Cel7A Engineering and Expression

By

Craig Matthew Dana

A dissertation submitted in partial satisfaction of

the requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Douglas S. Clark, Co-Chair
Professor Harvey W. Blanch, Co-Chair
Professor Jamie Doudna Cate, Outside

Fall 2013

Abstract

Cel7A Engineering and Expression

by

Craig Matthew Dana

Doctor of Philosophy in Chemical Engineering

University of California, Berkeley

Professor Douglas S. Clark, Co-Chair

Professor Harvey W. Blanch, Co-Chair

Renewable fuels produced from biomass-derived sugars are receiving increasing attention. Lignocellulose-degrading enzymes derived from fungi are attractive for saccharification of biomass because they can be produced at higher titers and at significantly less cost than those produced by bacteria or archaea. However, their properties can be suboptimal; for example, they are subject to product inhibition and are sensitive to small changes in pH. Furthermore, increased thermostability would be advantageous for saccharification as increased temperature may reduce the risk of microbial contamination. Therefore, there is a need for a generalized platform that can be applied to the engineering of these enzymes. Commercially available lignocellulose-degrading enzymes are produced using a hypersecreting strain of the filamentous fungus *Trichoderma reesei*. Among the enzymes secreted by this organism, the cellulase Cel7A is present in the highest concentration and is the only enzyme responsible for non-reducing end directed exo-acting cellulolytic activity. Additionally, the enzyme's presence is critical for growth of *T. reesei* on cellulosic substrates. Here, a general mutagenesis platform that employed the budding yeast *Saccharomyces cerevisiae* was developed to improve the properties of Cel7A. Secretion of Cel7A at titers of 26 mg/L with limited hyperglycosylation was achieved using an *S. cerevisiae* strain with upregulated protein disulfide isomerase, an engineered α-factor prepro leader, and the deletion of a plasma membrane ATPase. Because cellulase activities are difficult to screen in high-throughput, a DNA shuffling based library generation technique that results in a high percentage of active clones was developed called Biased Clique Shuffling (BCS). BCS allows for the control of DNA diversity during library generation. Applying this technique to 11 homologous Cel7A genes, we generated several libraries that were rich in activity and identified chimeras with improved thermostability, thermal activity, and product inhibition. The libraries generated using the BCS technique were far superior as a source of active and stable chimeras compared to an equimolar library prepared from the same 11 genes (as is classically prepared using DNA shuffling). Finally, we found that Cel7A expressed in the filamentous fungus *N. crassa* had twice the specific activity at 65°C and a 10°C higher T_m relative to Cel7A expressed in *S. cerevisiae*. Through a study of the three known post-translational modifications, namely glycosylation, disulfide bond formation, and N-terminal glutamine cyclization, we revealed that *S. cerevisiae* expressed Cel7A with an unmodified N-terminus, unlike native Cel7A which has an N-terminal pyroglutamate. Furthermore cyclizing the unmodified N-terminal glutamine in the Cel7A expressed in *S. cerevisiae* to form pryoglutamate *in-vitro* with glutaminyl cyclase increased the enzyme's specific activity and thermostability to match those Cel7A expressed in *N. crassa*. This unprecedented result demonstrates the importance of the hydrophobic pyroglutamate in the N-terminal position of Cel7A.

To

Mom and Dad

for your unconditional love and support.

Acknowledgments

Without the support, advice, and direction from Professors Douglas S. Clark and Harvey W. Blanch, this thesis would stand forever unwritten. I could not have asked for better mentors in this adventure. Thank you.

While there is formally a single author here, this work would be without any reasonable merit were it not for the contributions and camaraderie from countless scientists and engineers. In particular, to Poonam Saija, Sarala Kal, and Harshal Chokhawala, a whole-hearted thank you.

To Dr. Brad Sickenius, whose wisdom helped guide me through the most difficult of times, know that I'll always appreciate the value of a good pair of shoes.

To Professor Joel Sokol, your calm kindness and willingness to help inspire me. I've looked up to you all my life and will continue to do so for its remainder. Thank you for being so valuable an ally.

To my brothers Steve and Dave, who throughout my life have been role models of the highest caliber, I simply could not have made it this far without you. Thanks to you, I got it done.

Finally, to Rachel, your love and support have strengthened me immeasurably. Every baon was packed with love. Thank you so much.

Table of Contents

List of Tables..iv

List of Figures...v

1 Development of a *Saccharomyces cerevisiae* Expression Platform for Cel7A..1

1.2 Materials and Methods...7

1.3 Figures...9

2 DNA Shuffling Library Design and DNA Synthesis Methods for the Engineering of Cel7A...14

2.2 Materials and Methods...22

2.3 Tables..24

2.4 Figures...26

3 Biased Clique Shuffling for the Evolution of Cel7A..32

3.2 Materials and Methods...43

3.3 Tables..46

3.4 Figures...49

4 The Importance of Pyroglutamate in Cel7A...64

4.2 Materials and Methods...69

4.3 Tables..71

4.4 Figures...72

5 References...76

List of Tables

2-1 Source of Cel7A genes……………………………..…………………………………………….……24

2-2 Results of Clique Analysis……………………………..……………………………………………….25

2-3 UNIPROT accession numbers for the eleven Cel7A parental genes………………25

3-1 Mutations present in active and stable chimeras..……………………………………….…46

3-2 Mutations present in active but unstable chimeras...…………………………………..…47

3-3 Mutations present in inactive chimeras……………………………………………………..48

4-1 Melting temperatures (T_m) of TeCel7A preparations…………………………………71

List of Figures

1-1 Activity in yeast supernatant..........9

1-2 MuLac assay of the soluble portion of lysate from wild-type *S. cerevisiae* expressing different Cel7A genes under the control of GPD or CUP1 promoters..10

1-3 Anti-His_6 Western Blots of secreted TeCel7A from different strains of *S. cerevisiae*..........11

1-4 SDS PAGE gel of TeCel7A expressed in *S. cerevisiae* ΔOCH1 at 25°C..........12

1-5 MuLac activity of TeCel7A expressed in *S. cerevisiae* ΔOCH1 in high-throughput configurations..........12

2-1 Schematic representation of the DNA shuffling method..........26

2-2 Agarose gels depicting the construction of a DNA shuffled Cel7A library..........27

2-3 Representative sequence from the Equimolar Library..........28

2-4 Representative sequence from the Biased Library..........28

2-5 Agarose gels depicting the AMP method for DNA synthesis..........29

2-6 A cartoon depiction of the AMP assembly..........29

3-1 Biased Clique Shuffling Library Design Schematics..........49

3-2 High-throughput screen of 50% TeCel7A biased library in 3mM cellobiose..........50

3-3 Comparison of inhibited and uninhibited initial rates on MuLac..........51

3-4 Inhibition curve of J14 compared to that of TeCel7A and *T. longibrachiatum* Cel7A..........52

3-5 Example high-throughput data for the second generation screen..........53

3-6 Calculated rates from high-throughput data versus cellobiose concentration......54

List of Figures

3-7 Assay of 5A10 and TeCel7A supernatant without robotics....................................55

3-8 Schematic of Molecular Biology Steps in Biased Clique Shuffling..........................56

3-9 A subset of high-throughput data from the screen at 65°C..................................57

3-10 Temperature-activity profile of TeCel7a and mutants on MuLac...........................58

3-11 Residual activity after heating at 65°C...59

3-12 Temperature-activity profile of TeCel7A and mutants on Avicel............................60

3-13 Structure of TeCel7A...61

4-1 Activity of TeCel7A enzymes against Avicel...72

4-2 Crystal structure of TeCel7A catalytic domain...73

4-3 N-terminal glutamine cyclization...73

4-4 Differential scanning calorimetry data measuring T_m values of TeCel7A..............74

Chapter 1

Development of a *Saccharomyces cerevisiae* Expression Platform for Cel7A

The properties of heterologously-expressed Cel7A in *Saccharomyces cerevisiae* are unpredictable. For example, Smith and coworkers report an optimum temperature (T_{opt}) of 40°C for *Trichoderma reesei* Cel7A heterologously expressed in *S. cerevisiae*, while it is well known that the T_{opt} of the native enzyme is over 50°C (Baker et al., 1992; Smith et al., 2013). Reports of T_{opt} for recombinant Cel7A deviate widely even when deliberate attempts are made to use identical expression conditions (Heinzelman et al., 2010). Voutilainen and colleagues reported that Cel7A from *Talaromyces emersonii* (TeCel7A) expressed in *S. cerevisiae* maintains 100% of its optimal activity at 65°C while other reports indicate that optimal activity is reduced by more than half at 65°C (Dana et al., 2012; Heinzelman et al., 2010; Komor et al., 2012; Voutilainen et al., 2010).

An additional complication is that expression levels of Cel7A in *S. cerevisiae* are generally quite low and depend on the primary sequence of the particular Cel7A. Moreover, Cel7A is often hyperglycosylated, running as a high molecular weight smear upwards of 200 kDa on an SDS PAGE gel (the native enzyme runs at about 66kDa) (Ilmén et al., 2011; Penttilä et al., 1988). While particular blocks of primary sequence have been implicated (Heinzelman et al., 2010), and the unfolded protein response has been demonstrated to play a role (Ilmén et al., 2011), it remains unclear why there is such diversity in expression level and extent of hyperglycosylation among Cel7A sequences.

In this work, Cel7A expression in *S. cerevisiae* was achieved through the use of a multifaceted expression system involving optimized promoter choice, plasmid type, and strain, among other elements. This chapter discusses the basic platform of Cel7A expression in *S. cerevisiae*. However, the effect of the absence of the post-translational modification that catalyzes the cyclization of the N-terminal glutamine is left for discussion in Chapter 4.

Achieving Expression of Cel7A in *Saccharomyces cerevisiae*

The budding yeast *S. cerevisiae* was chosen as the expression host for Cel7A study and engineering. This organism was chosen because it has several advantages:

1. It can secrete enzymes.
2. It is easily transformed at high efficiency.
3. Its genome is easily manipulated.
4. It can perform both O-linked and N-linked glycosylation.
5. It can form disulfide bonds.

These advantages indicate that *S. cerevisiae* is suitable for both characterization and high-throughput engineering of Cel7A. However, as discussed in detail in this chapter, native *S. cerevisiae* is unfit for Cel7A expression; therefore, I applied several strategies to acheive expression and secretion of Cel7A.

Expression of Cel7A under the control of the constitutive GPD promoter was compared with that under the control of the copper inducible CUP1 promoter. Cel7A genes from *Talaromyces emersonii, Acremonium thermophilum,* and *Thermoascus aurantiacus,* were tested for expression. As shown in Figure 1-1, no activity on the soluble substrate β-1,4-Methylumbelliferyl-lactoside was present in the supernatant for any promoter/gene combination tested. Next, I assayed the soluble lysate obtained through several passes of the harvested cells through a french press. In the soluble lysate fraction, activity was achieved with the Cel7A from *T. emersonii* (TeCel7A) expressed under control of either promoter. However, activity was approximately ten-fold higher for TeCel7A expressed under the control of the CUP1 promoter as shown in Figure 1-2. The CUP1 promoter was selected for use in future expression studies of Cel7A.

It is interesting that the other Cel7A sequences do not express in active form under either promoter while the TeCel7A does. The expressability of TeCel7A in *S. cerevisiae* relative to other Cel7A sequences has been documented by other researchers as well. Ilmen and coworkers reported similar findings when expressing fourteen different Cel7A sequences in *S. cerevisiae*. TeCel7A (Ilmén et al., 2011). They show that TeCel7A expresses at the highest level among the sequences and that nine out of fourteen of the other Cel7A sequences express at least 100-fold less than TeCel7A. Notably, one of these nine is the Cel7A from *Trichoderma reesei*, a commonly found enzyme in industrial cellulase cocktails. They also go on to show that the unfolded protein response is activated strongly during the expression of *T. reesei* Cel7A but only very mildly during expression of *T. emersonii.* This suggests that there are differences in the abilities to fold of the different Cel7A proteins in the ER lumen of *S. cerevisiae*. The cause for this variation across Cel7A genes is unknown, although Chapter 6 of this thesis provides a hypothesis.

Intracellular expression of Cel7A would hinder high-throughput screening efforts; therefore, the signal peptide was replaced with an engineered α-factor sequence called APPS4. This strategy worked well and led to the secretion of TeCel7A in low titers (<1 mg/mL).

S. cerevisiae is known to hyperglycosylate recombinant enzymes although it is unclear why some enzymes are decorated with more glycans than others. Hyperglycosylation begins when the yeast glycosyl transferase OCH1 covalently links a mannose residue to the core N-linked oligosaccharides. This decorated core then becomes the substrate for several glycosyl transferases which act repeatedly to add on additional mannose moieties. In this way, as many as 200 mannose residues can be linked to a glycoprotein at each asparagine site within the NXS/T sequon.

While presented earlier as an advantage of the *S. cerevisiae* expression system, N-linked glycosylation can hinder activity of Cel7A when it is present as hyperglycosylation. Native Cel7A proteins possess both N-linked and O-linked glycans, but they are not hyperglycosylated. In native Cel7A, the N-linked glycans are commonly observed in truncated form where only a single N-acetyl glucosamine residue is present.

Previously, researchers have studied the effect of removing genes in the glycosylation pathway. From these studies several candidate genes were selected to be knocked out in S*. cerevisiae* to limit the extent of hyperglycosylation of recombinant Cel7A.

Firstly, the KRE2 knockout was studied. KRE2 denotes killer toxin resistance. The phenotype of *S. cerevisiae* ΔKRE2 displays a marked reduction in average content of N-linked glycosylation, indicating a deficiency in the N-linked glycosylation pathway (Hill et al., 1992). It follows then, that Cel7A expressed in *S. cerevisiae* ΔKRE2 is likely to displayed reduced N-linked glycosylation on average. Previously, researchers have reported that TeCel7A expressed in *S. cerevisiae* ΔKRE2 was not hyperglycosylated (Heinzelman et al., 2010). Still, our findings are that a significant fration of the secreted TeCel7A expressed in *S. cerevisiae* ΔKRE2 was hyperglycosylated. The polyacrylamide gel electrophoresis experiment revealed an extremeley heterogeneous population of TeCel7A protein molecules which were smeared over a range of 100kDa. Unlike the wild-type *S. cerevisiae*, the *S. cerevisiae* ΔKRE2 strain was able to limit hyperglycosylation in a subset of secreted Cel7A proteins. However, because a large fraction remained hyperglycosylated, *S. cerevisiae* ΔKRE2 was not selected for use in Cel7A engineering.

Next, the double knockout of genes MNN1 and MNN9 was studied. MNN1 and MNN9 genes encode mannosyl transferases involved in the decoration of the core N-linked oligosaccharide unit. Mnn9p, a membrane-bound protein, is required for the addition of α1,6 mannoses to the backbone of complex manna. Mnn1p, an α1,3 mannosyltransferase, is located in the golgi apparatus *S. cerevisiae* ΔMNN1ΔMNN9. This double knockout was selected for study because it has been reported that the secreted glycoprotein invertase has homogeneous Man10GlcNAc2 oligosaccharides and does not contain α1,3 linked mannoses (Yip et al., 1994). Surprisingly, *S. cerevisiae* ΔMNN1ΔMNN9 produced TeCel7A with similar glycosylation patterning to the wild-type *S. cerevisiae*, as measured by gel electrophoresis and shown in Figure 1-3**.** Because *S. cerevisiae* ΔMNN1ΔMNN9 provided no reduction in N-linked glycosylation of TeCel7A, it was not chosen for the study or engineering of Cel7A.

Next, we explored the use of *S. cerevisiae* ΔOCH1. OCH1 denotes outer chain formation, and Och1p initiates the construction of the outer chain high-mannose structure. This backbone then acts a substrate for downstream mannosyl transferases that covalently link mannose residues in serial fashion in the development of a hyperglycosylated structure (Nakayama et al., 1992). *S. cerevisiae* ΔOCH1 secretes invertase with an outer chain deficiency; that is, the outer chain is not formed and thus the invertase presents as a single band on a polyacrylamide gel rather than a smear. We expressed TeCel7A in *S. cerevisiae* ΔOCH1 and observed that the secreted TeCel7A protein product was not hyperglycosylated. When secreted from *S. cerevisiae* ΔOCH1, TeCel7A did not run above 100 kDa on a polyacrylamide gel; whereas TeCel7A secreted from wild-type *S. cerevisiae* exclusively ran above 100 kDa. In addition, the TeCel7A secreted from *S. cerevisiae* ΔOCH1 presented primarily as a single band. Some proteolysis products were detected with the Western Blot, perhaps indicating that the hyperglycosylation of TeCel7A observed in the wild-type *S. cerevisiae* and in *S. cerevisiae* ΔMNN1ΔMNN9 prevented proteolytic degradation, albeit at the expense of producing hyperglycosylated TeCel7A. Despite the limited extent of proteolysis products, *S. cerevisiae* ΔOCH1 appeared to be an excellent candidate for Cel7A engineering and study.

Finally, we investigated *S. cerevisiae* ΔPMR1. PMR1 denotes plasma membrane ATPase related. PMR1 encodes an ATPase that is involved in calcium and manganese homeostasis within the endoplasmic reticulum and the golgi apparatus. While Pmr1p is not a glycosyl transferase, removing it from the genome of *S. cerevisiae* has an indirect, but pronounced effect on the extent of N-linked glycosylation present in its secreted proteins (Antebi and Fink, 1992). This implies a global role for calcium homeostasis in the normal processing of glycoproteins through the secretory pathway of *S. cerevisiae*. We used *S. cerevisiae* ΔPMR1 to express TeCel7A and found that secreted TeCel7A had limited glycosylation. Similar to *S. cerevisiae* ΔOCH1, *S. cerevisiae* ΔPMR1 secreted TeCel7A that ran below 100 kDa on a polyacrylamide gel. TeCel7A from *S. cerevisiae* ΔPMR1 was also proteolysed to some extent, as in the case for TeCel7A secreted from *S. cerevisiae* ΔOCH1. This supports the hypothesis that hyperglycosylation wards off proteolytic attack. Because it secreted TeCel7A as primarily a single band, *S. cerevisiae* ΔPMR1 was chosen for use in the engineering and study of TeCel7A in this thesis.

Both *S. cerevisiae* ΔPMR1 and *S. cerevisiae* ΔOCH1 secreted TeCel7A with limited glyocylsation patterning; however, an unintended and undesirable consequence of this was the vulnerability of TeCel7A to proteolytic attack. To address this, the temperature was reduced from 30°C to 25°C after induction of TeCel7A expression in *S. cerevisiae* ΔPMR1. As shown in Figure 1-4, TeCel7A expressed at the lower temperature was both intact and free of hyperglycosylation.

While *S. cerevisiae* ΔPMR1 did secrete TeCel7A appended to an α-factor leader sequence, the expression titer was less than 1 mg/L, a value too low for facile high-throughput screening. Thus, while the selection of the copper-inducible CUP1 promoter and the PMR1 knockout were helpful, the *S. cerevisiae* expression system required further engineering before it could have been used for high-throughput engineering of Cel7A.

It has previously been shown that the formation of disulfide bonds can be limited in the production of recombinant proteins in *S. cerevisiae*. In the laboratory of Dane Wittrup, researchers have compared the expression titers of monoclonal antibodies in two different strains of *S. cerevisiae*. One strain was the wild-type while the other strain contained additional chromosomally-integrated copies of the protein disulfide isomerase gene (PDI) under a constitutive promoter. They found that the expression titer of the disulfide-bonded antibody increased substantially in the PDI upregulated strain relative to the wild-type strain (Robinson et al., 1994).

