefficient (bhat)
     bhat=nlinfit(x,y,fh,b0);
% plot the fit
     xf = linspace(x(1), x(length(x)));
     plot(xf,fh(bhat,xf),'linewidth',1,'color',[1,0,0]);
     legend('original data','fit data','location','Best')% the result
     xlabel('Temperature (C)')
     ylabel('Signal')
     bhat(1)
     bhat(2)
% The parameter bhat(3) is the desired TA50
TA50=bhat(3)
```

Linear regression model for HjCel5a chimera library: (With contact penalty)

```
clear all
close all

% thermostability values
C=[69.14815
.
.
75.6344]
% contact penalties
E=[9
.
.
3]
% chimeras
```

```
n(1)=base2dec('00012032',4)+1;
.
.
n(48)=base2dec('00000110',4)+1;

%First make a matrix with block indices for each chimera

A=zeros(4^8,24);

for k=1:4^8;
   %8th block
if mod(k,4)==2;
    A(k,22)=1;
end
if mod(k,4)==3;
    A(k,23)=1;
end
if mod(k,4)==0;
    A(k,24)=1;
end
%7th block

if mod(ceil(k/4),4)==2;
    A(k,19)=1;
end
if mod(ceil(k/4),4)==3;
    A(k,20)=1;
end
if mod(ceil(k/4),4)==0;
    A(k,21)=1;
end
%6th block
if mod(ceil(k/16),4)==2;
    A(k,16)=1;
end
if mod(ceil(k/16),4)==3;
    A(k,17)=1;
end
if mod(ceil(k/16),4)==0;
    A(k,18)=1;
end
%5th block
if mod(ceil(k/4^3),4)==2;
    A(k,13)=1;
end
if mod(ceil(k/4^3),4)==3;
    A(k,14)=1;
end
if mod(ceil(k/4^3),4)==0;
    A(k,15)=1;
end
```

```
%4th block
if mod(ceil(k/4^4),4)==2;
    A(k,10)=1;
end
if mod(ceil(k/4^4),4)==3;
    A(k,11)=1;
end
if mod(ceil(k/4^4),4)==0;
    A(k,12)=1;
end
%3th block
if mod(ceil(k/4^5),4)==2;
    A(k,7)=1;
end
if mod(ceil(k/4^5),4)==3;
    A(k,8)=1;
end
if mod(ceil(k/4^5),4)==0;
    A(k,9)=1;
end
%2nd block
if mod(ceil(k/4^6),4)==2;
    A(k,4)=1;
end
if mod(ceil(k/4^6),4)==3;
    A(k,5)=1;
end
if mod(ceil(k/4^6),4)==0;
    A(k,6)=1;
end
%1st block
if mod(ceil(k/4^7),4)==2;
    A(k,1)=1;
end
if mod(ceil(k/4^7),4)==3;
    A(k,2)=1;
end
if mod(ceil(k/4^7),4)==0;
    A(k,3)=1;
end
end

%assign chimeras to blocks, and do regression

D=zeros(size(n),26);
for i=1:size(n);
    m=n(i);
for y=1:24;
    D(i,y)=A(m,y);
end
end
for j=1:size(n);
D(j,25)=1;
end
```

```
for j=1:size(n);
    D(j,26)=E(j);
end
D

test = regress(C,D)

Predictedchim=D*test;
X=C;
Y=Predictedchim;

figure(1)
plot(X,Y,'o')
scatter(X,Y)
xlabel('Actual Chimera A50 (\circC)','FontSize',16,'FontName','Arial')
ylabel('Predicted Chimera A50
(\circC)','FontSize',16,'FontName','Arial')
title('Linear regression model','FontSize',16)
```

BIBLIOGRAPHY

1. Otto, S. P. (2009). The evolutionary enigma of sex. *the american naturalist* **174**, S1-S14.

2. Romero, P. A. & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology* **10**, 866-876.

3. Crameri, A., Raillard, S.-A., Bermudez, E. & Stemmer, W. P. (1998). DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **391**, 288-291.

4. Carbone, M. N. & Arnold, F. H. (2007). Engineering by homologous recombination: exploring sequence and function within a conserved fold. *Current opinion in structural biology* **17**, 454-459.

5. Jochens, H. & Bornscheuer, U. T. (2010). Natural diversity to guide focused directed evolution. *ChemBioChem* **11**, 1861-1866.

6. Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O. & Arnold, F. H. (2005). On the conservative nature of intragenic recombination. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5380-5385.

7. Voigt, C. A., Martinez, C., Wang, Z.-G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nature Structural & Molecular Biology* **9**, 553-558.

8. Romero, P. A. & Arnold, F. H. (2012). Random field model reveals structure of the protein recombinational landscape. *PLoS computational biology* **8**, e1002713.

9. Romero, P. A., Krause, A. & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences of the United States of America* **110**, E193-E201.