TeCel7A contains eleven disulfide bonds, a count which includes the two disulfide bonds that are present in the appended carbohydrate binding module (CBM) from the fungus *Agaricus bisporus*; therefore, it is likely that disulfide bond formation was also limiting the expression of TeCel7A. We obtained a strain of *S. cerevisiae* with upregulated PDI activity as a gift from Dane Wittrup.

In other work carried out by these researchers, the α-factor leader sequence was appended to a reporter protein and engineered to increase secreted titer using random error-prone PCR mutagenesis. One improved mutant, named APPS4, was selected for use in this study. In TeCel7A, the α-factor signal sequence was replaced with the hypersecreting signal APPS4.

Using the combination of these strategies, the expression titer of secreted TeCel7A reached over 26 mg/L. This titer, achieved in 2.5L shake flasks, was sufficient for high-throughput screening; however, high-throughput screening requires miniaturization of cultures in multi-well plates, and it was unclear how this would effect expression titer. We tested several configurations of multi-well plates and found that the configurations that maximized the surface area exposed to air to culture volume ratio performed best. As depicted in Figure 1-5, the deep-well 96 well plates nearly failed to secrete any active TeCel7A. In this configuration, the yeast grows beneath 1.5mL of media affixed to the bottom of the plate while only a small area at the top of the column is exposed to air. In the standard 96 well plates and large volume 6-well multiwell plates, the secreted expression titers were far higher, as indicated by the increased activity in the supernatant. The standard 96 well plates were chosen for high-throughput expression because they allow for high secreted expression titers without restricting the number of wells.

In conclusion, the expression of Cel7A in *S. cerevisiae* was optimized in an iterative process that ultimately combined several strategies. The inducible promoter CUP1 was found to be far superior to the constitutive GPD promoter**.** Secretion was obtained through the use of the engineered α-factor APPS4. The hyperglycosylating phenotype of wild-type yeast was suppressed by knocking out the PMR1 gene that encodes an ATPase which controlled calcium homeostasis in the ER and Golgi. And finally, the expression titer was increased by using a strain of *S. cerevisiae* that contained additional chromosmally integrated copies of the PDI gene that encodes a protein disulfide isomerase. Ultimately, a secreted titer of 26mg/L of TeCel7A with limited glycosylation patterning was achieved. The titer remained high when expressed in a high-throughput format provided that the surface area exposed to air versus culture volume ratio remained high.

Materials and Methods

Strain Engineering and Expression of TeCel7A in *S. cerevisiae*

S. cerevisiae with the following knockouts were obtained from the ATCC, transformed with the *T. emersonii* Cel7A, and characterized in terms of their glycosylation patterning of recombinant *T. emersonii* Cel7A using a Western blot: KRE2 (ATCC: 4034317), MNN1MNN9 (ATCC: 200180), OCH1 (ATCC: 4034406), PMR1 (ATCC: 4014534). The PDI overexpressing strain YVH10 was obtained from Prof. Dane Wittrup (Robinson et al., 1994). The PMR1 locus was disrupted using standard methods (Gueldener et al., 2002). Expression Protocol Either SC-Trp or SC-Leu was inoculated with

S. cerevisiae containing the Cel7A gene in high-copy number plasmids (pCu424 or pCu425; (Labbé and Thiele, 1999), 1999) and grown for 3 days at 308C, 220 rpm. Cultures were spun down at 5,000g for 5 min and resuspended in YPD supplemented with 500 mM Cu2SO4 and allowed to express for 3 days at 258C, 220 rpm. Supernatant was collected and purified over a GE HisTrap HP 5mL column. The centromeric plasmid tested was YCpLG, but this provided poor titers. For expression under the control of the GPD promoter, cultures were grown in YPAD medium for three days prior to harvesting. Clique Identification Glycosyl Hydrolase 7 (GH7) sequences were downloaded from the PFAM database and clustalw version 2.0.9 was used to identify all pairs that shared amino acid percent identity within the range of 68–86%. The list was then compiled and used as input for the Cliquer algorithm (Östergård, 2002) to identify the maximum GH7 clique. Where needed, linkers and CBMs from related Cel7A enzymes were appended to the catalytic domains identified. Selected genes are shown in Table I with Uniprot accessions in Supplemental Figure 4. Gene Synthesis For all genes, the native signal sequence was identified using SignalP and replaced with the mutant a-factor leader AppS4 (Rakestraw et al., 2009). In addition, six histidine residues were appended to the C-termini of the genes using the following nucleotide sequence CACCATCACCATCACCAT. The *T. emersonii* Cel7A was ordered from Genscript using their proprietary *S. cerevisiae* codon bias. The remaining 10 genes were ordered from DNA2.0 in the following manner. First, amino acid sequences were aligned to the *T. emersonii* Cel7A sequence. Where amino acids were identical, identical codons were chosen. Where amino acids differed, DNA2.0 used their proprietary *S. cerevisiae* codon optimization algorithm to select the codon. Genes were sublconed into pCu424 or pCu425 using standard methods.

Activity Assays on Methylumbelliferyl Lactoside

Supernatant activities were assayed by first concentrating using Amicon 30kDa spin concentrators at 4000xg for 1 hour at 4°C, to yield a 100-fold concentration. 10µL of concentrated supernatant was then mixed with 837µM MuLac in 50mM sodium acetate pH 5 in 200uL and incubated at 50°C.

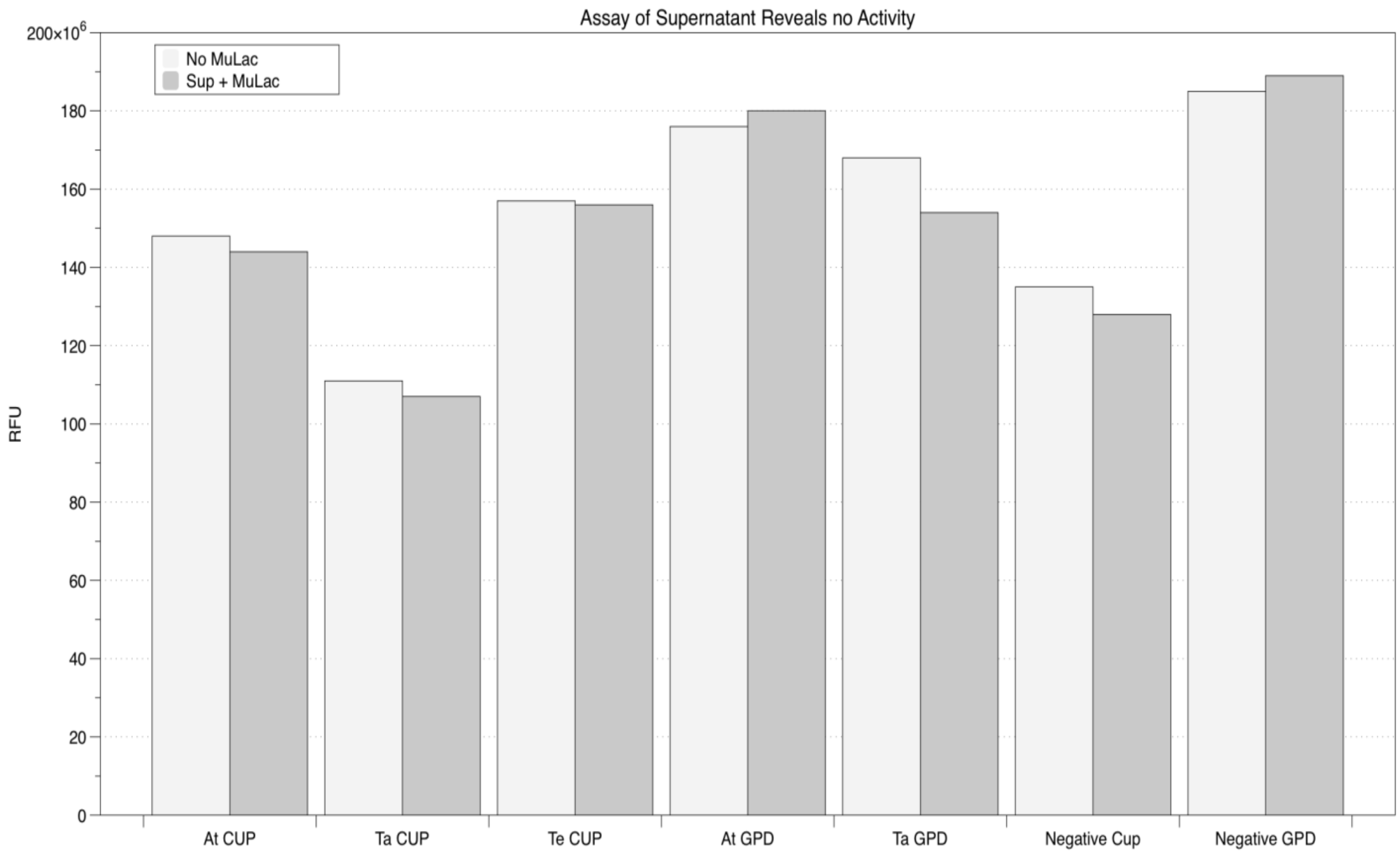

Assay of Supernatant Reveals no Activity
No MuLac
Sup + MuLac
RFU
200×10^{6}
180
160
140
120
100
80
60
40
20
0
At CUP
Ta CUP
Te CUP
At GPD
Ta GPD
Negative Cup
Negative GPD

Figure 1-1

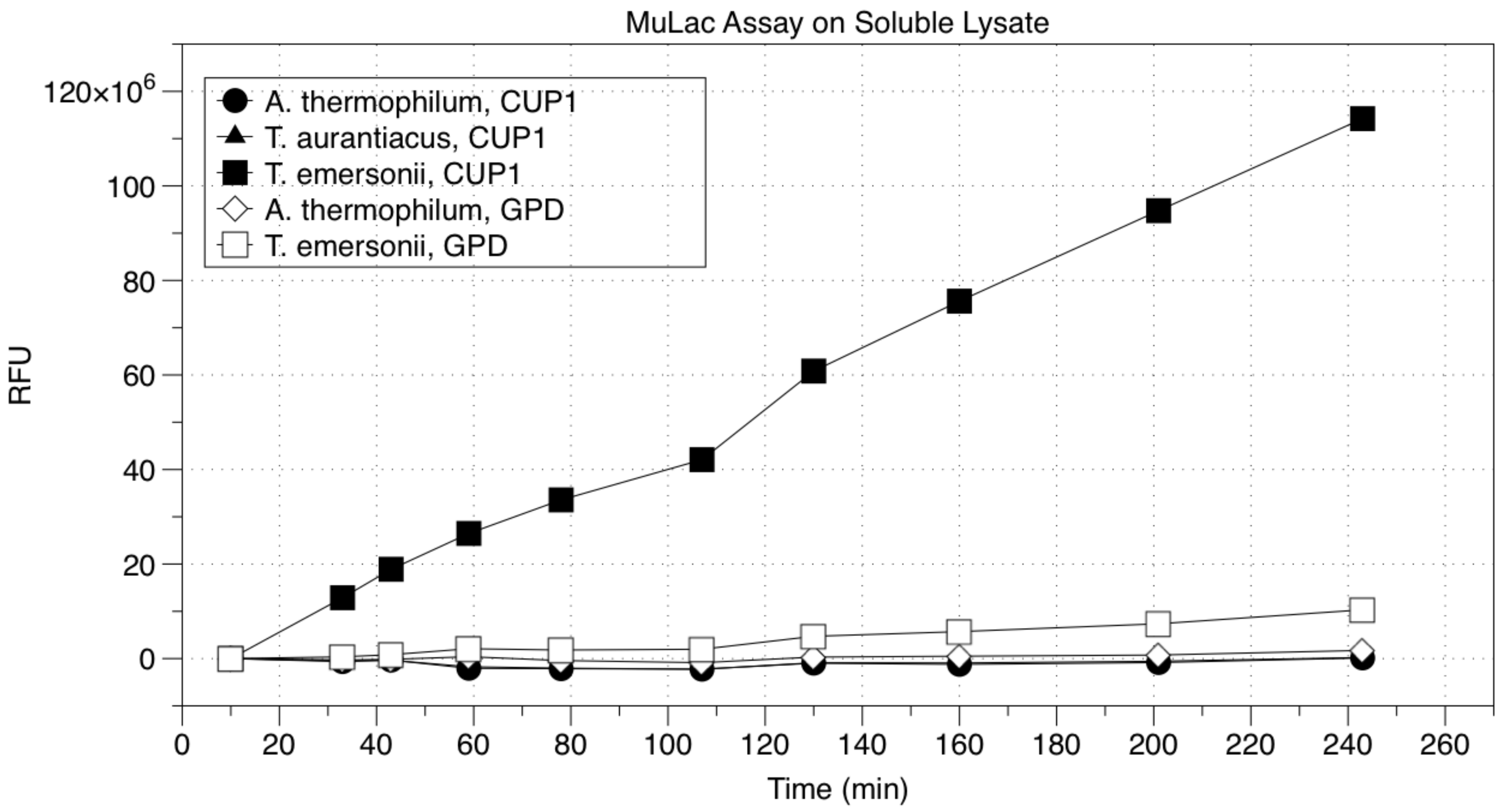

MuLac Assay on Soluble Lysate
RFU
120×10^6
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
200
220
240
260
Time (min)
● A. thermophilum, CUP1
▲ T. aurantiacus, CUP1
■ T. emersonii, CUP1
◇ A. thermophilum, GPD
□ T. emersonii, GPD

Figure 1-2

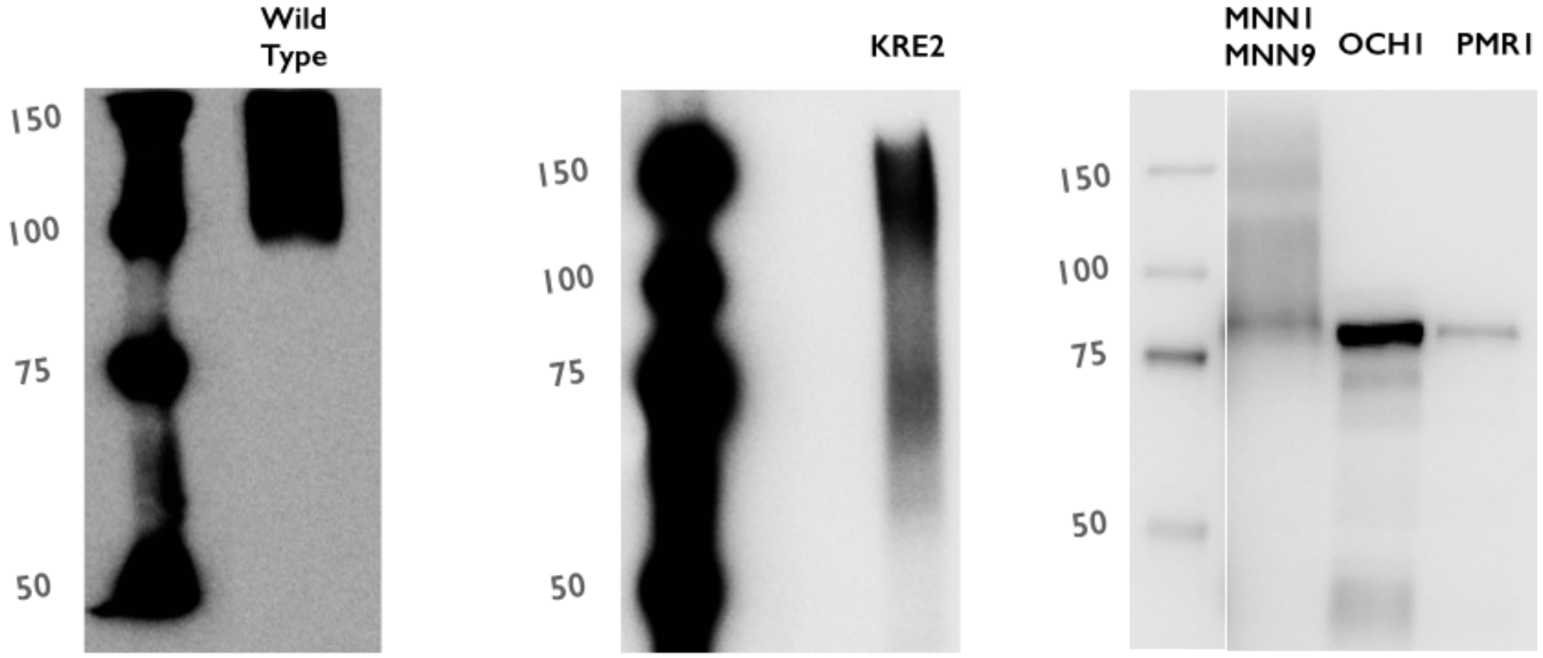

Wild
Type
150
100
75
50
KRE2
150
100
75
50
MNN1
MNN9
OCH1
PMR1
150
100
75
50

Figure 1-3

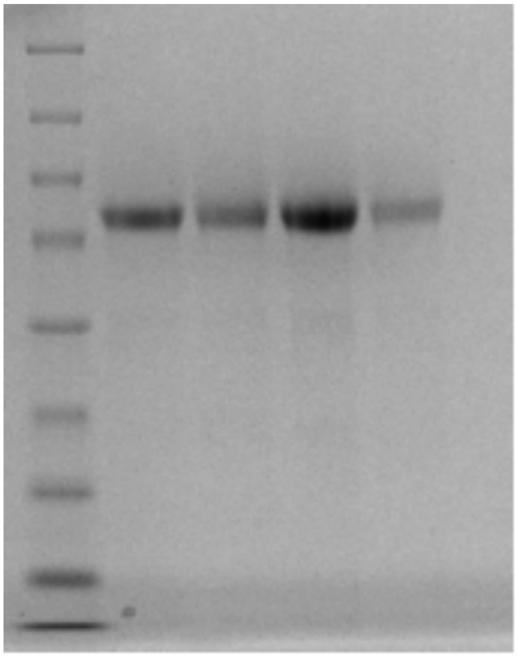


Figure 1-4

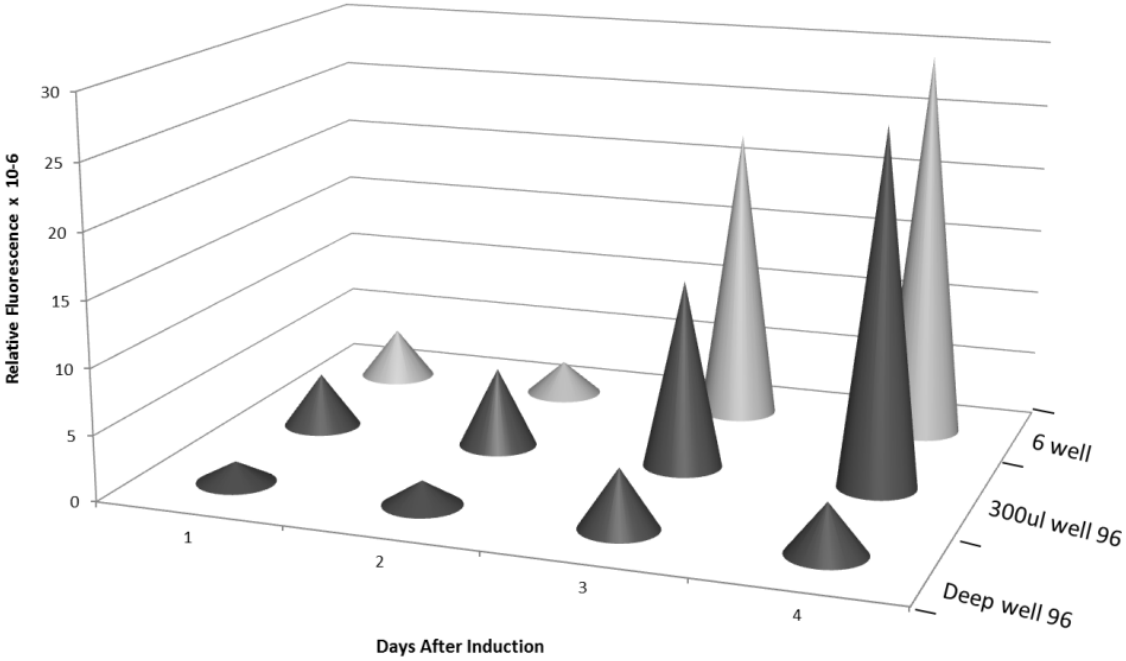


Figure 1-5

Figure 1-1. Activity in yeast supernatant. No activity was found in the yeast supernatant for any promoter/Cel7A combination tested. For all cases tested, the relative fluorescence in the presence of MuLac was about equal to the background fluorescence present in the rich media.

Figure 1-2. MuLac assay of the soluble portion of lysate from wild-type *S. cerevisiae* expressing different Cel7A genes under the control of GPD or CUP1 promoters. The TeCel7A expressed under the control of the CUP1 promoter far outperformed the other combinations. It is not clear why the Cel7A gene from *T. aurantiacus* or *A. thermophilum* do not express well.

Figure 1-3. Anti-His$_6$ Western Blots of secreted TeCel7A from different strains of *S. cerevisiae*. Expression of TeCel7A in wild-type *S. cerevisiae* resultd in a highly heterogeneous protein population, all of which is hyperglycosylated. Expression of TeCel7A in the glycosylation-defective KRE2 strain reduced the average level of glycosylation but did not sufficiently mitigate hyperglcyosylation, as shown in the middle image. Similarly MNN1MNN9 resulted in some TeCel7A with limited glycosylation, but did not eliminate high molecular weight TeCel7A. OCH1 and PMR1 produced TeCel7A without smeaing over 100 kDa, indicating that hyperglycosylation was substantially limited.

Figure 1-4. SDS PAGE gel of TeCel7A expressed in *S. cerevisiae* ΔOCH1 at 25°C. This expression protocol resulted in primarily a single TeCel7A band. Additional bands are replicate cultures. The proteolysis shown in Figure 1-2, OCH1 lane, was markedly reduced at 25°C relative to 30°C.

Figure 1-5. MuLac activity of TeCel7A expressed in *S. cerevisiae* ΔOCH1 in high-throughput configurations. Expression in deep well multiwell plates was poor, as indicated by the low level of activity in the supernatant. Expression in standard 300µL 96 well plates was far improved, as was expression in the larger 6 well plates. Because it is amenable to high-throughput screening, the standard 300µL plates were selected.

Chapter 2

DNA Shuffling Library Design and DNA Synthesis Methods for the Engineering of Cel7A

In order to facilitate industrial scale hydrolysis of cellulose for biofuel production, a multitude of cellulase attributes will require engineering including thermostability, thermal activity, pH-activity profile, product inhibition, specific activity, synergy, and expressibility in recombinant hosts, among others. Because of the breadth of attributes to be engineered into a single enzyme, a rational approach would be exceptionally resource and time intensive. Alternatively, a directed evolution approach allows each of these attributes to be addressed with a single and relatively simple methodology. The only variation in the methodology would be the development and use of appropriate screens that test for improvements in particular attributes. The success of the evolution experiment would also depend on the method to create diversity in the DNA encoding the cellulase enzyme. In this work, the cellulase Cel7A was chosen for engineering owing to its dominant presence in commercial cellulase cocktails; however, the tools presented in this thesis are general by design and applicable to the evolution of any fungal cellulase.

Many methods have been developed to generate diversity in the DNA of an enzyme in order to create a library of mutant enzymes that can be screened for improved fitness (Romero and Arnold, 2009). The goal of these methods is to identify a set of mutations that imparts an incremental improvement in fitness relative to the original enzyme. Two broad categories of library diversification methods can be defined. In one are methods that induce point mutations in DNA sequences. In the second are methods that result in the recombination of lengths of DNA from two or more parental sequences.

In the first category, the most popular method is called Error-Prone Polymerase Chain Reaction (epPCR). In epPCR, the DNA template is amplified using the polymerase chain reaction (PCR) under conditions that promote mismatching of DNA nucleotides. For example, biased proportions of nucleotides can be added to the PCR reaction or magnesium concentrations can be increased. Proprietary polymerases are also available that, themselves, have been evolved to introduce errors in DNA sequences (Cirino et al.). These specialized polymerases have been engineered to promote biased mutation rates (ie: more GC to AT changes). If the initial concentration of template is sufficiently low, the percentage of sequences containing point mutations will be high (Cirino, 2003). The products of this error-prone amplification are referred to as the library and contain a random distribution of mutations throughout the length of the DNA sequence. While the mutations occur randomly, they do not occur without bias (Neylon, 2004). In fact, this method strongly favors particular types of mutations owing to the simple fact that some pairs of codons encoding distinct amino acids are closer in sequence than other pairs. Because it is unlikely that two or three mutations will occur in the same codon, amino acid substitutions accessible via single mutations tend to dominate the resultant libraries. On the other hand, amino acid substitutions that require two or three mutations in a single codon are rare. This constraint cannot be addressed through polymerase engineering or nucleotide bias. Therefore, libraries constructed using epPCR are inherently biased.

Another challenge associated with epPCR and other point mutation techniques is the fact that 30-50% of mutations arrived at via epPCR tend to be strongly deleterious (Romero and Arnold, 2009). While not surprising, this leads to a significant fraction of inactive enzymes in the resultant library. The screening and/or selection protocol then becomes critical.

Some selection and screening protocols permit the exploration of thousands of billions sequences. One percent of this vast sequence space is still remarkably vast. Examples of these screening and selection strategies include phage display methods, ribosome display, and others. One example of particular interest here is a chemical complementation method for cellulase enzymes (Peralta-Yahya et al., 2008). In this method, researchers linked growth to the β1-4 glycosidic bond cleavage through the URA3 counter-selection system in the yeast *Saccharomyces cerevisiae*. When an active cellulase is expressed intracellularly, the bond within a cellobiose linker is cleaved, ultimately preventing the transcription of the toxic URA3 gene. Because this is a selection process, the number of DNA sequences that can be tested is only limited by the potential transformation efficiency of *S. cerevisiae*. While this selection method does allow for the use of high diversity protocols like epPCR, it is unsuitable for the evolution of Cel7A. For both native and industrial purposes, Cel7A acts on lignocellulosic substrates which have radically different chemical and physical environments than that presented in the above selection. Temperature and pH also cannot be controlled.

In the case of Cel7A, and many other enzymes, current technology limits screening capabilities to thousands of sequences (Cherry et al., 2009). In low or medium throughput screens, the relative importance of having a high percentage of active mutants in a library is high. Therefore, careful consideration must be used before selecting epPCR as the gene diversification method.

Techniques of DNA recombination create chimeric DNA sequences that contain lengths of DNA from several parental templates. The resultant libraries are tiled with segments of DNA from two or more parental templates, but in general all sequences have identical lengths. The advantage of using recombinant methods to create diversity in the DNA of enzymes for engineering is that the percentage of active mutants in the library tends to be higher. For example, Arnold and coworkers created a recombinant library of more than 6,000 P450 enzymes and found that approximately half of them were properly folded. Among those at least 75% were active (Otey et al., 2006). One reason for this is that, as opposed to the mutations that can arise in epPCR, the amino acid at each position in the recombinant sequences are as they were in functional sequences. Relative to any parental sequence, the mutations that arise through recombination are conservative in that they already existed in related and active sequences. In this way, recombinant approaches limit the scope of diversity that is introduced into the library and produce a higher percentage of active sequences. Because it is difficult to screen libraries of cellulase enzymes in high-throughput, I opted to use a DNA recombination technique called DNA Shuffling.

The DNA shuffling protocol relies on the fundamental principles in the polymerase chain reaction, namely denaturation, annealing, and extension (Stemmer, 1994). If a length of DNA is subjected to random cleavage by DNAseI, and the fragments, together with polymerase and nucleotide bases, are subjected to a primerless PCR temperature schedule, the DNA would reassemble itself. If instead, two genes sharing very low percentage nucleotide identity are cleaved randomly, mixed, and subjected to primerless PCR they would each reassemble independently. If, however, the genes share a high level of identity then crossover annealing events would occur where a fragment from one template anneals to a fragment from another. If the annealing produces free 3' ends, the polymerase will then act to extend the DNA surrounding the crossover during the extension step. In this way, libraries of chimeric enzymes can be created. This workflow is presented schematically in Figure 2-1**.** The number and distribution of crossover events is controlled by many parameters, including the percentage nucleotide identity between the parental sequences, the size distribution of randomly cleaved fragments, the temperature schedule during the annealing step, and the concentration of magnesium (Moore et al., 2000).

In this thesis, a general method for DNA shuffling was developed and applied to Cel7A. The method, called Biased Clique Shuffling differs from previously described protocols in that the level of diversity in the shuffled library can be tuned by altering template concentrations.

The selection of Cel7A DNA sequences for DNA shuffling is the first step in Biased Clique Shuffling. As discussed above, sequences for the DNA shuffling protocol must have a high proportion of identical nucleotides in order for crossover annealing events to occur. On the other hand, if the sequences share too high a level of identity, the resultant library would have low diversity. The window that was selected in this study was 68-86% identity on the amino acid level. Selecting two sequences that share this level of identity is simple. Selecting the largest sets of sequences wherein each member shares this level of identity with every other member in the set is difficult. In fact, the problem is NP-hard. This logic reduces to the clique problem in graph theory.

In the clique problem, points on a graph may or may not be connected. Connected points are said to share an edge. A clique is a set of points such that each point shares an edge with every other point in the set. If we define the points to be sequences from glycosyl hydrolase family 7, then we can define an edge to connect any two sequences that share 68-86% homology on the amino acid level. Our goal is to identify the cliques.

To generate data for the analysis, I downloaded the PFAM00840 database which contained 182 non-redundant non-truncated sequences from glycosyl hydrolase family 7, which contains both Cel7A and Cel7B enzymes. Then, an all-against-all protein BLAST (basic local alignment search tool) was performed to generate a percentage identity for every pair of sequences. A script was then coded to identify and list every pair of sequences that shared an edge. The solution of the maximal clique problem is NP-hard, and accordingly graph theorists have designed several algorithms to solve it provided the number of points is small. A publically available program that compiles on UNIX machines is called Cliquer. This program exploits a previously developed exact branch-and-bound algorithm (Östergård, 2002) and solves this maximal clique problem essentially instantaneously provided there are are relatively small number of points on the graph (i.e. hundreds).

The solution of the maximal clique problem on glycosyl family 7 hydrolases produced a convenient result. The clique was of size 11 and contained two moderately thermophilic Cel7A genes. This result presented two advantages. First, applying the DNA shuffling protocol to 11 sequences improves the probability of crossover events and increases the diversity of the library relative to one created with fewer sequences. Second, thermostability is a key engineering target for fungal cellulases. Because two sequences are already moderately thermostable, they do not have as far to go to reach stability at the target temperature (65-70°C). For these reasons, this set of sequences was selected for DNA shuffling.

The DNA encoding these 11 genes can either be obtained via isolating the genomic DNA of each respective organism or by artificial DNA synthesis. I chose to use artificial synthesis for two primary reasons. First, the genes were selected based on homology on the amino acid level. Given the differences in codon usage across organisms, the percentage of identity between amino acid sequences will be higher than that between the encoding native DNA sequences. This would have reduced the likelihood of crossover events. Using artificial DNA synthesis, the percentage nucleotide identity between DNA sequences can be made to be higher than the percentage amino acid identity between protein sequences. Therefore, the eleven genes in the glycosyl hydrolase family 7 clique were synthesized rather than isolated from their respective organisms. The genes were synthesized using *S. cerevisiae* codon bias by Genscript and DNA 2.0. The full gene sequences are presented in this theis.

The DNA shuffling protocol was applied to the set of eleven Cel7A genes that were artificially synthesized. The genes were first amplified using PCR. Then isolated bands were extracted from the agarose gel and mixed in equimolar proportions before being subjected to random digestion by DNAseI. The digestion was loaded onto another agarose gel to remove the enzymes and fragments between x and xbp were extracted from the gel. To generate crossovers, PCR was applied to these fragments without the addition of exogenous primers. The reaction is self-priming, and produced a smear that extended above and below the appropriate molecular weight of intact Cel7A (xbp). In order to isolate the properly assembled full-length Cel7A sequences, an additional PCR amplification was performed using flanking universal primers that prime the beginnings and ends of all eleven parental sequences. The agarose gels depicting the DNA at each step of this process are shown in Figure 2-2. The single band shown in the far right gel for the final amplification step represents a chimeric library of DNA with remarkable diversity, as discussed next.

To explore the efficiency of the DNA shuffling protocol described above, the DNA library was cloned into *Escherichia coli* for separation. When individual clones were sequenced, an unexpectedly high level of diversity was observed. A representative sequence is shown in Figure 2-3. This sequence contains DNA fragments from nine different parental templates and twenty-six crossovers. This high of a level of diversity and crossovers in a DNA shuffling library has not been reported previously, perhaps because DNA shuffling experiments tend to use only two parental templates that have been amplified from genomic DNA or complementary DNA (refs?). In the following chapter, the effect of this level of diversity on enzyme activity is discussed. In what follows in this chapter is a strategy to reduce the level of diversity in the DNA shuffled library in order to develop a more conservative approach. This was motivated in large part because the screening assays discussed in the next chapter only allow for the screening of several thousand chimeras.

The experiment discussed above may approach the maximum obtainable diversity when shuffling these eleven Cel7A genes because they were mixed in equimolar proportions. I hypothesized that strongly skewing, or biasing, of the proportions would result in a reduction in diversity. This theory was applied in this work. To create a biased library, the eleven amplified full-length genes were mixed in unequal proportions. One template, the moderately thermophilic and well-expressing TeCel7A comprised 50% of the total mixture. Each of the remaining ten templates comprised 5% of the mixture. The DNA shuffling protocol was then carried out exactly as described above. This time, the library was far less diverse with many sequences containing only a few mutations. A representative sequence is shown in Figure 2-4**.** In this sequence, as was typical from sequences in this biased library, the dominant contribution is from the TeCel7A sequence. This is expected because it was present at ten times the concentration during the DNA shuffling protocol. However, there are small contributions from two other parental sequences and a total of only four crossovers. These substitutions resulted in only four point mutations relative to the TeCel7A parent sequence (P144T, Y146L, D267N, E269M). This reduced level of diversity relative to that present in the equimolar library improves the likelihood that the sequence will be active. This approach to DNA shuffling was termed Biased Clique Shuffling. It allows for the control of the level of diversity and is helpful for improving the activity of a chimeric library.

Interestingly, the resultant amino acid sequences were reminiscent of those created via error-prone PCR; however, the mutations that were present arose from crossover events rather than errors in PCR, and so they existed in closely related Cel7A sequences. This increases the likelihood of obtaining active sequences. For example, the active site residues cannot be removed during a crossover event because they are conserved in all eleven sequences, along with other key amino acids. In this method, areas of high diversity across the eleven sequences see mutations more often than areas with low diversity. Additionally, the sequence will tend towards the consensus sequence. This may be an advantage as researchers have shown previously that mutating an enzyme towards the consensus sequence can result in a substantial thermostability increase (Lehmann et al., 2000).

The A Million Pieces (AMP) Method for DNA Synthesis

The DNA shuffling protocol was developed to generate diversity in DNA sequences so that they could later be screened for improved fitness. The protocol, however, can be adapted to solve a challenge associated with DNA synthesis. Commonly, researchers require a method to link two or more polynucleotide sequences together. The overlap extension PCR protocol is often used to accomplish this (Bryksin and Matsumura, 2010). In the simplest conception of overlap extension PCR, two lengths of DNA are prepared. In one, the 3' end contains identical bases to the 5' end of the other. When these DNA fragments are used as templates for PCR, they will anneal to each other and the polymerase will extend them. In this way, a DNA fragment is constructed that is the connected product of the two templates. In practice however, I found this protocol was inconsistently successful. Often, overlap-extension PCR was difficult to apply to more than two fragments and required instead a stepwise approach. Kim and colleagues were able to apply overlap-extension PCR to connect several DNA fragments for the cell-free expression of a library of modular cellulases (Kim et al., 2010). However, the multi-step PCR scheme was not obvious and was specific for this particular system. In addition, this scheme was arrived at through trial and error (personal communication).

In my work, I attempted to apply a similar overlap-extension PCR scheme to connect DNA fragments that would ultimately server as an expression construct for Cel7A in *Neurospora crassa*. The construct consisted of a 1kb 5' flanking region, a 0.7kb GPAD promoter, a 1.8kb TeCel7A gene, a 0.5kb terminator, and a 1.7kb 3' flanking region. Despite several permutations of the overlap-extension PCR method, the construction of the full length construct was unsuccessful. Presumably, the cause was the inability of the DNA fragments to cross-anneal during PCR. This may have been caused by interfering secondary-structural elements of the DNA.

An alternative strategy was pursued that applied the DNA shuffling protocol to the problem of connecting DNA fragments. The five overlapping DNA fragments were digested with DNAseI and mixed in equimolar proportions. This digested pool of DNA was subjected to primerless PCR followed by another PCR with added primers that were complementary to the intended 5' and 3' ends of the full length sequence. This resulted in a PCR product at the correct size (5.7 kb) which was the intact assembly of the five fragments. Annotated agarose gels for each step in this protocol are show in Figure 2-5 and the protocol is schematized in Figure 2-6**.** No errors were found in the sequence.

Because the method relies on digesting the overlapping DNA fragments before connecting them, it was termed the “A Million Piece” method or the AMP method.

Interestingly, by increasing the entropy of the system first (through digestion of the five fragments), I was able to ultimately reassemble the full-length product. One explanation for why the AMP method was successful when the overlap-extension method was not, is that the digested fragments had many more paths available for reassembly. In the case of overlap-extension PCR, there is only one way for two fragments to anneal. If the secondary structure of these fragments prevents that annealing, the method is likely to fail. In AMP however, the annealing between fragments may occur when no prohibitive secondary structural feature is present. In this way, the AMP method provides a robust alternative to overlap-extension PCR for the assembly of DNA fragments.

Materials and Methods

Library Design and Construction

DNA family shuffling requires that the parent genes share a high level of identity at the nucleotide level in order for crossover annealing events to occur (Stemmer, 1994). However, too high identity will limit the diversity of the shuffled library. To satisfy these constraints, we chose a window of 68–86% identity at the amino acid level. The challenge of identifying sets of GH7 sequences that share this level of identity with all other sequences in the set reduces to a well-studied NP-hard problem in graph theory: the Clique Problem (Luce and Perry, 1949). Accordingly, algorithms exist to solve it. We applied the Cliquer algorithm on 182 non-redundant non-truncated sequences from PFAM GH7 (Niskanen and Ostergard, 2003). A total of 623 pairs satisfied the identity constraint, and the maximum clique contained 11 Cel7A genes, all from organisms in the Trichocomaceae family. Because this methodology only considers the GH7 catalytic domain, eight sequences did not include a linker and CBM. A linker and CBM were appended to these to create the final set of 11 genes shown in Table I. Gene synthesis technology allowed us to generate a set of 11 DNA sequences amenable to shuffling. First, the DNA encoding the Te Cel7A sequence was synthesized by Genscript with *S. cerevisiae* codon bias. The amino acid sequences of the remaining 10 genes were then individually aligned to Te Cel7A. Where amino acids were identical, an identical codon was used (unless this resulted in poor modeled expression by DNA 2.0). Where they differed, DNA2.0 assigned a codon based on their proprietary algorithm for maximizing expression in *S. cerevisiae*. In this way, the average identity between pairs was 5% higher on the nucleotide level than on the amino acid level, and the probability and uniformity of crossover annealing events during DNA shuffling was enhanced. Two DNA family-shuffled libraries were constructed from DNAseI digested 75–200bp segments of 11 Cel7A genes. The first library was generated from an equimolar mixture of parental templates; thus, each template had a near-equal likelihood of appearing in a chimera. The second library, however, was biased towards the highly expressed TeCel7A sequence. In this library, the template proportions were skewed such that 50% was comprised of Te Cel7A and the remaining 50% consisted of an equimolar mixture of the remaining 10 sequences.