10. Li, Y., Drummond, D. A., Sawayama, A. M., Snow, C. D., Bloom, J. D. & Arnold, F. H. (2007). A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nature biotechnology* **25**, 1051-1056.

11. Buske, F. A., Their, R., Gillam, E. M. & Bodén, M. (2009). In silico characterization of protein chimeras: relating sequence and function within the same fold. *Proteins: Structure, Function, and Bioinformatics* **77**, 111-120.

12. Heinzelman, P., Snow, C. D., Wu, I., Nguyen, C., Villalobos, A., Govindarajan, S., Minshull, J. & Arnold, F. H. (2009). A family of thermostable fungal cellulases created by structure-guided recombination. *Proc Natl Acad Sci U S A* **106**, 5610-5.

13. Heinzelman, P., Snow, C. D., Smith, M. A., Yu, X., Kannan, A., Boulware, K., Villalobos, A., Govindarajan, S., Minshull, J. & Arnold, F. H. (2009). SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *Journal of Biological Chemistry* **284**, 26229-26233.

14. Heinzelman, P., Komor, R., Kanaan, A., Romero, P., Yu, X., Mohler, S., Snow, C. & Arnold, F. (2010). Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Engineering Design and Selection* **23**, 871-880.

15. Smith, M. A., Rentmeister, A., Snow, C. D., Wu, T., Farrow, M. F., Mingardon, F. & Arnold, F. H. (2012). A diverse set of family 48 bacterial glycoside hydrolase cellulases created by structure-guided recombination. *FEBS Journal* **279**, 4453-4465.

16. Romero, P. A., Stone, E., Lamb, C., Chantranupong, L., Krause, A., Miklos, A. E., Hughes, R. A., Fechtel, B., Ellington, A. D. & Arnold, F. H. (2012). SCHEMA-designed variants of human arginase I and II reveal sequence elements important to stability and catalysis. *ACS synthetic biology* **1**, 221-228.

17. Ho, M. L., Adler, B. A., Torre, M. L., Silberg, J. J. & Suh, J. (2013). SCHEMA computational design of virus capsid chimeras: calibrating how genome packaging, protection, and transduction correlate with calculated structural disruption. *ACS synthetic biology*.

18. Saraf, M. C., Horswill, A. R., Benkovic, S. J. & Maranas, C. D. (2004). FamClash: a method for ranking the activity of engineered enzymes. *Proceedings of the National Academy of Sciences* **101**, 4142-4147.

19. Pantazes, R. J., Saraf, M. C. & Maranas, C. D. (2007). Optimal protein library design using recombination or point mutations based on sequence-based scoring functions. *Protein Engineering Design and Selection* **20**, 361-373.

20. Ye, X., Friedman, A. M. & Bailey-Kellogg, C. (2007). Hypergraph model of multi-residue interactions in proteins: sequentially-constrained partitioning algorithms for optimization of site-directed protein recombination. *J Comput Biol* **14**, 777-90.

21. Zheng, W., Friedman, A. M. & Bailey-Kellogg, C. (2008). *Research in Computational Molecular Biology*.

22. Smith, M. A., Romero, P. A., Wu, T., Brustad, E. M. & Arnold, F. H. (2013). Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein Science* **22**, 231-238.

23. Smith, M. A., Bedbrook, C. N., Wu, T. & Arnold, F. H. (2013). Hypocrea jecorina Cellobiohydrolase I Stabilizing Mutations Identified Using Noncontiguous Recombination. *ACS Synthetic Biology*.

24. Bharat, T. A., Eisenbeis, S., Zeth, K. & Höcker, B. (2008). A βα-barrel built by the combination of fragments from different folds. *Proceedings of the National Academy of Sciences* **105**, 9942-9947.

25. Eisenbeis, S., Proffitt, W., Coles, M., Truffault, V., Shanmugaratnam, S., Meiler, J. & Höcker, B. (2012). Potential of fragment recombination for rational design of proteins. *Journal of the American Chemical Society* **134**, 4019-4022.

26. Shanmugaratnam, S., Eisenbeis, S. & Höcker, B. (2012). A highly stable protein chimera built from fragments of different folds. *Protein Engineering Design and Selection* **25**, 699-703.

27. Zheng, W., Griswold, K. E. & Bailey-Kellogg, C. (2010). Protein fragment swapping: a method for asymmetric, selective site-directed recombination. *Journal of Computational Biology* **17**, 459-475.

28. Sharma, D., Cao, Y. & Li, H. (2006). Engineering proteins with novel mechanical properties by recombination of protein fragments. *Angewandte Chemie International Edition* **45**, 5633-5638.

29. Balamurali, M., Sharma, D., Chang, A., Khor, D., Chu, R. & Li, H. (2008). Recombination of protein fragments: A promising approach toward engineering proteins with novel nanomechanical properties. *Protein Science* **17**, 1815-1826.