A Million Pieces (AMP) Assembly of DNA

The gene encoding TeCel7A for expression in *Neurospora crassa* was synthesized separately using an *N. crassa* codon bias. The five elements of the gene expression cassette were amplified so that a total of 3µg of DNA was purified and recovered from the agarose gel. The 5' and 3' elements are homologous to regions surrounding the *csr-1* locus so that the construct will be directed for gene integration into this locus. The promoter element is the constitutive *Myceliophthora thermophila gpdA* promoter and the terminator. Genes were amplified individually and then mixed in equimolar proportions so that the total amount of DNA was 3µg. The remaining protocol is identical to that above used for DNA shuffling. The difference is that the product is not a chimeric library but rather an intact 5.7kb gene cassette.

Catalytic domain source	Linker source	CBM source	DNA % Identity to Te	Amino Acid % Identity to Te
Aspergillus terreus	*Aspergillus clavatus*		77	71
Aspergillus fischerianus	*Penicillium decumbens*		79	73
Penicillium chrysogenum			80	75
Aspergillus terreus			80	74
Thermoascus aurantiacus	*Acremonium thermophilum*	*Aspergillus terreus*	84	83
Aspergillus nidulans	*Humicola grisea var thermoida*		78	72
Aspergillus oryzae	*Penicillium chrysogenum*		80	72
Penicillium decumbens	*Aspergillus fumigatus*		79	72
Aspergillus clavatus	*Aspergillus fumigatus*		78	71
Aspergillus fischerianus			77	74
Talaromyces emersonii	*Acremonium thermophilum*	*Agaricus bisporus*	100	100

Table 2-1. Source of Cel7A genes. Where a linker and carbohydrate binding module (CBM) were not present naturally, one was appended. The clique analysis described in the text considered only the catalytic domain; therefore crossovers that occur during DNA shuffling are likely to occur only in this domain. Still, the linker and CBM can effectively be swapped during this protocol if the crossover occurs near the end of the catalytic domain.

Input	PFAM00840
Number of non-redundant sequences	182
Percent Identity Range (amino acid)	68%-86%
Pairs Satisfying Identity Constraint	623
Maximum Clique Size	11

Table 2-2. Results of Clique Analysis. The cliquer algorithm was applied to the list of glycosyl hydrolase family 7 sequences obtained from the PFAM database. This table describes statistics associated with this analysis.

ID	Uniprot Catalytic Domain	Uniprot Linker	Uniprot CBM	%Nucleotide Identity to Te	%Amino Acid Identity to Te
SH1	Q0CRF7	A1CU44		77	71
SH2	A1DMA5	C9EI49		79	73
SH3	Q5S1P9			80	75
SH4	Q0CMT2			80	74
SH5	Q8TG37	A7WNT9	Q0CMT1	84	83
SH6	Q5B2Q4	Q12621		78	72
SH7	Q2UBM3	B6HE71		80	72
SH8	A3RG86	Q4WM08		79	72
SH9	A1CE97	B0Y8K2		78	71
SH10	A1DNL0			77	74
Te	Q8TFL9	A7WNT9	Q92400	100	100

Table 2-3. UNIPROT accession numbers for the eleven Cel7A parental genes. The clique analysis was carried out only on the catalytic domain. Later, a linker and CBM were appended from related sequences. For example, sequence SH6 contains the catalytic domain from uniprot accession Q5B2Q4 and a linker and CBM from sequence with uniprot accession Q12621.

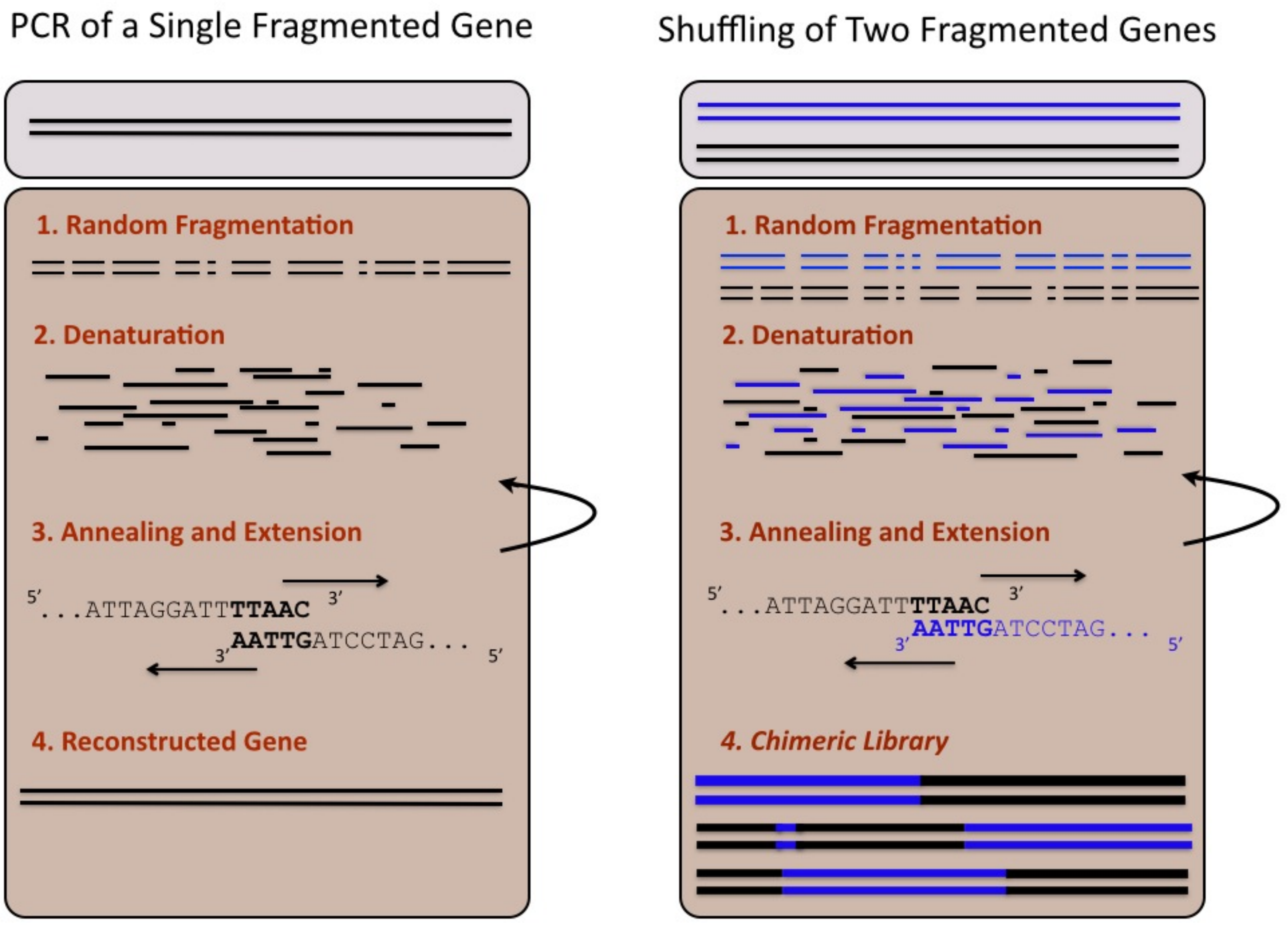

PCR of a Single Fragmented Gene
1. Random Fragmentation
2. Denaturation
3. Annealing and Extension
5′ ...ATTAGGATTTTAAC 3′
3′ AATTGATCCTAG... 5′
4. Reconstructed Gene
Shuffling of Two Fragmented Genes
1. Random Fragmentation
2. Denaturation
3. Annealing and Extension
5′ ...ATTAGGATTTTAAC 3′
3′ AATTGATCCTAG... 5′
4. Chimeric Library

Figure 2-1

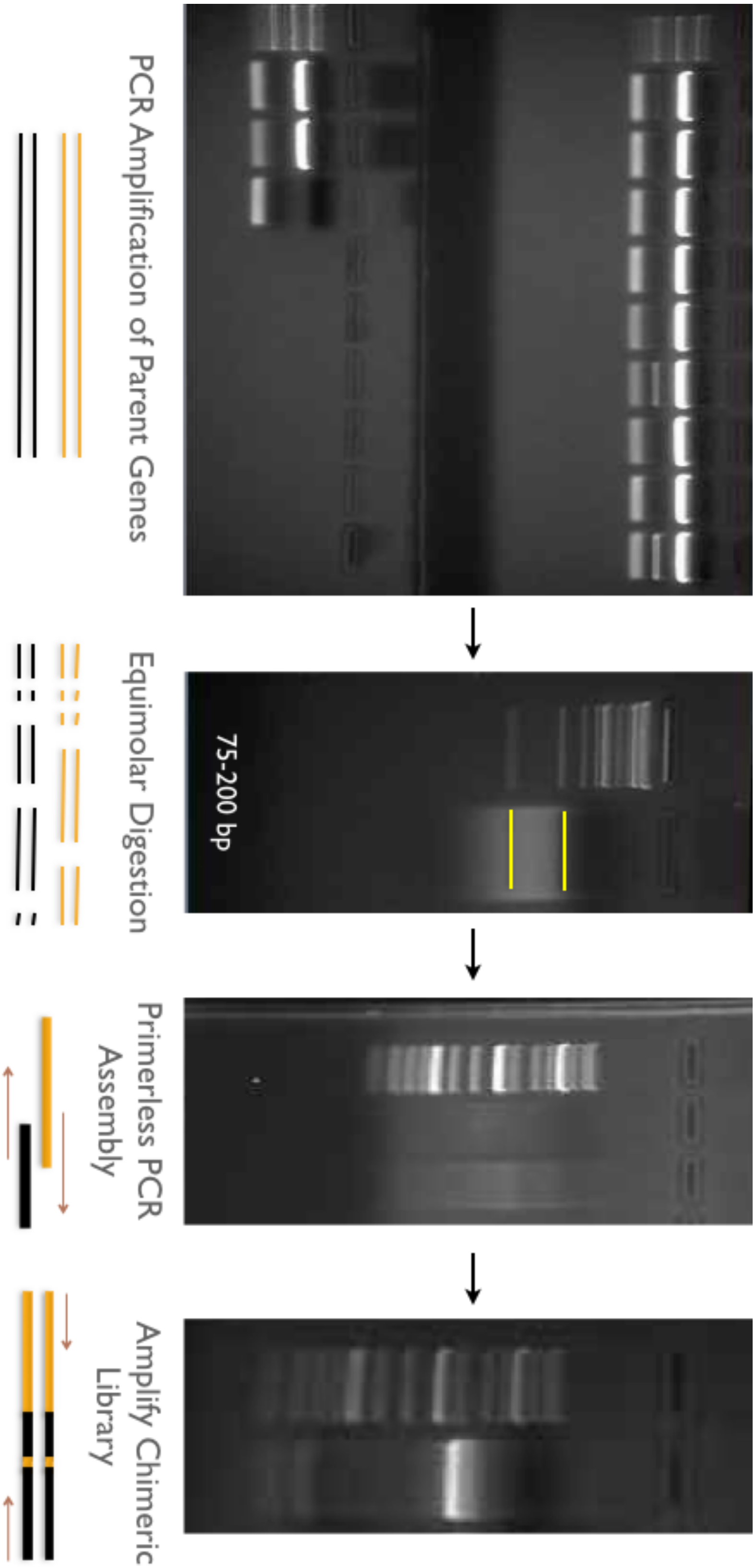
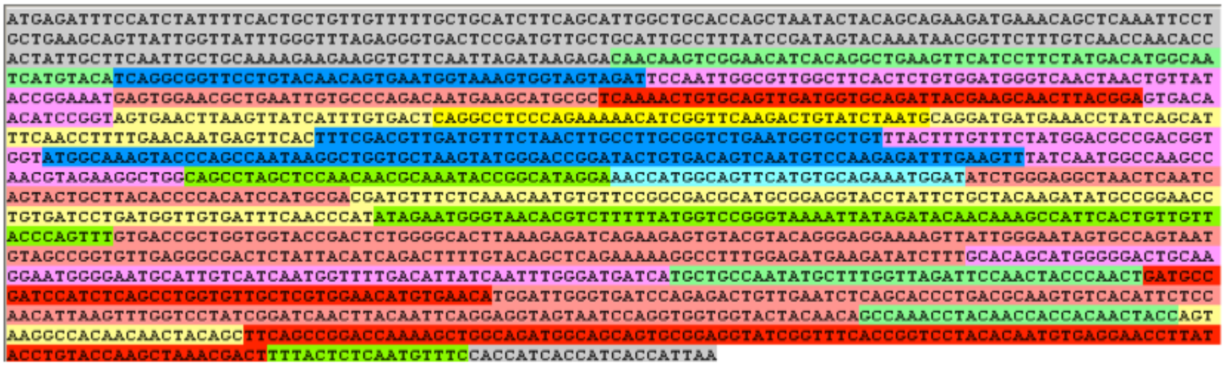
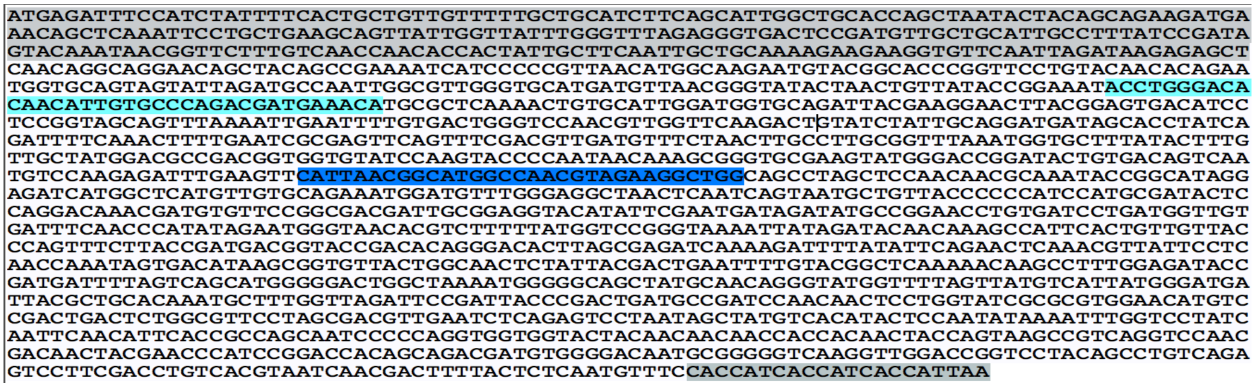

PCR Amplification of Parent Genes
Equimolar Digestion
75-200 bp
Primerless PCR Assembly
Amplify Chimeric Library

Figure 2-2



```
ATGAGATTTCCATCTATTTTCACTGCTGTTGTTTTTGCTGCATCTTCAGCATTGGCTGCACCAGCTAATACTACAGCAGAAGATGAAACAGCTCAAATTCCT
GCTGAAGCAGTTATTGGTTATTTGGGTTTAGAGGGTGACTCCGATGTTGCTGCATTGCCTTTATCCGATAGTACAAATAACGGTTCTTTGTCAACCAACACC
ACTATTGCTTCAATTGCTGCAAAAGAAGAAGGTGTTCAATTAGATAAGAGACAACAAGTCGGAACATCACAGGCTGAAGTTCATCCTTCTATGACATGGCAA
TCATGTACATCAGGCGGTTCCTGTACAACAGTGAATGGTAAAGTGGTAGTAGATTCCAATTGGCGTTGGCTTCACTCTGTGGATGGGTCAACTAACTGTTAT
ACCGGAAATGAGTGGAACGCTGAATTGTGCCCAGACAATGAAGCATGCGCTCAAAACTGTGCAGTTGATGGTGCAGATTACGAAGCAACTTACGGAGTGACA
ACATCCGGTAGTGAACTTAAGTTATCATTTGTGACTCAGGCCTCCCAGAAAAACATCGGTTCAAGACTGTATCTAATGCAGGATGATGAAACCTATCAGCAT
TTCAACCTTTTGAACAATGAGTTCACTTTCGACGTTGATGTTTCTAACTTGCCTTGCGGTCTGAATGGTGCTGTTTACTTTGTTTCTATGGACGCCGACGGT
GGTATGGCAAAGTACCCAGCCAATAAGGCTGGTGCTAAGTATGGACCGGATACTGTGACAGTCAATGTCCAAGAGATTTGAAGTTTATCAATGGCCAAGCC
AACGTAGAAGGCTGGCAGCCTAGCTCCAACAACGCAAATACCGGCATAGGAAACCATGGCAGTTCATGTGCAGAAATGGATATCTGGGAGGCTAACTCAATC
AGTACTGCTTACACCCCACATCCATGCGACGATGTTTCTCAAACAATGTGTTCCGGCGACGCATGCGGAGGTACCTATTCTGCTACAAGATATGCCGGAACC
TGTGATCCTGATGGTTGTGATTTCAACCCATATAGAATGGGTAACACGTCTTTTTATGGTCCGGGTAAAATTATAGATACAACAAAGCCATTCACTGTTGTT
ACCCAGTTTGTGACCGCTGGTGGTACCGACTCTGGGGCACTTAAAGAGATCAGAAGAGTGTACGTACAGGGAGGAAAAGTTATTGGGAATAGTGCCAGTAAT
GTAGCCGGTGTTGAGGGCGACTCTATTACATCAGACTTTTGTACAGCTCAGAAAAAGGCCTTTGGAGATGAAGATATCTTTGCACAGCATGGGGGACTGCAA
GGAATGGGGAATGCATTGTCATCAATGGTTTTGACATTATCAATTTGGATGATCATGCTGCCAATATGCTTTGGTTAGATTCCAACTACCCAACTGATGCC
GATCCATCTCAGCCTGGTGTTGCTCGTGGAACATGTGAACATGGATTGGGTGATCCAGAGACTGTTGAATCTCAGCACCCTGACGCAAGTGTCACATTCTCC
AACATTAAGTTTGGTCCTATCGGATCAACTTACAATTCAGGAGGTAGTAATCCAGGTGGTGGTACTACAACAGCCAAACCTACAACCACCACAACTACCAGT
AAGGCCCACAACAACTACAGCTTCAGCCCGGACCAAAAGCTGGCAGATGGCAGCACTGCGGAGGTATCGGTTTCACCGGTCCTACACAATGTGAGGAACCTTAT
ACCTGTACCAAGCTAAACGACTTTTACTCTCAATGTTTCCACCATCACCATCACCATTAA
```