30. Lu, W., Negi, S. S., Oberhauser, A. F. & Braun, W. (2012). Engineering proteins with enhanced mechanical stability by force-specific sequence motifs. *Proteins: Structure, Function, and Bioinformatics* **80**, 1308-1315.

31. Ng, S. P., Billings, K. S., Ohashi, T., Allen, M. D., Best, R. B., Randles, L. G., Erickson, H. P. & Clarke, J. (2007). Designing an extracellular matrix protein with enhanced mechanical stability. *Proceedings of the National Academy of Sciences* **104**, 9633-9637.

32. Ng, S. P., Billings, K., Randles, L. & Clarke, J. (2008). Manipulating the stability of fibronectin type III domains by protein engineering. *Nanotechnology* **19**, 384023.

33. Clouthier, C. M., Morin, S., Gobeil, S. M., Doucet, N., Blanchet, J., Nguyen, E., Gagné, S. M. & Pelletier, J. N. (2012). Chimeric β-Lactamases: Global Conservation of Parental Function and Fast Time-Scale Dynamics with Increased Slow Motions. *PloS one* **7**, e52283.

34. Chen, C. K., Berry, R. E., Shokhireva, T., Murataliev, M. B., Zhang, H. & Walker, F. A. (2010). Scanning chimeragenesis: the approach used to change the substrate selectivity of fatty acid monooxygenase CYP102A1 to that of terpene omega-hydroxylase CYP4C7. *J Biol Inorg Chem* **15**, 159-74.

35. Campbell, E., Chuang, S. & Banta, S. (2013). Modular exchange of substrate-binding loops alters both substrate and cofactor specificity in a member of the aldo-keto reductase superfamily. *Protein Eng Des Sel* **26**, 181-6.

36. van Beek, H. L., de Gonzalo, G. & Fraaije, M. W. (2012). Blending Baeyer–Villiger monooxygenases: using a robust BVMO as a scaffold for creating chimeric enzymes with novel catalytic properties. *Chemical Communications* **48**, 3288-3290.

37. Jones, D. D. (2011). Recombining low homology, functionally rich regions of bacterial subtilisins by combinatorial fragment exchange. *PloS one* **6**, e24319.

38. Yizhar, O., Fenno, L. E., Prigge, M., Schneider, F., Davidson, T. J., O'Shea, D. J., Sohal, V. S., Goshen, I., Finkelstein, J. & Paz, J. T. (2011). Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171-178.

39. Wang, H., Sugiyama, Y., Hikima, T., Sugano, E., Tomita, H., Takahashi, T., Ishizuka, T. & Yawo, H. (2009). Molecular determinants differentiating photocurrent properties of two channelrhodopsins from chlamydomonas. *Journal of Biological Chemistry* **284**, 5685-5696.

40. Lin, J. Y., Lin, M. Z., Steinbach, P. & Tsien, R. Y. (2009). Characterization of engineered channelrhodopsin variants with improved properties and kinetics. *Biophysical journal* **96**, 1803-1814.

41. Wen, L., Wang, H., Tanimoto, S., Egawa, R., Matsuzaka, Y., Mushiake, H., Ishizuka, T. & Yawo, H. (2010). Opto-current-clamp actuation of cortical neurons using a strategically designed channelrhodopsin. *PLoS One* **5**, e12893.

42. Heinzelman, P., Romero, P. A. & Arnold, F. H. (2013). Efficient sampling of SCHEMA chimera families to identify useful sequence elements. *Methods Enzymol* **523**, 351-68.

43. Trudeau, D. L., Smith, M. A. & Arnold, F. H. (2013). Innovation by homologous recombination. *Current Opinion in Chemical Biology*.

44. Heinzelman, P., Komor, R., Kanaan, A., Romero, P., Yu, X., Mohler, S., Snow, C. & Arnold, F. (2010). Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng Des Sel* **23**, 871-80.

45. Karypis, G. & Kumar, V. (2000). Multilevel k-way hypergraph partitioning. *VLSI design* **11**, 285-300.

46. Suominen, P. L., Mantyla, A. L., Karhunen, T., Hakola, S. & Nevalainen, H. (1993). High frequency one-step gene replacement in Trichoderma resei. II. Effects of deletions of individual cellulase genes. *Molecular Genetics and Genomics* **241**, 523-530.

47. Komor, R. S., Romero, P. A., Xie, C. B. & Arnold, F. H. (2012). Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. *Protein Eng Des Sel* **25**, 827-33.

48. Wu, I. & Arnold, F. H. (2013). Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. *Biotechnology and Bioengineering* **110**, 1874-1883.

49. Lee, T. M., Farrow, M. F., Arnold, F. H. & Mayo, S. L. (2011). A structural study of Hypocrea jecorina Cel5A. *Protein Science* **20**, 1935-1940.

50. Smith, M. A., Romero, P. A., Wu, T., Brustad, E. M. & Arnold, F. H. (2013). Chimeragenesis of distantly-related proteins by noncontiguous recombination. *Protein Sci* **22**, 231-8.

51. Zhao, J., Shi, P., Huang, H., Li, Z., Yuan, T., Yang, P., Luo, H., Bai, Y. & Yao, B. (2012). A novel thermoacidophilic and thermostable endo-β-1, 4-glucanase from Phialophora sp. G5: its thermostability influenced by a distinct β-sheet and the carbohydrate-binding module. *Applied microbiology and biotechnology* **95**, 947-955.

52. Wei, X.-M., Qin, Y.-Q. & Qu, Y.-B. (2010). Molecular cloning and characterization of two major endoglucanases from Penicillium decumbens. *Journal of microbiology and biotechnology* **20**, 265-270.

53. Jeya, M., Joo, A. R., Lee, K. M., Sim, W. I., Oh, D. K., Kim, Y. S., Kim, I. W. & Lee, J. K. (2010). Characterization of endo-beta-1,4-glucanase from a novel strain of Penicillium pinophilum KMJ601. *Appl Microbiol Biotechnol* **85**, 1005-14.

54. Chang, C.-J., Chang, H.-S., Lee, C.-C., Trudeau, D. L., Smith, M. A., Yua, S.-M., Ho, T.-H. D., Wang, A. H. J., Arnold, F. H. & Chao, Y.-C. (2013). Structure characterization and stability improvement of bacterial GH5 cellulases by SCHEMA structure-guided recombination.

55. Dana, C. M., Saija, P., Kal, S. M., Bryan, M. B., Blanch, H. W. & Clark, D. S. (2012). Biased clique shuffling reveals stabilizing mutations in cellulase Cel7A. *Biotechnology and Bioengineering* **109**, 2710-2719.

56. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994). Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *Journal of Molecular Biology* **240**, 188-192.

57. Guerois, R., Nielsen, J. E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology* **320**, 369-387.

58. Lee, T. M., Trudeau, D. L., Arnold, F. H. & Mayo, S. L. (in preparation). Thermostable Variants of Hypocrea jecorina Cel5A Engineered Through Various Means.

59. Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends in Genetics* **27**, 323-331.

60. Reetz, M. T. (2013). The Importance of Additive and Non-Additive Mutational Effects in Protein Engineering. *Angewandte Chemie International Edition*.

61. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *science* **312**, 111-114.

62. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535-538.

63. Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.

64. Krause, A. (2010). Sfo: A toolbox for submodular function optimization. *The Journal of Machine Learning Research* **11**, 1141-1144.

65. Georgescu, R., Bandara, G. & Sun, L. (2003). Saturation mutagenesis. In *Directed evolution library creation*, pp. 75-83. Springer.

66. Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J. & Govindarajan, S. (2006). Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *Bmc Bioinformatics* **7**, 285.

67. Gibson, D. G. (2011). Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol* **498**, 349-61.

68. Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D. & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In *The proteomics protocols handbook*, pp. 571-607. Springer.

69. Park, J. T. & Johnson, M. J. (1949). A submicrodetermination of glucose. *J Biol Chem* **181**, 149-51.

70. Smogyi, M. (1952). Notes on sugar determination. *J Biol Chem* **195**, 19-23.

71. Nelson, N. (1944). A photometric adaptation of the Somogyi method for the determination of glucose. *J. biol. Chem* **153**, 375-379.

72. Turner, P., Mamo, G. & Karlsson, E. N. (2007). Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microb Cell Fact* **6**, 1-23.

73. Viikari, L., Alapuranen, M., Puranen, T., Vehmaanperä, J. & Siika-Aho, M. (2007). Thermostable enzymes in lignocellulose hydrolysis. In *Biofuels*, pp. 121-145. Springer.

74. Mingardon, F., Bagert, J. D., Maisonnier, C., Trudeau, D. L. & Arnold, F. H. (2011). Comparison of family 9 cellulases from mesophilic and thermophilic bacteria. *Applied and environmental microbiology* **77**, 1436-1442.

75. Tomme, P., Warren, R. & Gilkes, N. (1995). Cellulose hydrolysis by bacteria and fungi. *Advances in microbial physiology* **37**, 1-81.

76. Murray, P., Aro, N., Collins, C., Grassick, A., Penttilä, M., Saloheimo, M. & Tuohy, M. (2004). Expression in Trichoderma reesei and characterisation of a thermostable family 3 β-glucosidase from the moderately thermophilic fungus Talaromyces emersonii. *Protein expression and purification* **38**, 248-257.

77. Suominen, P. L., Mäntylä, A. L., Karhunen, T., Hakola, S. & Nevalainen, H. (1993). High frequency one-step gene replacement in Trichoderma reesei. II. Effects of deletions of individual cellulase genes. *Molecular and General Genetics MGG* **241**, 523-530.