Figure 2-3



```
ATGAGATTTCCATCTATTTTCACTGCTGTTGTTTTTGCTGCATCTTCAGCATTGGCTGCACCAGCTAATACTACAGCAGAAGATGA
AACAGCTCAAATTCCTGCTGAAGCAGTTATTGGTTATTTGGGTTTAGAGGGTGACTCCGATGTTGCTGCATTGCCTTTATCCGATA
GTACAAATAACGGTTCTTTGTCAACCAACACCACTATTGCTTCAATTGCTGCAAAAGAAGAAGGTGTTCAATTAGATAAGAGAGCT
CAACAGGCAGGAACAGCTACAGCCGAAAATCATCCCCCGTTAACATGGCAAGAATGTACGGCACCCGGTTCCTGTACAACACAGAA
TGGTGCAGTAGTATTAGATGCCAATTGGCGTTGGGTGCATGATGTTAACGGGTATACTAACTGTTATACCGGAAATACCTGGGACA
CAACATTGTGCCCAGACGATGAAACATGCGCTCAAAACTGTGCATTGGATGGTGCAGATTACGAAGGAACTTACGGAGTGACATCC
TCCGGTAGCAGTTTAAAATTGAATTTTGTGACTGGGTCCAACGTTGGTTCAAGACTGTATCTATTGCAGGATGATAGCACCTATCA
GATTTTCAAACTTTTGAATCGCGAGTTCAGTTTCGACGTTGATGTTTCTAACTTGCCTTGCGGTTTAAATGGTGCTTTATACTTTG
TTGCTATGGACGCCGACGGTGGTGTATCCAAGTACCCCAATAACAAAGCGGGTGCGAAGTATGGGACCGGATACTGTGACAGTCAA
TGTCCAAGAGATTTGAAGTTCATTAACGGCATGGCCAACGTAGAAGGCTGGCAGCCTAGCTCCAACAACGCAAATACCGGCATAGG
AGATCATGGCTCATGTTGTGCAGAAATGGATGTTTGGGAGGCTAACTCAATCAGTAATGCTGTTACCCCCCATCCATGCGATACTC
CAGGACAAACGATGTGTTCCGGCGACGATTGCGGAGGTACATATTCGAATGATAGATATGCCGGAACCTGTGATCCTGATGGTTGT
GATTTCAACCCATATAGAATGGGTAACACGTCTTTTTATGGTCCGGGTAAAATTATAGATACAACAAAGCCATTCACTGTTGTTAC
CCAGTTTCTTACCGATGACGGTACCGACACAGGGACACTTAGCGAGATCAAAAGATTTTATATTCAGAACTCAAACGTTATTCCTC
AACCAAATAGTGACATAAGCGGTGTTACTGGCAACTCTATTACGACTGAATTTTGTACGGCTCAAAAACAAGCCTTTGGAGATACC
GATGATTTTAGTCAGCATGGGGGACTGGCTAAAATGGGGGCAGCTATGCAACAGGGTATGGTTTTAGTTATGTCATTATGGGATGA
TTACGCTGCACAAATGCTTTGGTTAGATTCCGATTACCCGACTGATGCCGATCCAACAACTCCTGGTATCGCGCGTGGAACATGTC
CGACTGACTCTGGCGTTCCTAGCGACGTTGAATCTCAGAGTCCTAATAGCTATGTCACATACTCCAATATAAAATTTGGTCCTATC
AATTCAACATTCACCGCCAGCAATCCCCCAGGTGGTGGTACTACAACAACAACCACCACAACTACCAGTAAGCCGTCAGGTCCAAC
GACAACTACGAACCCATCCGGACCACAGCAGACGATGTGGGGACAATGCGGGGGTCAAGGTTGGACCGGTCCTACAGCCTGTCAGA
GTCCTTCGACCTGTCACGTAATCAACGACTTTTACTCTCAATGTTTCCACCATCACCATCACCATTAA
```

Figure 2-4

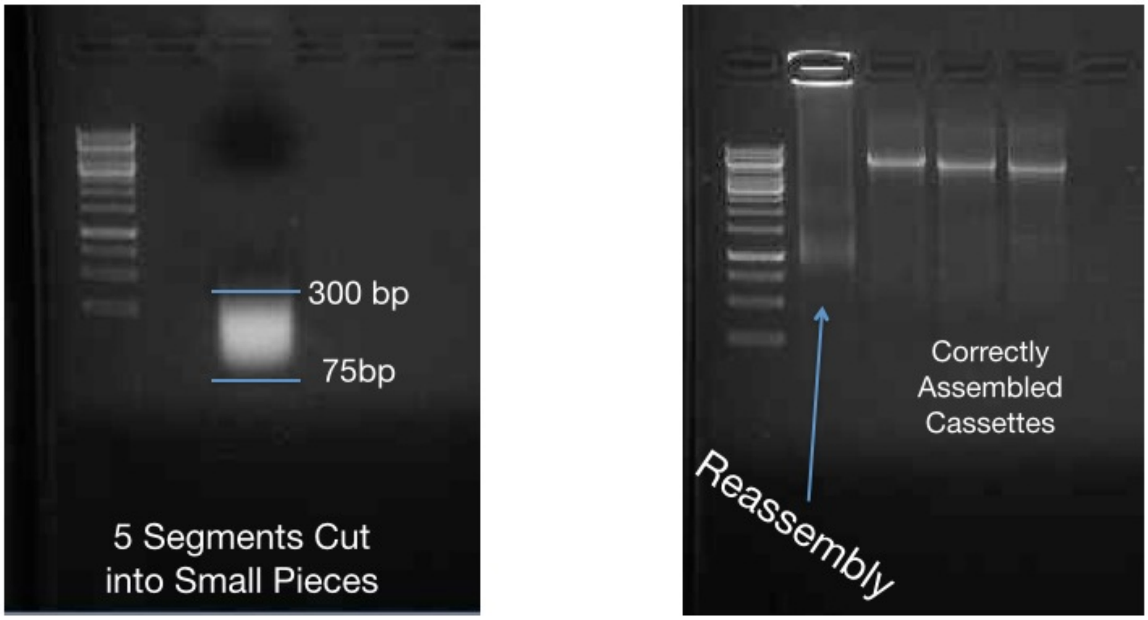


Figure 2-5

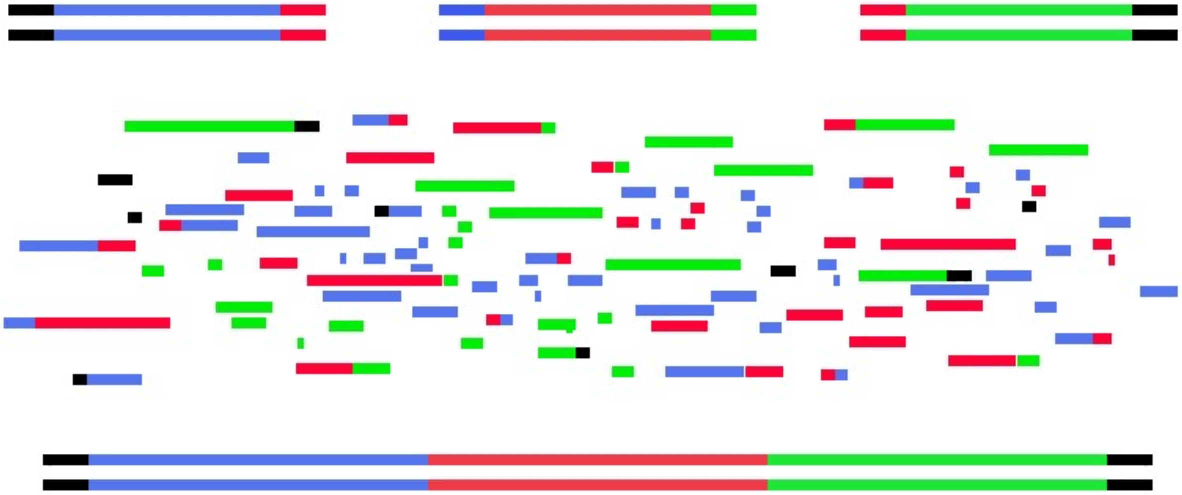


Figure 2-6

Figure 2-1. Schematic representation of the DNA shuffling method. Left: Random fragments of a single gene are subjected to PCR and reassemble. Right: Random fragments of two genes with high percentage identities are subjected to PCR. Rather than independently reassembling, the genes reassemble as hybrid chimeras because crossover annealing events will occur.

Figure 2-2. Agarose gels depicting the construction of a DNA shuffled Cel7A library. First Image: Initial amplification of the 11 Cel7A genes that share a high level of identity. The 12th lane represents a negative control. Second Image: The genes were extracted and mixed in equimolar proportions before digestion with DNAseI in the presence of manganese. Digestion was allowed to run until the distribution of fragments was between 75-200 base pairs. Third Image: The portion of fragments bracketed by the bars (between 75-200 base pairs) was extracted from the gel, purified, and used as templates for a primerless PCR reaction in this reassembly step. The reassembly is a smear with a slight emphasis at the location of the full length Cel7A gene (1.8kb). Fourth Image: In this final step called reamplification, the PCR products of the reassembly reaction are used as a template in a separate PCR reaction with flanking primers. The reassembly is not purified prior to this step. The band present at the correct size represents the DNA shuffled library.

Figure 2-3. Representative sequence from the Equimolar Library. Each color corresponds to a unique parental DNA template. There are nine colors here, indicating that this chimeric sequence combines lengths of DNA from nine different Cel7A genes. In all, there are twenty-six crossovers. The grey areas represent constant regions throughout all Cel7A parents (N-terminal APPS4 signal sequence and C-terminal 6x-His tag). The sequences from this library tended to be too diverse to be useful for screening because they were inactive as discussed in Chapter 3.

Figure 2-4. Representative sequence from the Biased Library. White corresponds to TeCel7A sequence. Each color corresponds to a unique parental DNA template. There are lengths of DNA from only two Cel7A genes in addition to TeCel7A. There are only 4 crossover points. The grey areas represent constant regions throughout all Cel7A parents (N-terminal APPS4 signal sequence and C-terminal 6x-His tag). The sequences from this library tended to have low levels of diversity.

Figure 2-5. Agarose gels depicting the AMP method for DNA synthesis. First Image: Five amplicons of various sizes were mixed in equimolar proportions and subjected to DNAseI digestion until the distribution of fragments lied between 300 and 75 base pairs. The amplicons contained overlapping regions so that the percent identity was 100% on

the ends. Second Image: As in DNA shuffling, the digested fragments were excised and purified from the gel and subjected to a primerless PCR protocol in the reassembly step which resulted in a smear. This unpurified reassembly was used as a template in a reamplification step which contained flanking primers for the full length construct. The three bands represent three correctly sized constructs. The AMP method exploits the DNA shuffling protocol to solve the problem of DNA assembly.

Figure 2-6. A cartoon depiction of the AMP assembly. Three example DNA amplicons are shown in the first row. The colors represent the identical sequences at the ends of the fragments, as is often designed for overlap-extension PCR experiments. In the AMP method, the fragments are digested before being allowed to cross-anneal, improving the probability that the annealing event will occur without hindrance from secondary structural elements.

Chapter 3

Biased Clique Shuffling for the Evolution of Cel7A

Hypersecretion of fungal cellulases is possible at titers over 100 g/L (Wilson, 2009). However, native fungal cellulases tend to have sharp temperature and pH-activity profiles, are inhibited by their reaction products, and are not as stable as many archaeal or bacterial cellulases (Bommarius et al., 2011; Bu et al., 2011; Bu et al., 2011; Graham et al., 2011; Den Haan et al., 2007; Qin et al., 2008; Ruttersmith and Daniel, 1991). Increasing the saccharification temperature (e.g., over 65°C) would reduce the risk of bacterial contamination in non-aseptic saccharification reactors. In addition, increasing enzyme thermostability creates the opportunity for recovery and recycle of active enzymes after hydrolysis. We have developed a simple and effective mutagenesis method to tune the properties of fungal cellulases and have applied it to increase the optimal temperature and enhance the thermostability of Cel7A. Cel7A comprises up to 40wt% of the secretome of *Neurospora crassa*, and removing it from the genome results in severe growth reduction on cellulose. Moreover, when Cel7A is removed from the Trichoderma reesei genome, the filter paper activity of its culture filtrate is reduced by 70% (Ghose, 1987; Phillips et al., 2011; Suominen et al., 1993; Tian et al., 2009). Thus, Cel7A is a key enzyme in these and similar cellulase systems and is a practical target for improvement by protein engineering. Previous efforts to stabilize and increase the operating temperature of *Talaromyces emersonii* Cel7A (Te Cel7A) using rational mutagenesis and SCHEMA recombination were carried out by (Heinzelman et al., 2010; Voutilainen et al., 2010). Both of these techniques require detailed knowledge of protein structure; here we present a method that only requires knowledge of primary sequence.

Cel7A activities are difficult to measure in high-throughput, as even robotics-assisted screens tend to be limited to thousands per round (Cherry et al., 2009). Therefore, it is especially important to navigate efficiently through functionally rich portions of sequence space. Recombinant methods are a common way of constraining sequence space in the search for improved enzymes (Bommarius et al., 2011; Lutz and Patrick, 2004). In this work, we improved on standard DNA shuffling (Stemmer, 1994) using an approach we call biased clique shuffling (BCS). Biasing the template proportions provided direct control over the level of diversity in the shuffled library. This concept has been modeled previously (Moore et al., 2000) in order to increase the probability of obtaining a target sequence with a specified number and placement of crossovers. Here, we show experimentally that biasing the library towards one sequence increases the percentage of active chimeras by reducing library diversity. Using this technique, we generated and screened a relatively small (469) and minimally mutated biased library from 11 Cel7A genes. Of these chimeras, 86% were active and 51 exhibited improved thermostability at 65°C.

Engineering *S. cerevisiae* to Produce Cel7A

As discussed in Chapter 1, we found S. cerevisiae without modification to be ineffective for expression of Cel7A due to low titers (<1 mg/L) and hyperglycosylation. Therefore, several techniques were combined to improve expression titers and limit hyperglycosylation of recombinant Cel7A. To address glycosylation, we targeted the following genes associated with glycan decorating of glycoproteins: KRE2, OCH1, PMR1, and MNN1/MNN9. *S. cerevisiae* strains harboring knockouts in each locus were obtained from the ATCC, and glycosylation patterning of recombinant C-terminally His-tagged Te Cel7A was monitored using Western blots. Only OCH1 and PMR1 knockouts significantly limited hyperglycosylation as shown in Figure 1-3. While these knockouts produced additional lower MW bands at 30°C, expression at 25°C resulted primarily in a single band with reduced smearing as shown in Figure 1-4. We also explored the effects of plasmid copy number, promoter, PDI activity, and signal peptide on expression titer. With a centromeric plasmid, native signal peptide, and GPD promoter, we found only traces of activity in the supernatant and cell lysate fractions. However, with the high-copy pRS424 plasmid, supernatant from yeast transformed with Te Cel7A under control of the copper inducible promoter Cup1 had five-times more activity than with the constitutive GPD promoter (data not shown). To improve expression further, we used strain YVH10, which was developed by Wittrup and coworkers (Robinson et al., 1994) to exhibit upregulated PDI activity. Finally, instead of using the native yeast a-factor prepro leader to direct secretion, we employed the engineered a-factor sequence AppS4, which was also developed by Wittrup and coworkers (Rakestraw et al., 2009). The titer of secreted Cel7A achieved with AppS4 was threefold higher than that obtained using the native a-factor. Using this expression system, 26 mg/L of purified recombinant His-tagged *T. emersonii* Cel7A catalytic domain was recoverable from culture supernatant, which represents a ~100-fold improvement in expression titer when compared to our original configurations using a centromeric plasmid, the native Cel7A signal peptide or a-factor, and the GPD promoter in strain BY4743. Finally, the PMR1 locus was removed from strain YVH10 using standard procedures. With this strain, named YVH10DPMR1, 17 mg/L of Te Cel7A with appended linker and carbohydrate-binding module (CBM) was recovered in pure form with limited hyperglycosylation. It was also convenient that this strain grew colonially on the bottom of 96-well plates, as if it were on agar, with virtually no media clouding and secreted relatively pure Te Cel7A into the medium above, making it easily adaptable to high-throughput screening. The secreted titer met our requirements for facile high-throughput screening. The remaining 10 Cel7A genes listed in Table 2-1 were still not, however, expressible at sufficient titers, which motivated a biased library design.

We found that yeast-expressed Te Cel7A was not as stable as natively expressed Te Cel7A described in the literature. The Tm for natively expressed Te Cel7A is reported to be 74°C (Voutilainen et al., 2010). Based on CD measurements, the T_m of our yeast-expressed Te Cel7A is 64°C (Supplemental Fig. 5). The T_m of yeast-expressed Te Cel7A reported by Heinzelman et al. (2010) is similar at 65°C. Interestingly, both these values are substantially less than 75°C, the Tm for yeast-expressed Te Cel7A reported by Voutilainen et al. (2010). This discrepancy is elucidated in Chapter 4.

Library Design and Construction

DNA family shuffling requires that the parent genes share a high level of identity at the nucleotide level in order for crossover annealing events to occur (Stemmer, 1994). However, too high identity will limit the diversity of the shuffled library. To satisfy these constraints, we chose a window of 68–86% identity at the amino acid level. The challenge of identifying sets of GH7 sequences that share this level of identity with all other sequences in the set reduces to a well-studied NP-hard problem in graph theory: the Clique Problem (Luce and Perry, 1949). Accordingly, algorithms exist to solve it. We applied the Cliquer algorithm on 182 non-redundant non-truncated sequences from PFAM GH7 (Östergård, 2002). A total of 623 pairs satisfied the identity constraint, and the maximum clique contained 11 Cel7A genes, all from organisms in the *Trichocomaceae* family. Because this methodology only considers the GH7 catalytic domain, eight sequences did not include a linker and CBM. A linker and CBM were appended to these to create the final set of 11 genes shown in Table 2-1. Gene synthesis technology allowed us to generate a set of 11 DNA sequences amenable to shuffling. First, the DNA encoding the Te Cel7A sequence was synthesized by Genscript with *S. cerevisiae* codon bias. The amino acid sequences of the remaining 10 genes were then individually aligned to Te Cel7A. The average identity between pairs was 5% higher on the nucleotide level than on the amino acid level, and the probability and uniformity of crossover annealing events during DNA shuffling was enhanced. Two DNA family-shuffled libraries were constructed from DNAseI digested 75–200 bp segments of 11

Cel7A genes. The first library was generated from an equimolar mixture of parental templates; thus, each template had a near-equal likelihood of appearing in a chimera. The second library, however, was biased towards the highly expressed Te Cel7A sequence. In this library, the template proportions were skewed such that 50% was comprised of Te Cel7A and the remaining 50% consisted of an equimolar mixture of the remaining 10 sequences. This design is schematized in Figure 3-1.

Engineering TeCel7A for Reduced Product Inhibition by Cellobiose

The initial biased clique shuffling library of size 469 (50% bias towards TeCel7A sequence) was screened for activity at the elevated temperature of 65°C. In addition, it was also screened for retention of activity on Methylumbelliferyl-β-D-1,4-Lactoside in the presence of cellobiose.

A common difficulty encountered in high-throughput screening is the lack of accounting for variations in protein concentration across wells. Variations in protein concentration can arise through differences in inoculum size, evaporation, as well as the inherent expressibility of the chimeric DNA sequence. This variability can in turn cause variability in signal when no variability in the specific activity of the enzymes exists. In this way, some noise is inherently introduced into a high-throughput screen of this nature. Previously, the library was screened at a temperature where the parental enzyme, TeCel7A, only retained marginal activity. In the cellobiose screen however, the parental enzyme retains a higher percentage of its activity. This leads to a higher incidence of noise.