78. Rosgaard, L., Pedersen, S., Langston, J., Akerhielm, D., Cherry, J. R. & Meyer, A. S. (2007). Evaluation of minimal Trichoderma reesei cellulase mixtures on differently pretreated barley straw substrates. *Biotechnology progress* **23**, 1270-1276.

79. Zhang, Y. H. P. & Lynd, L. R. (2004). Toward an aggregated understanding of enzymatic hydrolysis of cellulose: noncomplexed cellulase systems. *Biotechnology and bioengineering* **88**, 797-824.

80. Wood, T. M. & McCrae, S. I. (1978). The cellulase of Trichoderma koningii. Purification and properties of some endoglucanase components with special reference to their action on cellulose when acting alone and in synergism with the cellobiohydrolase. *Biochem J* **171**, 61-72.

81. Baker, J. O., Ehrman, C. I., Adney, W. S., Thomas, S. R. & Himmel, M. E. (1998). Hydrolysis of cellulose using ternary mixtures of purified celluloses. In *Biotechnology for Fuels and Chemicals*, pp. 395-403. Springer.

82. Meyer, A. S., Rosgaard, L. & Sørensen, H. R. (2009). The minimal enzyme cocktail concept for biomass processing. *Journal of Cereal Science* **50**, 337-344.

83. Van Dyk, J. & Pletschke, B. (2012). A review of lignocellulose bioconversion using enzymatic hydrolysis and synergistic cooperation between enzymes—Factors affecting enzymes, conversion and synergy. *Biotechnology advances* **30**, 1458-1480.

84. Woodward, J., Hayes, M. K. & Lee, N. E. (1988). Hydrolysis of Cellulose by Saturating and Non–Saturating Concentrations of Cellulase: Implications for Synergism. *Nature Biotechnology* **6**, 301-304.

85. Hoshino, E., Shiroishi, M., Amano, Y., Nomura, M. & Kanda, T. (1997). Synergistic actions of exo-type cellulases in the hydrolysis of cellulose with different crystallinities. *Journal of Fermentation and Bioengineering* **84**, 300-306.

86. Henrissat, B., Vigny, B., Buleon, A. & Perez, S. (1988). Possible adsorption sites of cellulases on crystalline cellulose. *FEBS letters* **231**, 177-182.

87. Srisodsuk, M., Kleman-Leyer, K., Keränen, S., Kirk, T. K. & Teeri, T. T. (1998). Modes of action on cotton and bacterial cellulose of a homologous endoglucanase-exoglucanase pair from Trichoderma reesei. *European Journal of Biochemistry* **251**, 885-892.

88. Boisset, C., Fraschini, C., Schülein, M., Henrissat, B. & Chanzy, H. (2000). Imaging the enzymatic digestion of bacterial cellulose ribbons reveals the endo character of the cellobiohydrolase Cel6A from Humicola insolens and its mode of synergy with cellobiohydrolase Cel7A. *Applied and environmental microbiology* **66**, 1444-1452.

89. Väljamäe, P., Pettersson, G. & Johansson, G. (2001). Mechanism of substrate inhibition in cellulose synergistic degradation. *European Journal of Biochemistry* **268**, 4520-4526.

90. Medve, J., Ståhlberg, J. & Tjerneld, F. (1994). Adsorption and synergism of cellobiohydrolase I and II of Trichoderma reesei during hydrolysis of microcrystalline cellulose. *Biotechnology and bioengineering* **44**, 1064-1073.

91. Banerjee, G., Car, S., Scott-Craig, J. S., Borrusch, M. S., Aslam, N. & Walton, J. D. (2010). Synthetic enzyme mixtures for biomass deconstruction: production and optimization of a core set. *Biotechnology and bioengineering* **106**, 707-720.

92. Rivera, E. C., Rabelo, S. C., dos Reis Garcia, D. & da Costa, A. C. (2010). Enzymatic hydrolysis of sugarcane bagasse for bioethanol production: determining optimal enzyme loading using neural networks. *Journal of Chemical Technology and Biotechnology* **85**, 983-992.

93. Hsu, T.-C., Guo, G.-L., Chen, W.-H. & Hwang, W.-S. (2010). Effect of dilute acid pretreatment of rice straw on structural properties and enzymatic hydrolysis. *Bioresource technology* **101**, 4907-4913.

94. Coelho, P. S., Wang, Z. J., Ener, M. E., Baril, S. A., Kannan, A., Arnold, F. H. & Brustad, E. M. (2013). A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins in vivo. *Nat Chem Biol* **9**, 485-7.

95. Savile, C. K., Janey, J. M., Mundorff, E. C., Moore, J. C., Tam, S., Jarvis, W. R., Colbeck, J. C., Krebber, A., Fleitz, F. J. & Brands, J. (2010). Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science* **329**, 305-309.

96. Berman-Frank, I., Lundgren, P. & Falkowski, P. (2003). Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Research in Microbiology* **154**, 157-164.

97. Brown, O. R., Smyk-Randall, E., Draczynska-Lusiak, B. & Fee, J. A. (1995). Dihydroxy-acid dehydratase, a [4Fe-4S] cluster-containing enzyme in Escherichia coli: effects of intracellular superoxide dismutase on its inactivation by oxidant stress. *Arch Biochem Biophys* **319**, 10-22.

98. Brinkmann-Chen, S., Flock, T., Cahn, J. K., Snow, C. D., Brustad, E. M., McIntosh, J. A., Meinhold, P., Zhang, L. & Arnold, F. H. (2013). General approach to reversing ketol-acid reductoisomerase cofactor dependence from NADPH to NADH. *Proc Natl Acad Sci U S A* **110**, 10946-51.

99. Bastian, S., Liu, X., Meyerowitz, J. T., Snow, C. D., Chen, M. M. & Arnold, F. H. (2011). Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli. *Metab Eng* **13**, 345-52.

100. Atsumi, S., Hanai, T. & Liao, J. C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**, 86-89.

101. Gerlt, J. A., Babbitt, P. C. & Rayment, I. (2005). Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity. *Archives of biochemistry and biophysics* **433**, 59-70.

102. Rakus, J. F., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Hubbard, B. K., Delli, J. D., Babbitt, P. C., Almo, S. C. & Gerlt, J. A. (2008). Evolution of Enzymatic Activities in the Enolase Superfamily: L-Rhamnonate Dehydratase. *Biochemistry* **47**, 9944-9954.

103. Bloom, J. D. & Arnold, F. H. (2009). In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci U S A* **106 Suppl 1**, 9995-10000.

104. Gulick, A. M., Hubbard, B. K., Gerlt, J. A. & Rayment, I. (2001). Evolution of enzymatic activities in the enolase superfamily: identification of the general acid catalyst in the active site of D-glucarate dehydratase from Escherichia coli. *Biochemistry* **40**, 10054-62.

105. Rakus, J. F., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Vick, J. E., Babbitt, P. C., Almo, S. C. & Gerlt, J. A. (2007). Evolution of enzymatic activities in the enolase superfamily: D-Mannonate dehydratase from Novosphingobium aromaticivorans. *Biochemistry* **46**, 12896-908.

106. Fasan, R., Meharenna, Y. T., Snow, C. D., Poulos, T. L. & Arnold, F. H. (2008). Evolutionary history of a specialized P450 propane monooxygenase. *Journal of molecular biology* **383**, 1069-1080.

107. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869-74.

108. Nelson, D. L., Lehninger, A. L. & Cox, M. M. (2008). *Lehninger principles of biochemistry*, Macmillan.

109. Richter, F., Leaver-Fay, A., Khare, S. D., Bjelic, S. & Baker, D. (2011). De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230.

110. Dellus-Gur, E., Toth-Petroczy, A., Elias, M. & Tawfik, D. S. (2013). What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J Mol Biol* **425**, 2609-21.

111. Schmidt, D. M., Mundorff, E. C., Dojka, M., Bermudez, E., Ness, J. E., Govindarajan, S., Babbitt, P. C., Minshull, J. & Gerlt, J. A. (2003). Evolutionary potential of (β/α) 8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42**, 8387-8393.

112. Szekely, P., Sheftel, H., Mayo, A. & Alon, U. (2013). Evolutionary tradeoffs between economy and effectiveness in biological homeostasis systems. *PLoS Comput Biol* **9**, e1003163.

113. Savir, Y., Noor, E., Milo, R. & Tlusty, T. (2010). Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc Natl Acad Sci U S A* **107**, 3475-80.

114. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* **2**.

115. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982). *Molecular cloning: a laboratory manual*, 545, Cold Spring Harbor Laboratory Cold Spring Harbor, NY.

116. Yew, W. S., Fedorov, A. A., Fedorov, E. V., Rakus, J. F., Pierce, R. W., Almo, S. C. & Gerlt, J. A. (2006). Evolution of enzymatic activities in the enolase superfamily: L-fuconate dehydratase from Xanthomonas campestris. *Biochemistry* **45**, 14582-14597.

117. DeLano, W. L. (2002). The PyMOL molecular graphics system.

118. Cirino, P. C., Mayer, K. M. & Umeno, D. (2003). Generating mutant libraries using error-prone PCR. In *Directed Evolution Library Creation*, pp. 3-9. Springer.

119. Lewis, J. C., Mantovani, S. M., Fu, Y., Snow, C. D., Komor, R. S., Wong, C. H. & Arnold, F. H. (2010). Combinatorial alanine substitution enables rapid optimization of cytochrome P450BM3 for selective hydroxylation of large substrates. *ChemBioChem* **11**, 2502-2505.

120. Fellouse, F. A., Barthelemy, P. A., Kelley, R. F. & Sidhu, S. S. (2006). Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *Journal of molecular biology* **357**, 100-114.