The first round cellobiose screen was carried out with 837μM MuLac in the presence of 3mM cellobiose. After 2 hours, the relative fluorescence values were measured and recorded. As shown in Figure 3-2, there was a wide distribution of fluorescence across the library. Only sixteen chimeric enzymes, or about 3%, produced fluorescence values above that of the parental control TeCel7A. Four of these enzymes were expressed and purified using nickel affinity chromatography. The initial rates of the chimeric enzymes on MuLac were measured in the presence of 250μM cellobiose and in its absence. The relative extents of inhibition by 250μM cellobiose was used as an indicator of the chimeric enzyme's ability to tolerate cellobiose. The results of this measurement are shown in Figure 3-3. The chimeras G8 and J22 did not have improved tolerance to cellobiose relative to the parental TeCel7A enzyme. Thus, the fact that they yielded higher fluorescence values during the high-throughput assay may be more easily explained by higher specific activities than by reduced cellobiose inhibition. The data in Figure 3-3 are consistent with this explanation as these two enzymes displayed over 50% increases in activity relative to TeCel7A. However, the increase in specific activity on MuLac is not likely to correlate with a commensurate, or any, increase in specific activity on a solid and/or lignocellulosic substrate. Therefore, because G8 and J22 did not display reduced product inhibition and had irrelevant increases in specific activities towards MuLac, they were abandoned from further study.

The chimeras J4 and J14 displayed reduced product inhibition relative to the parental enzyme TeCel7A. In the presence of 250µM cellobiose, the activity of the parental enzyme TeCel7A is reduced by 37%. The activities of chimeras J4 and J14 are reduced by only 20% and 17%, respectively.

To further investigate the product inhibition of J14 by cellobiose, a similar measurement was carried out with more cellobiose concentrations. The Cel7A from *Trichoderma longibrachiatum* (*T long* CBHI) was included in this study to provide additional context, as it is closely related to the well-studied Cel7A from *Trichoderma reesei*. As shown in Figure 3-4, the 1J14 chimera was less inhibited by cellobiose than either the parental TeCel7A enzyme or the Cel7A from *T. longibrachiatum*.

Chimera 1J14 was used to enrich a second generation biased clique shuffling library. Previously, maintaining a 50% bias towards well expressing sequences produced highly active libraries. To continue this strategy, this portion of the library was split equally between TeCe7A and 1J14 so that each comprised 25% of the total library. The remaining 50% was again split evenly among the remaining 10 parental templates, as discussed previously, so that each comprised 5% of the DNA pool. This library design is depicted in Figure 3-1.

In this second generation screen, a larger number of chimeras were assayed and a more thorough approach was employed compared to the first generation screen. Separation of specific activity measurements or total activity measurements from inhibition properties was a primary target of this screen. In the first generation screen, the product inhibition signal in the single point measurement for one cellobiose concentration was somewhat difficult to discern from the noise. Therefore, in this second generation screen, a set of four cellobiose concentrations were used: 0, 100µM, 500µM, 2000µM. In addition, linear initial rates were measured via linear regression through three timepoints: 0 minutes, 22.5 minutes, and 43.5 minutes. A total of 1,152 enzymes were screened in this way, yielding 13,824 data points for analysis.

These data are illustrated in Figure 3-5 for TeCel7A and the second generation mutant 5A10. This data set was generated for each of the 1,152 enzymes in the screen. Linear regression was applied to generate slopes, or rates in relative fluorescence per time units, for each cellobiose concentration. Note the difference in rate reductions for TeCel7A and the mutant 5A10 shown in Figure 3-5. These data are made concise by replotting as rates versus cellobiose concentrations graphs as shown in Figure 3-6. The trend is made clear in this arrangement as the intersection of interpolation lines indicates that 5A10 is less inhibited than TeCel7A. These high-throughput data indicate that there is a significant difference in rates between the two enzymes at low concentrations of cellobiose; however, the screening technique employed here does not allow for the measurement of specific activities and it therefore cannot be determined if this rate differential was caused by differences in enzyme concentration or specific activity. It should be noted that despite the large gap in activities at 0µM cellobiose, the well associated with the chimeric enzyme 5A10 had more activity than TeCel7A at 2000µM cellobiose. This observation is afforded by the data-intensive nature of this screening technique and is a strong indication that the chimera 5A10 is less inhibited that the parental enzyme TeCel7A.

Despite efforts to increase the information content of the high-throughput screening data, it is still impossible to eliminate noise entirely. The robotic pipetting techniques employed by the Biomek FXP robotic liquid handling unit are precise, but not without error. At this scale, many entry points for error still exist, such as splashing, cross-contamination during cell growth, plate reader misreads and slight differences in actual hydrolysis incubation times. To gather more evidence that the chimera 5A10 is less inhibited that TeCel7A, the supernatant-containing 384-well plates containing both 5A10 and TeCel7A were thawed and assayed by hand. The results of this measurement are presented in Figure 3-7. The left panel displays the absolute fluorescence values while the right panel displays the normalized values. These results are consistent with the high-throughput data in that they also indicate 5A10 is less inhibited than the parental TeCel7A enzyme.

Engineering Cel7A for Increased Thermal Activity and Thermostability

Of 84 chimeras sampled from the equimolar library, only one had measurable activity against MuL at 37°C and none had measurable activity at 60 or 65°C. Conversely, 404 of 469 (86%) chimeras from the biased library were measurably active in the 65°C screen. Thus, the BCS library was a superior source of active chimeras compared to the equimolar one. While Te Cel7A was inactive after 13 h at 65°C, a high percentage of the BCS library remained active. In fact, between 35 and 63 h, 51 chimeras produced fluorescence at a rate at least five times the background rate and 27 had a rate of at least 16 times the background. High-throughput screening data for the most stable 12 chimeras and the parent Te Cel7A are shown in Figure 3-8.

Activities and Stabilities of Purified Chimeras

The three chimeras that produced the most product during the 65°C screen were expressed and purified. Temperature-activity profiles on MuL were then measured and compared to the profile for Te Cel7A (Figure 3-9). The maximum rate for these chimeras was achieved at 55–60°C followed by a steep decline at higher temperatures. 1G21 and 2E10 had higher MuL activity than Te Cel7A at all temperatures. 2I13 had lower activity than Te Cel7A below 60°C but higher activity at 65 and 70°C. At 65°C, the screening temperature, 1G21, 2I13, and 2E10 had 2.4-, 2.1-, and 4.0-fold higher activity, respectively. Moreover, at this temperature, 1G21, 2I13, and 2E10 retained 41%, 47%, and 65% of their optimal activities while Te Cel7A retained only 22%. The nature of the high-throughput screen emphasized stability by virtue of providing rate data over extended times. Therefore, we also investigated whether stability of the chimeras, rather than activity at 65°C, was improved. To this end, the purified enzymes were preincubated at 65°C, then immediately assayed on MuL at 50°C. As shown in Figure 3-10, two chimeras showed a different thermal denaturation pattern compared to Te Cel7A. While the initial rate decrease was rapid for all four enzymes, the activities for 1G21 and 2I13 approached constant values after 90 min. Even after 24 h at 65°C, a condition that completely inactivates Te Cel7A and 2E10, the chimeras 1G21 and 2I13 still retained 23% and 30% of their activity at 50°C, respectively. Interestingly, this is 74% of the activity each enzyme had after a 3.5-h incubation. This stability experiment therefore revealed that chimeras 1G21 and 2I13 were substantially more stable than 2E10. Long-time (22-h) hydrolysis of Avicel was carried out between 45°C and 70°C using

Te Cel7A and the three stabilized chimeras (Figure 3-11). 1G21 and 2I13 were both more active at 65 and 70°C. At 65°C, 1G21, 2I13, and 2E10 released 5.5-, 6.3-, and 2.9-fold more glucose equivalents, respectively, than Te Cel7A. At 70°C, the chimeras released 13, 15, and 2.8-fold more glucose. For chimeras 1G21 and 2I13, much of this improvement can be attributed to enzyme stability rather than to a higher catalytic rate. For example, 2E10 and 2I13 had comparable Avicelase activities at 45–50°C, but because 2E10 was relatively unstable at 65°C (Figure 3-10) it exhibited a marked decrease in long-time Avicelase activity at 65°C, losing 67% of its optimal activity achieved at 50°C. Similarly, Te Cel7A lost 88%. On the other hand, 2I13 lost only 30% (Figure 3-11) and 1G21 lost 46%. Thus, chimeras 1G21 and 2I13 are better suited for long-term hydrolysis at elevated temperatures.

Sequencing Analysis

Six sequences randomly drawn from the equimolar library were patterned by a diverse tiling of segments from the parental templates. They contained 15–21 crossovers that incorporated segments from eight to all 11 of the parents. On the amino acid level, the crossovers resulted in 70–110 mutations away from the most similar parent. While this provided evidence of an efficient family shuffling protocol, the library proved too diverse to be of value because it had so few active chimeras. Thus, the biased library approach was developed to reduce diversity. The sequences from the biased library were dominated by the Te Cel7A sequence, and even the most diverse was only 29 mutations away from the Te Cel7A sequence. Many sequences were found that had fewer than six mutations. Thus, the BCS library was far less diverse than the equimolar one. Sequences from the BCS library were categorized by the properties of the enzyme they encode: inactive, active but unstable at 65°C, active and stable at 65°C.

Seven inactive clones were sequenced. One was an empty plasmid, and another had a frameshift mutation. The five others had 4–10 crossovers yielding 29, 15, 18, 12, and 9 amino acid mutations. The inactive chimera with nine mutations also had a single insertion of three amino acids. These sequences tended to be more diverse than the active ones, which underscores the importance of controlling diversity in the library. In the active but unstable at 65°C category, six chimeras were sequenced, each of which had a flat activity profile similar to that of Te Cel7A in Figure 3-9 (data not shown). Three were the intact Te sequence, a natural contaminant in the biased library. The remaining three had four, two, and two crossovers yielding seven, six, and six amino acid mutations, respectively. In addition to the six mutations, a small three amino acid insertion was present in the third one. In the active and stable category, 12 of the chimeras that hydrolyzed the most MuL between 35 and 63 h at 65°C were sequenced (Figure 3-9). Three identical pairs were found among the 12 sequences, an artifact of plasmid duplication when cloning through E. coli. Hence, there were nine unique sequences containing three to eight crossovers. These sequences contained a total of 50 unique amino acid mutations at 44 sites. Interestingly, some sites were mutated in four of these sequences. These most prevalent mutations that did not appear in the inactive or unstable sequences were P58T or P58S, Y60L and D181N. Eight other sites were mutated in multiple stable sequences. That such enrichments emerge from a relatively sparse set of sequences suggests that assaying and sequencing a BCS library may be an effective general method to uncover sequence-activity relationships with much less effort than is required by site-directed mutagenesis, cloning, expression, and purification of individual mutants. The active stable sequences are listed in Table 3-1, and the remaining sequences appear in Table 3-2 and Table 3-3. Although many of the stable sequences contained few mutations, they are unlike sequences generated from error-prone PCR experiments. In the latter case, only a fraction of all possible amino acid mutations are likely to occur. In addition, mutations from error-prone PCR are artifactually biased towards particular amino acids (Rasila et al., 2009; Wong et al., 2006). In the BCS library, mutations in the Te Cel7A sequence were derived from the clique of homologous parental templates, and the mutational loading was controlled by the bias. Guiding mutagenesis in this way led to the sampling of sequence space that was rich in activity. The three chimeras that were purified and assayed (1G21, 2I13, and 2E10) had four to five mutations each. Eleven out of 14 total mutations among them were unique and occurred over 10 sites. All 10 were in the extensively disulfide-bonded loop network that ultimately extends over the active site to create a tunnel as shown in Figure 3-13. Chimera 1G21 contained an A3V substitution and four mutations from the A1DNL0 sequence V152M, N157T, D181N, and E183Q. These last four mutations represented a length of the A1DNL0 parent sequence from *Aspergillus fischerianus*, whereas alanine and valine were nearly equally distributed among the parents in position 3. Although these mutations spanned 180 amino acids in primary sequence, N157 and E183 are as close as 5.4A˚ in the Te Cel7A structure while A3 and N157 are within 7.2A˚ of each other. These five positions vary in conservation. Asparagine at site 181 is highly conserved, existing in all parents except Te Cel7A. Therefore, D181N represents a step towards the consensus sequence. Site 183, however, is highly diverse among the parents and has positively charged lysine, negatively charged glutamic acid, hydrophobic methionine, alanine, and leucine, as well as glutamine. Together, D181N and E183Q resulted in the elimination of two negative surface charges while preserving hydrophilicity. Position 152 is populated among the parental sequences by three leucines, three valines, and five methionines, hence V152M was relatively conservative and may have resulted in more efficient hydrophobic packing. At site 157, Te Cel7A is the only parent containing asparagine. Of the others, five are serine or threonine, three are glycine, and two are alanine. Chimera 2I13 contained two sets of mutations. First, a P58T/Y60L pair of mutations was identified from the Q0CMT2 parent sequence from *Aspergillus terreus*. Position 58, occurring in an outer loop turn, is particularly diverse among the parental sequences, and includes proline, positively charged lysine and arginine, negatively charged glutamic acid, hydrophilic serine and threonine, and hydrophobic alanine. Proline is relatively rare, however, occurring in only 15 out of 230 sequences in the PFAM GH7 alignment, while threonine is the most common at about 33%. Position 60 is highly conserved among the parents as nine out of 10 contain leucine and one contains isoleucine. Thus, the aromatic tyrosine with partial hydrophilic character at position 60 is quite different. In fact, of the 230 non-redundant sequences available in the PFAM database for GH7, only 11 contain tyrosine while about half contain leucine. The pair of mutations P58T/Y60L therefore represented a significant step towards the consensus sequence. Chimera 2I13 also contained a three amino acid substitution set S236Q, N246S, and D247T derived from the Q8TG37 sequence from *Thermoascus aurantiacus*. Q8TG37 is the only parental sequence containing a glutamine in position 236, and only 7 out of 230 sequences from the GH7 PFAM alignment have glutamine there. Of the parental sequences, four contain negatively charged glutamic or aspartic acid, five contain hydrophilic serine or threonine, and one contains glycine. Thus, in position 236, there is a large diversity of size and charge that is sampled during DNA shuffling and the occurrence of glutamine in 2I13 actually represented a move away from the consensus. In position 246, Te Cel7A is the only parent sequence that contains an asparagine while seven others contain a serine or threonine. And at position 247, Te Cel7A and three other sequences contain aspartic acid, one contains glutamic acid, and five contain threonine. As with D181N and E183Q, D247T removed another negatively charged residue and replaced it with a neutral hydrophilic one. The third chimera, 2E10, was mutated only at sites mutated in 1G21 and 2I13 yet was less stable than either. Three out of the four mutations were exactly as they were in 1G21 and 2I13: the pair P58T/Y60L and D181N. E183M, originating from Q0CRF7, A1DMA5, or Q0CMT2, was the only unique mutation present in 2E10. Notably, E183M was somewhat aggressive as it replaced a charged hydrophobic residue with a hydrophilic one. This mutation therefore may have been destabilizing in 2E10 even though the net effect of the four mutations was marginally stabilizing.

Conclusions

A biased approach to DNA shuffling was developed for engineering fungal cellulases. The BCS technique resulted in a more active library than one generated from classic equimolar DNA shuffling. The BCS library, which samples sets of consensus mutations at a probability proportional to their frequency and bias in the parental templates, was rich in active chimeras with low mutational loadings. Even this small screen of a few hundred chimeras revealed several sites that are implicated in thermostability, many of which are confined to one region of the tertiary structure. This method made use of an improved *S. cerevisiae* expression system geared for high-throughput screening of fungal cellulases. Although we demonstrated the efficiency of the BCS technique by screening only a small library, thousand more variants can be readily screened using this roboticized platform. In addition, since only 2 mL out of 250 mL of supernatant were required, a single library can be screened over 100 times for improvements in various traits. Presumably, improved chimeras from separate screens can be recombined to obtain chimeras with multiple improvements in orthogonal properties.

Materials and Methods

Saccharomyces cerevisiae Strain Engineering

S. cerevisiae with the following knockouts were obtained from the ATCC, transformed with the *T. emersonii* Cel7A, and characterized in terms of their glycosylation patterning of recombinant *T. emersonii* Cel7A using a Western blot: KRE2 (ATCC: 4034317), MNN1MNN9 (ATCC: 200180), OCH1 (ATCC: 4034406), PMR1 (ATCC: 4014534). The PDI overexpressing strain YVH10 was obtained from Prof. Dane Wittrup (Robinson et al., 1994). The PMR1 locus was disrupted using standard methods (Gueldener et al., 2002). Expression Protocol Either SC-Trp or SC-Leu was inoculated with *S. cerevisiae* containing the Cel7A gene in high-copy number plasmids (pCu424 or pCu425; (Labbé and Thiele, 1999), 1999) and grown for 3 days at 30°C, 220 rpm.

Cultures were spun down at 5,000g for 5 min and resuspended in YPD supplemented with 500 µM Cu2SO4 and allowed to express for 3 days at 25°C, 220 rpm. Supernatant was collected and purified over a GE HisTrap HP 5mL column. The centromeric plasmid tested was YCpLG, but this provided poor titers. Clique Identification Glycosyl Hydrolase 7 (GH7) sequences were downloaded from the PFAM database and clustalw version 2.0.9 was used to identify all pairs that shared amino acid percent identity within the range of 68–86%. The list was then compiled and used as input for the Cliquer algorithm (Östergård, 2002) to identify the maximum GH7 clique. Where needed, linkers and CBMs from related Cel7A enzymes were appended to the catalytic domains identified. Gene Synthesis For all genes, the native signal sequence was identified using SignalP and replaced with the mutant a-factor leader AppS4 (Rakestraw et al., 2009). In addition, six histidine residues were appended to the C-termini of the genes using the following nucleotide sequence CACCATCACCATCACCAT. The *T. emersonii* Cel7A was ordered from Genscript using their proprietary *S. cerevisiae* codon bias. The remaining 10 genes were ordered from DNA2.0 in the following manner. First, amino acid sequences were aligned to the *T. emersonii* Cel7A sequence. Where amino acids were identical, identical codons were chosen. Where amino acids differed, DNA2.0 used their proprietary S. cerevisiae codon optimization algorithm to select the codon. We subcloned these genes into pCu424 or pCu425 using standard methods.

High-Throughput Screening

Individual yeast clones were picked from bioassay dishes using a Genetix Qpix2 colony picker and transferred into round bottom 96-well plates (Corning, NY, 3359) containing 250 mL SC-Trp liquid media supplemented with 100 mg/L adenine hemisulfate. Plates were incubated for 3 days at 30°C with orbital shaking at 250 rpm. Because yeast grows on the bottom of the wells in these conditions, supernatants were readily removed with multichannel pipetting and cells were resuspended in 250 mL YPD supplemented with 500 µM Cu_2SO_4. Ten microliters of the resuspension were replicated to a stock 96-well plate containing YPD with 15% glycerol. Stock and expression plates were incubated at 25°C for 48 and 72 h, respectively. Stocks were then frozen at -80°C. Dilution 384-well plates were prepared using a Biomek FXP liquid handling robot by adding expression plate supernatant to nanopure water. Ninety-six-well expression plates were frozen at -20°C for storage, and dilution 384-well plates were carried forward for screening. Black, clear-bottom, sterile 384-well screening plates (Corning 3958) were prepared at room temperature by the Biomek. The screen was in 70 mL total volume which contained 50mM sodium acetate pH 4.8, 837 µM methylumbelliferyl lactoside (MuL; Sigma), and 1:35 final dilution of yeast supernatant to reduce background YPD fluorescence. Plates were hand sealed and incubated in a 65°C water bath. Fluorescence (excitation: 320nm, 25 nm bandwidth; emission: 450nm, 35 nm bandwidth) was measured using a Beckman plate reader. Clones of interest were grown, miniprepped (Zymo, Irvine, CA, D2004) and genes were sequenced by Elim Biopharmaceuticals, Inc., Hayward, CA. MuL Temperature Profiles of Purified Enzymes in PCR tubes, 100 mL of 0.6 µM enzyme in 50mM sodium acetate pH5 was preheated to 50°C. In separate tubes, 95 mL of 881 µM MuL, 50mM sodium acetate pH 5 was preheated to 45, 50, 55, 60, 65, or 70°C. Five microliters of enzyme solution was added to 95 mL substrate, mixed, and allowed to react for 10 min followed by transfer to a thermocycler set to 95°C and incubation for 5 min. Samples were set on ice and measured together as above.