121. Sims, R. E., Mabee, W., Saddler, J. N. & Taylor, M. (2010). An overview of second generation biofuel technologies. *Bioresource Technology* **101**, 1570-1580.

122. Blanch, H. W., Simmons, B. A. & Klein-Marcuschamer, D. (2011). Biomass deconstruction to sugars. *Biotechnology Journal* **6**, 1086-1102.

123. Klinke, H. B., Thomsen, A. & Ahring, B. K. (2004). Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. *Applied Microbiology and Biotechnology* **66**, 10-26.

124. Mikulášová, M., Vodný, Š. & Pekarovičová, A. (1990). Influence of phenolics on biomass production by Candida utilis and Candida albicans. *Biomass* **23**, 149-154.

125. Modig, T., Lidén, G. & Taherzadeh, M. (2002). Inhibition effects of furfural on alcohol dehydrogenase, aldehyde dehydrogenase and pyruvate dehydrogenase. *Biochem. J* **363**, 769-776.

126. Wang, X., Miller, E., Yomano, L., Zhang, X., Shanmugam, K. & Ingram, L. (2011). Increased furfural tolerance due to overexpression of NADH-dependent oxidoreductase FucO in Escherichia coli strains engineered for the production of ethanol and lactate. *Applied and environmental microbiology* **77**, 5132-5140.

127. Liu, Z. L. (2011). Molecular mechanisms of yeast tolerance and in situ detoxification of lignocellulose hydrolysates. *Applied microbiology and biotechnology* **90**, 809-825.

128. Dunlop, M. J., Dossani, Z. Y., Szmidt, H. L., Chu, H. C., Lee, T. S., Keasling, J. D., Hadi, M. Z. & Mukhopadhyay, A. (2011). Engineering microbial biofuel tolerance and export using efflux pumps. *Molecular systems biology* **7**.

129. Peng, X., Shindo, K., Kanoh, K., Inomata, Y., Choi, S.-K. & Misawa, N. (2005). Characterization of Sphingomonas aldehyde dehydrogenase catalyzing the conversion of various aromatic aldehydes to their carboxylic acids. *Applied microbiology and biotechnology* **69**, 141-150.

130. Almeida, J., Modig, T., Röder, A., Lidén, G. & Gorwa-Grauslund, M.-F. (2008). Pichia stipitis xylose reductase helps detoxifying lignocellulosic hydrolysate by reducing 5-hydroxymethyl-furfural (HMF). *Biotechnol Biofuels* **1**, 12.

131. Larsson, S., Cassland, P. & Jönsson, L. J. (2001). Development of a Saccharomyces cerevisiae strain with enhanced resistance to phenolic fermentation inhibitors in lignocellulose hydrolysates by heterologous expression of laccase. *Applied and environmental microbiology* **67**, 1163-1170.

132. (2013). ENZYME entry: EC 1.1.1.195.

133. Valencia, E., Larroy, C., Ochoa, W. F., Parés, X., Fita, I. & Biosca, J. A. (2004). Apo and Holo Structures of an NADP (H)-dependent Cinnamyl Alcohol Dehydrogenase from Saccharomyces cerevisiae. *Journal of molecular biology* **341**, 1049-1062.

134. Larroy, C., Fernandez, M., González, E., Parés, X. & Biosca, J. (2002). Characterization of the Saccharomyces cerevisiae YMR318C (ADH6) gene product as a broad specificity NADPH-dependent alcohol dehydrogenase: relevance in aldehyde reduction. *Biochem. J* **361**, 163-172.

135. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D. & Bussey, H. (1999). Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-906.

136. Tawfik, O. K. & S, D. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual review of biochemistry* **79**, 471-505.

137. Martín, C. & Jönsson, L. J. (2003). Comparison of the resistance of industrial and laboratory strains of Saccharomyces and Zygosaccharomyces to lignocellulose-derived fermentation inhibitors. *Enzyme and Microbial Technology* **32**, 386-395.

138. Van Dijken, J., Bauer, J., Brambilla, L., Duboc, P., Francois, J., Gancedo, C., Giuseppin, M., Heijnen, J., Hoare, M. & Lange, H. (2000). An interlaboratory comparison of physiological and genetic properties of four Saccharomyces cerevisiae strains. *Enzyme and microbial technology* **26**, 706-714.

139. Laadan, B., Almeida, J. R., Rådström, P., Hahn-Hägerdal, B. & Gorwa-Grauslund, M. (2008). Identification of an NADH-dependent 5-hydroxymethylfurfural-reducing alcohol dehydrogenase in Saccharomyces cerevisiae. *Yeast* **25**, 191-198.

140. Liu, Z. L. & Moon, J. (2009). A novel NADPH-dependent aldehyde reductase gene from Saccharomyces cerevisiae NRRL Y-12632 involved in the detoxification of aldehyde inhibitors derived from lignocellulosic biomass conversion. *Gene* **446**, 1-10.

141. Moon, J. & Liu, Z. L. (2012). Engineered NADH-dependent GRE2 from Saccharomyces cerevisiae by directed enzyme evolution enhances HMF reduction using additional cofactor NADPH. *Enzyme and microbial technology* **50**, 115-120.

142. Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S. M., Roodveldt, C. & Tawfik, D. S. (2004). The'evolvability'of promiscuous protein functions. *Nature genetics* **37**, 73-76.

143. Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S. & Milo, R. (2011). The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402-4410.

144. Gietz, R. D. & Schiestl, R. H. (2007). Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**, 38-41.

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

146. Wang, L., Jackson, W. C., Steinbach, P. A. & Tsien, R. Y. (2004). Evolution of new nonantibody proteins via iterative somatic hypermutation. *Proc Natl Acad Sci U S A* **101**, 16745-9.

147. Camps, M., Naukkarinen, J., Johnson, B. P. & Loeb, L. A. (2003). Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proceedings of the National Academy of Sciences* **100**, 9727-9732.

148. Makeyev, E. V. & Bamford, D. H. (2004). Evolutionary potential of an RNA virus. *Journal of virology* **78**, 2114-2120.

149. Davis, J. N. & van den Pol, A. N. (2010). Viral mutagenesis as a means for generating novel proteins. *Journal of virology* **84**, 1625-1630.

150. Das, A. T., Zhou, X., Vink, M., Klaver, B., Verhoef, K., Marzio, G. & Berkhout, B. (2004). Viral evolution as a tool to improve the tetracycline-regulated gene expression system. *Journal of Biological Chemistry* **279**, 18776-18782.

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

152. Papavasiliou, F. N. & Schatz, D. G. (2002). Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell* **109**, S35-S44.

153. Neuberger, M. S., Harris, R. S., Di Noia, J. & Petersen-Mahrt, S. K. (2003). Immunity through DNA deamination. *Trends in biochemical sciences* **28**, 305-312.

154. Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y. & Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553-563.

155. Di Noia, J. M. & Neuberger, M. S. (2007). Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1-22.

156. Peled, J. U., Kuang, F. L., Iglesias-Ussel, M. D., Roa, S., Kalis, S. L., Goodman, M. F. & Scharff, M. D. (2008). The biochemistry of somatic hypermutation. *Annu. Rev. Immunol.* **26**, 481-511.

157. Martin, A. & Scharff, M. D. (2002). Somatic hypermutation of the AID transgene in B and non-B cells. *Proceedings of the National Academy of Sciences* **99**, 12304-12308.

158. Odegard, V. H., Kim, S. T., Anderson, S. M., Shlomchik, M. J. & Schatz, D. G. (2005). Histone modifications associated with somatic hypermutation. *Immunity* **23**, 101-110.

159. Basu, U., Chaudhuri, J., Alpert, C., Dutt, S., Ranganath, S., Li, G., Schrum, J. P., Manis, J. P. & Alt, F. W. (2005). The AID antibody diversification enzyme is regulated by protein kinase A phosphorylation. *Nature* **438**, 508-511.

160. Nambu, Y., Sugai, M., Gonda, H., Lee, C.-G., Katakai, T., Agata, Y., Yokota, Y. & Shimizu, A. (2003). Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science* **302**, 2137-2140.

161. Petersen-Mahrt, S. K., Harris, R. S. & Neuberger, M. S. (2002). AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99-104.

162. Ramiro, A. R., Stavropoulos, P., Jankovic, M. & Nussenzweig, M. C. (2003). Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. *Nat Immunol* **4**, 452-6.

163. Kunkel, T. A. (1985). Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences* **82**, 488-492.

164. Coker, H. A., Morgan, H. D. & Petersen-Mahrt, S. K. (2006). Genetic and in vitro assays of DNA deamination. *Methods Enzymol* **408**, 156-70.

165. Bai, L., Santangelo, T. J. & Wang, M. D. (2006). Single-molecule analysis of RNA polymerase transcription. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 343-360.

166. Lee, J., Natarajan, M., Nashine, V. C., Socolich, M., Vo, T., Russ, W. P., Benkovic, S. J. & Ranganathan, R. (2008). Surface sites for engineering allosteric control in proteins. *Science* **322**, 438-442.

167. Faili, A., Aoufouchi, S., Guéranger, Q., Zober, C., Léon, A., Bertocci, B., Weill, J.-C. & Reynaud, C.-A. (2002). AID-dependent somatic hypermutation occurs as a DNA single-strand event in the BL2 cell line. *Nature immunology* **3**, 815-821.

168. Poltoratsky, V., Heacock, M., Kissling, G. E., Prasad, R. & Wilson, S. H. (2010). Mutagenesis dependent upon the combination of activation-induced deaminase expression and a double-strand break. *Molecular immunology* **48**, 164-170.

169. Inoue, H., Nojima, H. & Okayama, H. (1990). High efficiency transformation of Escherichia coli with plasmids. *Gene* **96**, 23-28.