Stability Measurement at 65°C of Purified Enzymes in PCR tubes

97.35mL of 50mM sodium acetate pH 5 was preheated to 65°C. 2.65 mL of 11.3 µM enzyme was initially added, forming a 100 mL solution containing 0.3 µM enzyme at 65°C. For the zero time point, this step was carried out at 50°C. For each time point, 7mL of enzyme was added to 63 mL of 930 µM MuL in 50mM sodium acetate pH 5 preheated to 50°C and allowed to react for 10 min followed by transfer to 95°C for 5 min and immediate icing. Fluorescence was recorded as above.

Enzyme Activity of Purified Enzymes Toward Avicel

In PCR tubes, 5mL of 11.4 µM enzyme in 50mM sodium acetate pH 5 preheated to 50°C was added to 95 mL of 4.21 g/L Avicel in 50mM sodium acetate pH 5 preheated to 50°C. Reactions were mixed and incubated for 22 h in a thermocycler. Soluble cellodextrins were measured using the glucose oxidase/peroxidase method with Amplex Red (Biovision, Milpitas, CA, 1572) as the peroxide detector (Kim et al., 2010). Supernatants were appropriately diluted and incubated with 2.5 mg/mL desalted b-glucosidase (Sigma: G0395) for 1 h at 37°C. Cellobiose controls were included to ensure complete hydrolysis. Then, Amplex Red cocktail was added so that the final concentration was 0.2mM Amplex Red, 10 units/mL glucose oxidase (Sigma: G2133), 10 units/mL horseradish peroxidase (Sigma: P6782), 100mM Hepes pH 7.45, and this solution was incubated 20 min at 37°C protected from light. Fluorescence was measured at excitation 535 and emission 595, and glucose was calculated from a linear standard curve.

Table 3-1: Mutations Present in Active and Stable Chimeras

Sites mutated in multiple stable chimeras

Site	2I13	2E4 (2M19)	2J14 (2K15)	1G21 (1I7)	2C17	2K8	2J9	2J17	2E10
A3			V	V	V				
A6			S			L	L		
P58	T				T	S			T
Y60	L				L	L			L
L108		M				M			
Q109		E				E			
N157				T		A			
D181		N		N				N	N
E183				Q					M
N246	S		E					A	
D247	T		T					T	

Sites mutated in only one stable chimera

	2I13	2E4 (2M19)	2J14 (2K15)	1G21 (1I7)	2C17	2K8	2J9	2J17	2E10
	S236Q			V152M	D57N	D43S	N10T	T299S	
					T59E	N45D	V212I	S301K	
					D64N	T48K		F306V	
					T66S	T55E			
					L73V	E65A			
					E79A	Q69T			
					S125T	E79S			
					I396V	S86T			
					P402D	S89N			
					T403I	S90A			
					D404S	K92R			
					S409A	S99A			
						S112E			
						V144T			
						A145S			
Total	**5**	**3**	**4**	**5**	**15**	**21**	**3**	**6**	**4**

Table 3-2: Mutations Present in Inactive Chimeras

	1G18	1C18	1M15	1K6	1I11
	L107M	V41T	Q190E	S87N	G151N
	L108M	D43S	S193D	S89D	V152L
	Q109K	N45D	N194S	K92T	N157G
	S112T	E183Q	N195D	N94K	N220S
	Q115E	S193A	A196K	L105V	V222F
	I116M	N195D	T198A	S193A	A354S
	R122Q	A196P	I200V	N195D	K355A
	S125T	T198A	D202G	A196P	---bt 357,358DAA
	V130A	I200V	A208P	T198A	A358S
	N132K	D202N	N246A	I200V	
	L141V	C206S	D247T	D202N	
	A145S	I396V	N327D	C206S	
	S153A	P402D	D384N		
	P156S	T403I	D388T		
	N157T	D404S	D390S		
	Q190K		P391S		
	S193D		T392S		
	N195D		I396V		
	A196K				
	T198A				
	I200V				
	D202P				
	S236Q				
	I396A				
	T400S				
	P402D				
	T403I				
	D404S				
	V407E				
Total	**29**	**15**	**18**	**12**	**12**

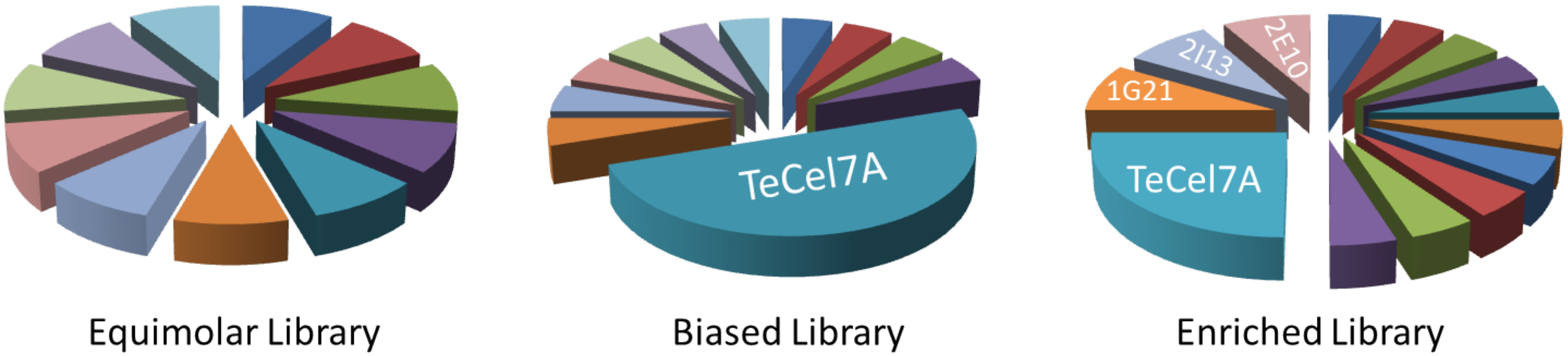
Table 3-3: Mutations Present in Active but Unstable Chimeras

	2H4	2G8	2C6	3 were TeCel7A
	Q190K	S86T	T229D	
	S193D	S89N	G231A	
	N195D	K92R	M234R	
	A196K	G98Q	S236T	
	T198A	S99A	D239S	
	I200T	Insert between 99/100: SQK	N246S	
	D202N	V101I		
Total	**7**	**9**	**6**	**0**

Table 3-1. Mutations present in active and stable chimeras. Each column corresponds to a single chimera. The top section lists sites mutated in multiple chimeras. For example, chimeras 2E4 and 2K8 contain Q109E, a glutamine to glutamate mutation in the 109th position of the TeCel7A protein without the leader sequence. The bottom section lists additional mutations that only appeared in one chimera.

Table 3-2. Mutations present in active but unstable chimeras. Each column corresponds to a single chimera. Of the chimeras sequenced in this category, three were the parental TeCel7A enzyme.

Table 3-3. Mutations present in inactive chimeras. Each column corresponds to a single chimera. On average, these chimeras had more mutations than active ones. It may be that each chimera contains at least one mutation that inactives the protein, or that novel contacts are formed between mutated residues that inactivate the enzyme.



Equimolar Library

Biased Library

Enriched Library

Figure 3-1

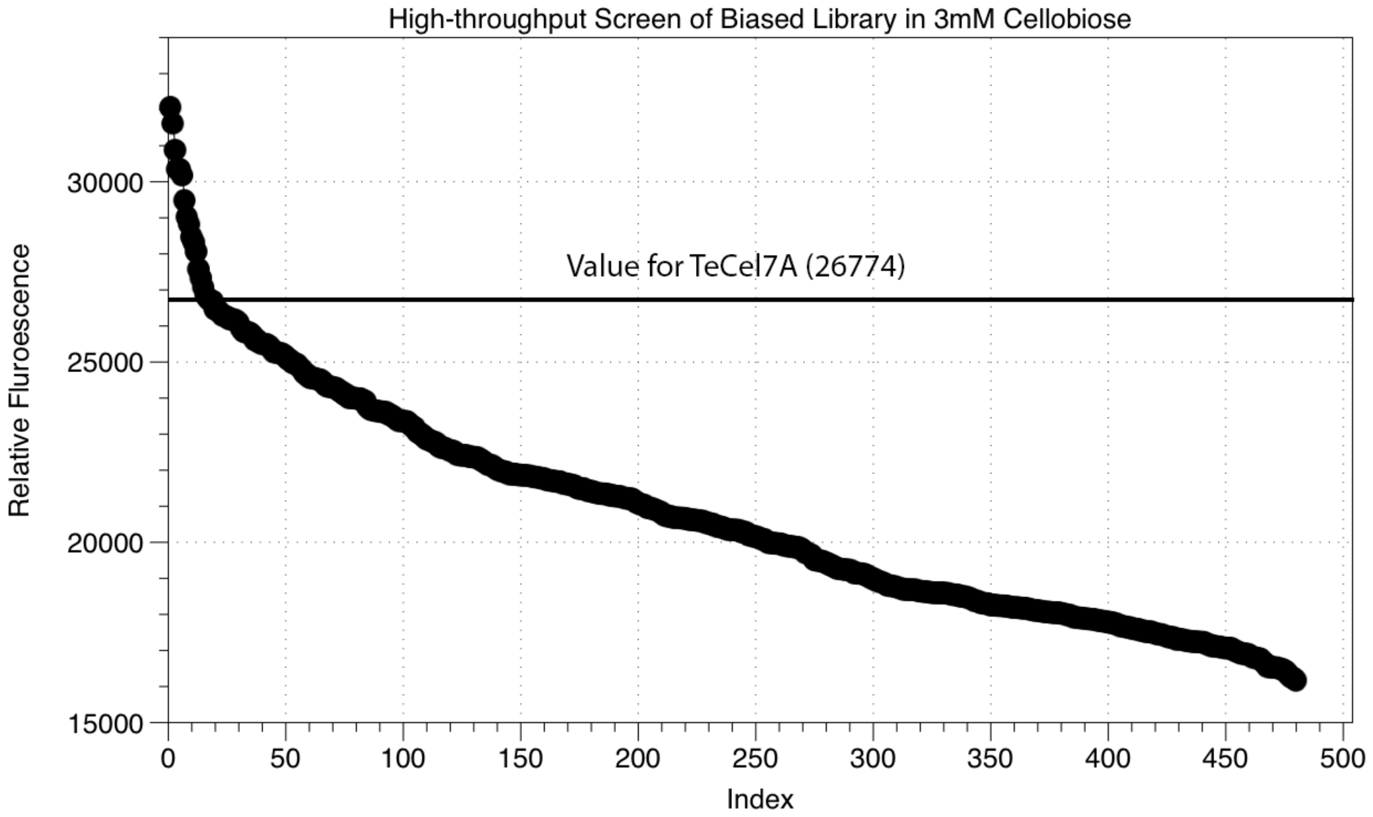

High-throughput Screen of Biased Library in 3mM Cellobiose
Value for TeCel7A (26774)
Relative Fluroescence
15000
20000
25000
30000
0
50
100
150
200
250
300
350
400
450
500
Index

Figure 3-2

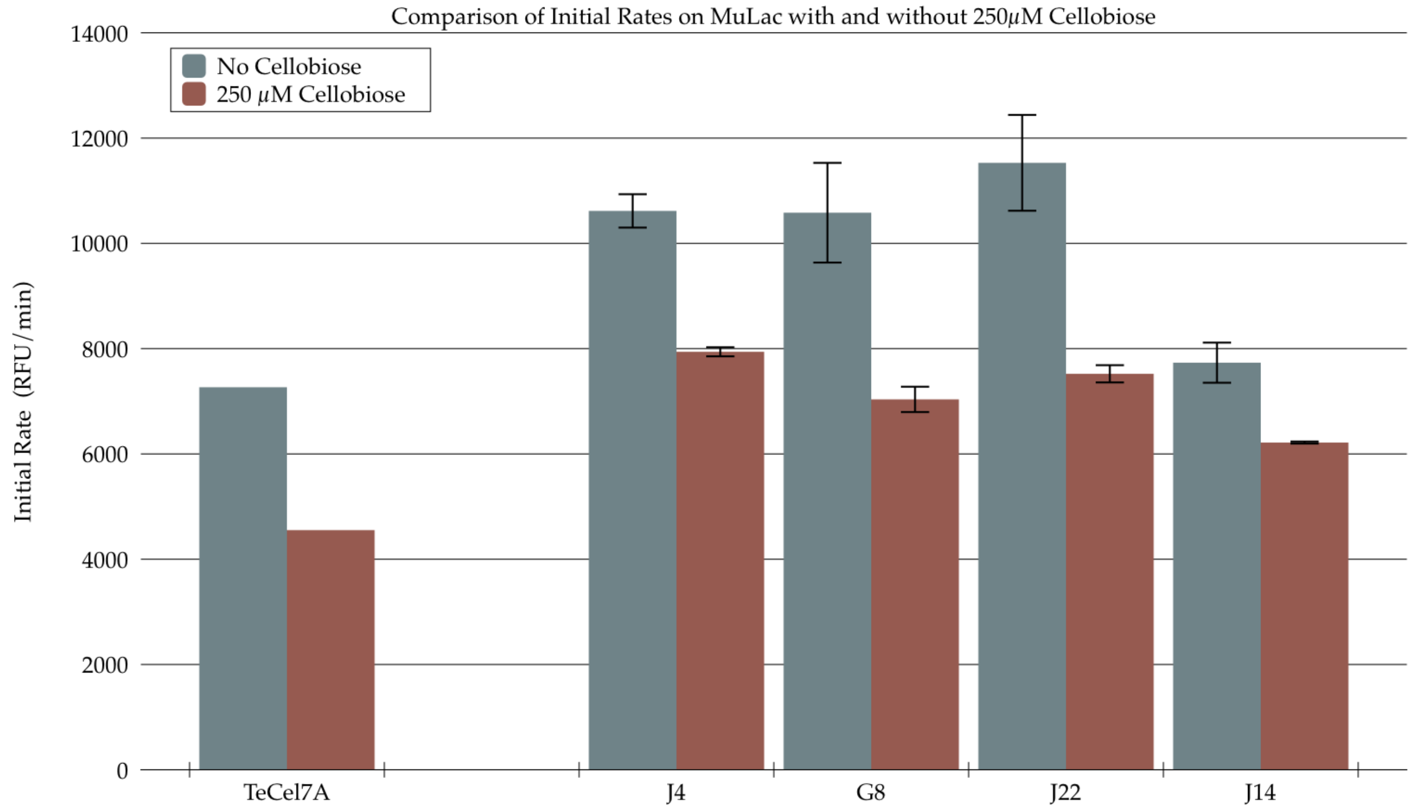

Comparison of Initial Rates on MuLac with and without 250μM Cellobiose
Initial Rate (RFU/min)
14000
12000
10000
8000
6000
4000
2000
0
No Cellobiose
250 μM Cellobiose
TeCel7A
J4
G8
J22
J14

Figure 3-3

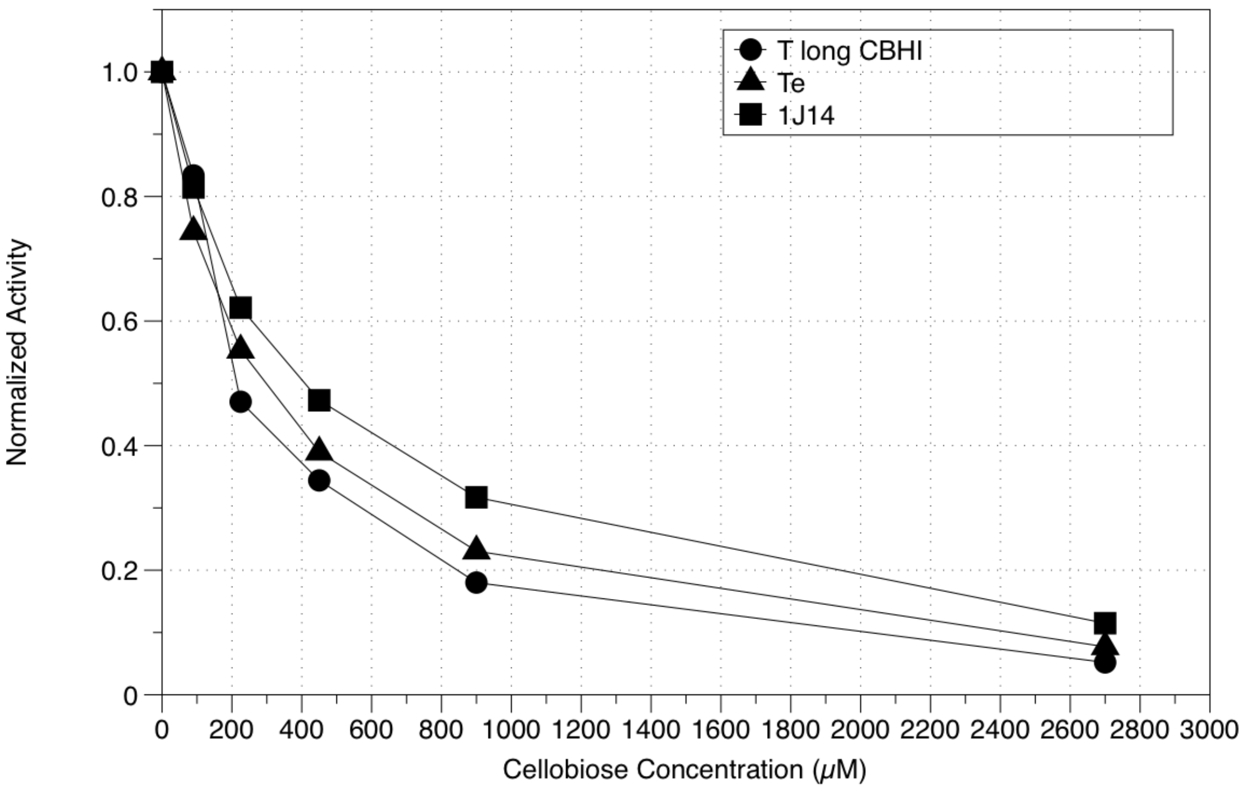

Normalized Activity
1.0
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
Cellobiose Concentration (μM)
● T long CBHI
▲ Te
■ 1J14

Figure 3-4

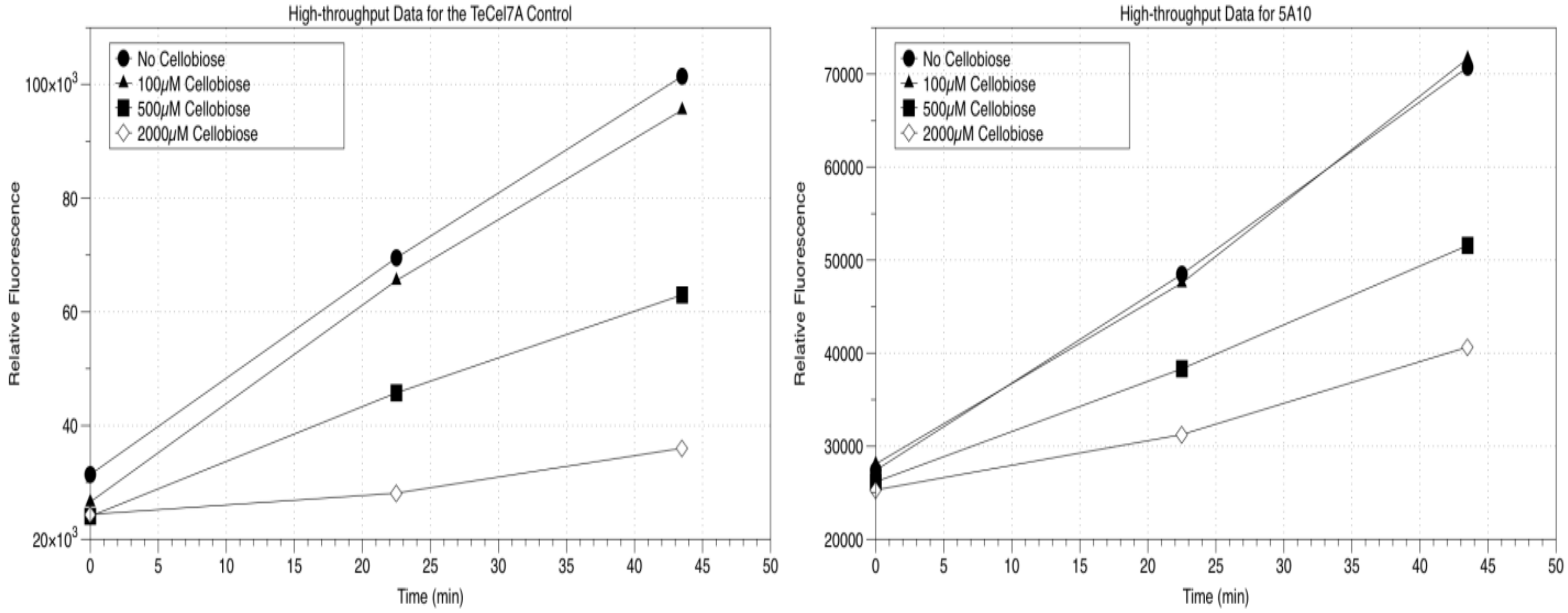

High-throughput Data for the TeCel7A Control
● No Cellobiose
▲ 100μM Cellobiose
■ 500μM Cellobiose
◇ 2000μM Cellobiose
Relative Fluorescence
100x10^3
80
60
40
20x10^3
0
5
10
15
20
25
30
35
40
45
50
Time (min)
High-throughput Data for 5A10
● No Cellobiose
▲ 100μM Cellobiose
■ 500μM Cellobiose
◇ 2000μM Cellobiose
Relative Fluorescence
70000
60000
50000
40000
30000
20000
0
5
10
15
20
25
30
35
40
45
50
Time (min)

Figure 3-5

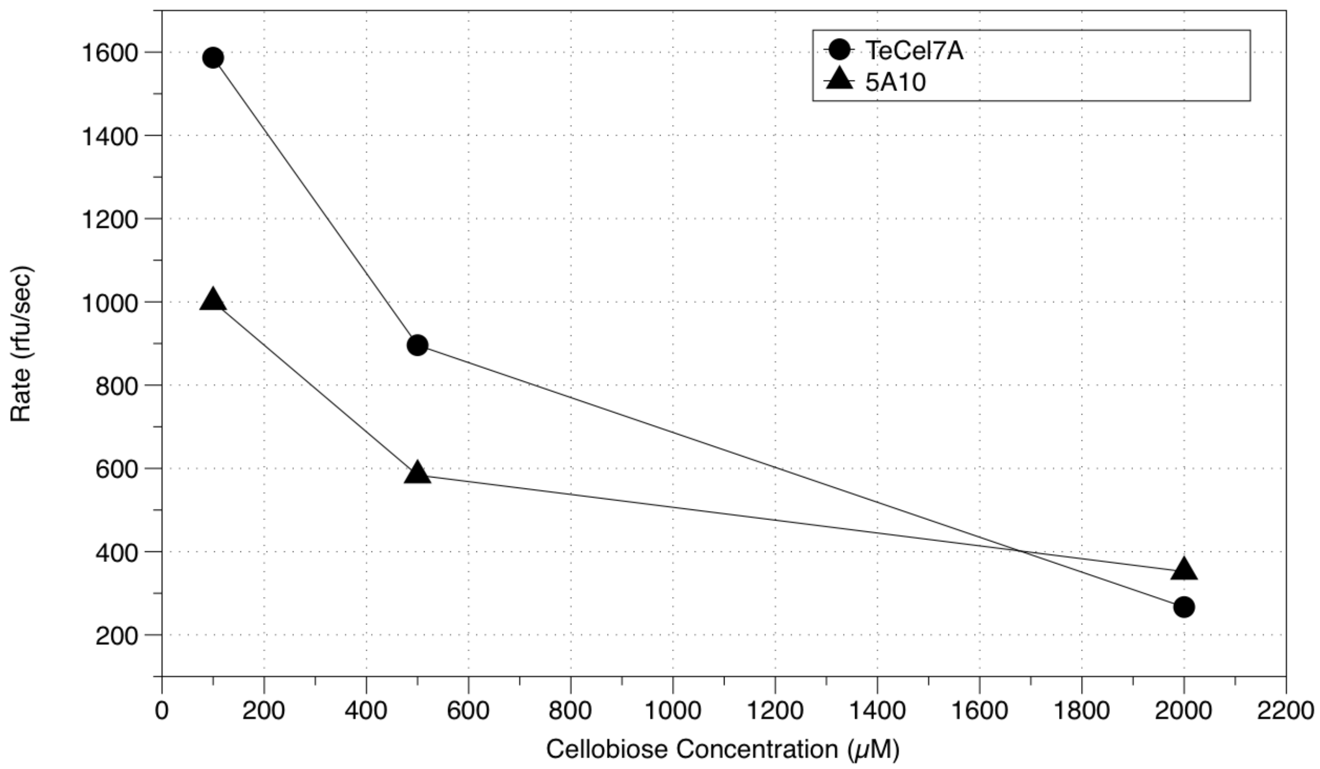

● TeCel7A
▲ 5A10
Rate (rfu/sec)
200
400
600
800
1000
1200
1400
1600
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
Cellobiose Concentration (μM)

Figure 3-6

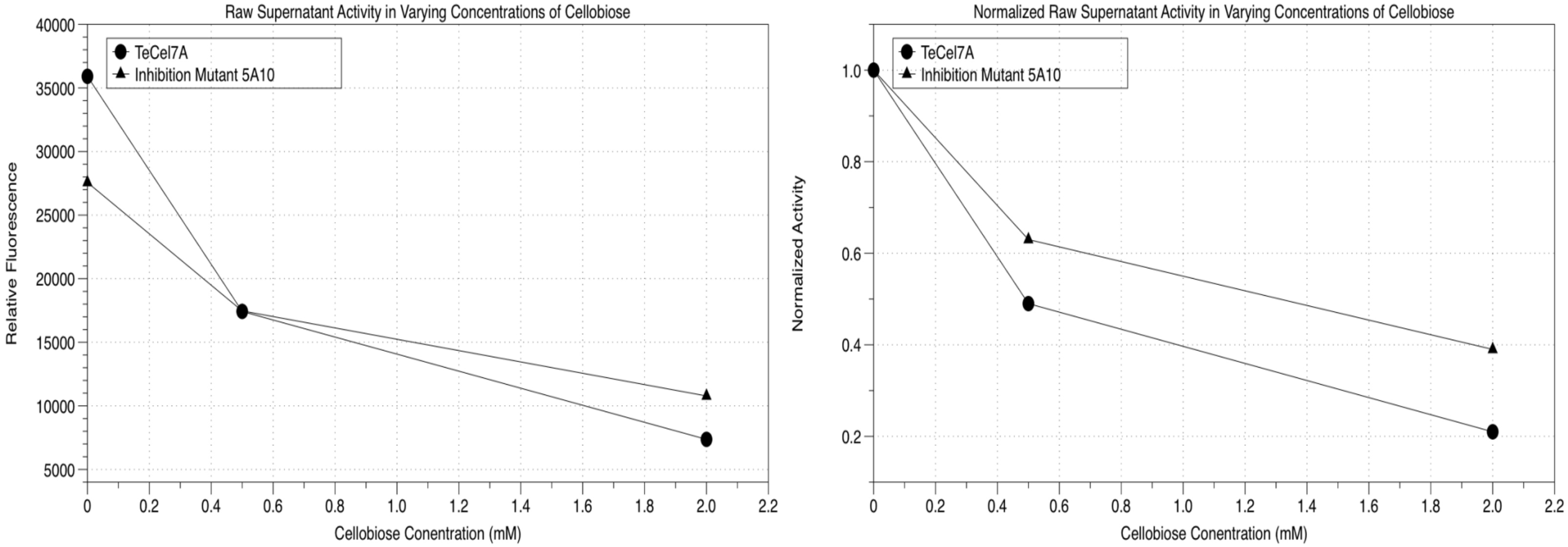

Raw Supernatant Activity in Varying Concentrations of Cellobiose
Relative Fluorescence
40000
35000
30000
25000
20000
15000
10000
5000
● TeCel7A
▲ Inhibition Mutant 5A10
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Cellobiose Conentration (mM)
Normalized Raw Supernatant Activity in Varying Concentrations of Cellobiose
Normalized Activity
1.0
0.8
0.6
0.4
0.2
● TeCel7A
▲ Inhibition Mutant 5A10
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Cellobiose Conentration (mM)

Figure 3-7

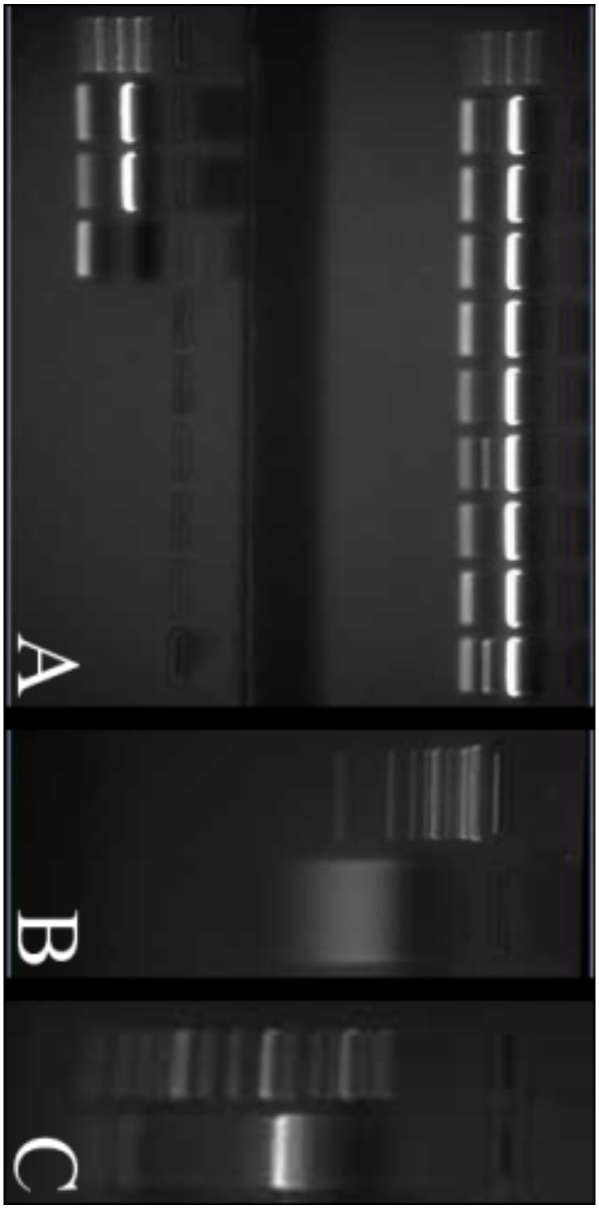


Figure 3-8

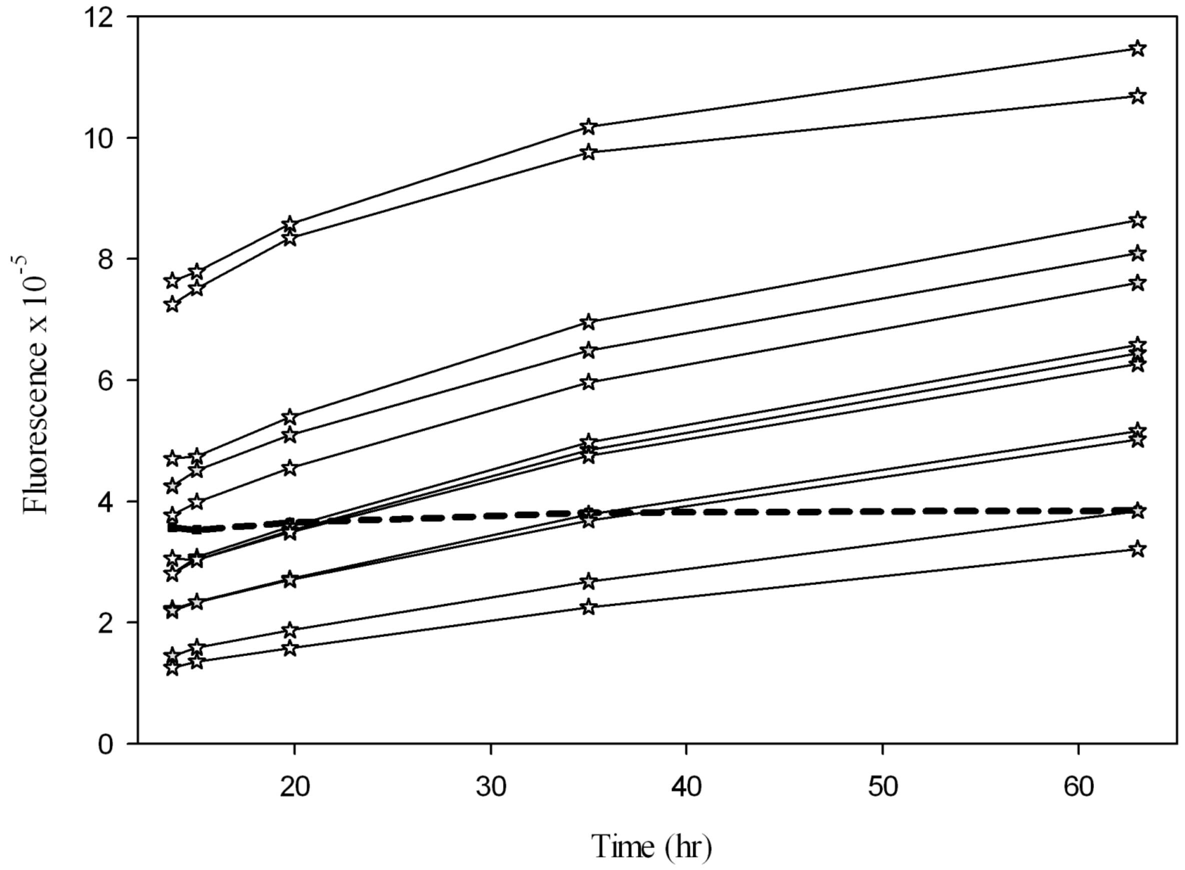


Figure 3-9

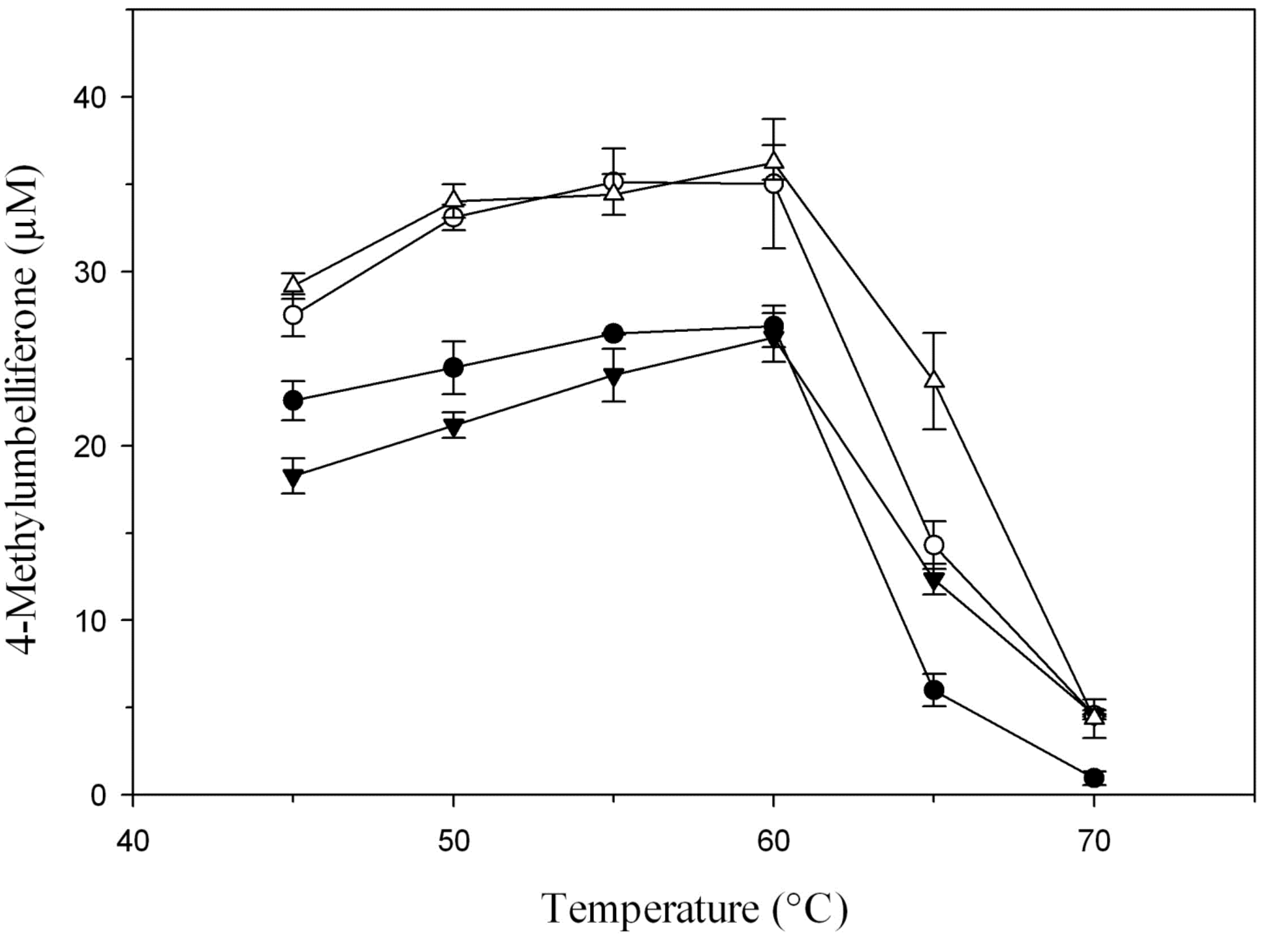

4-Methylumbelliferone (μM)
0
10
20
30
40
40
50
60
70
Temperature (°C)

Figure 3-10

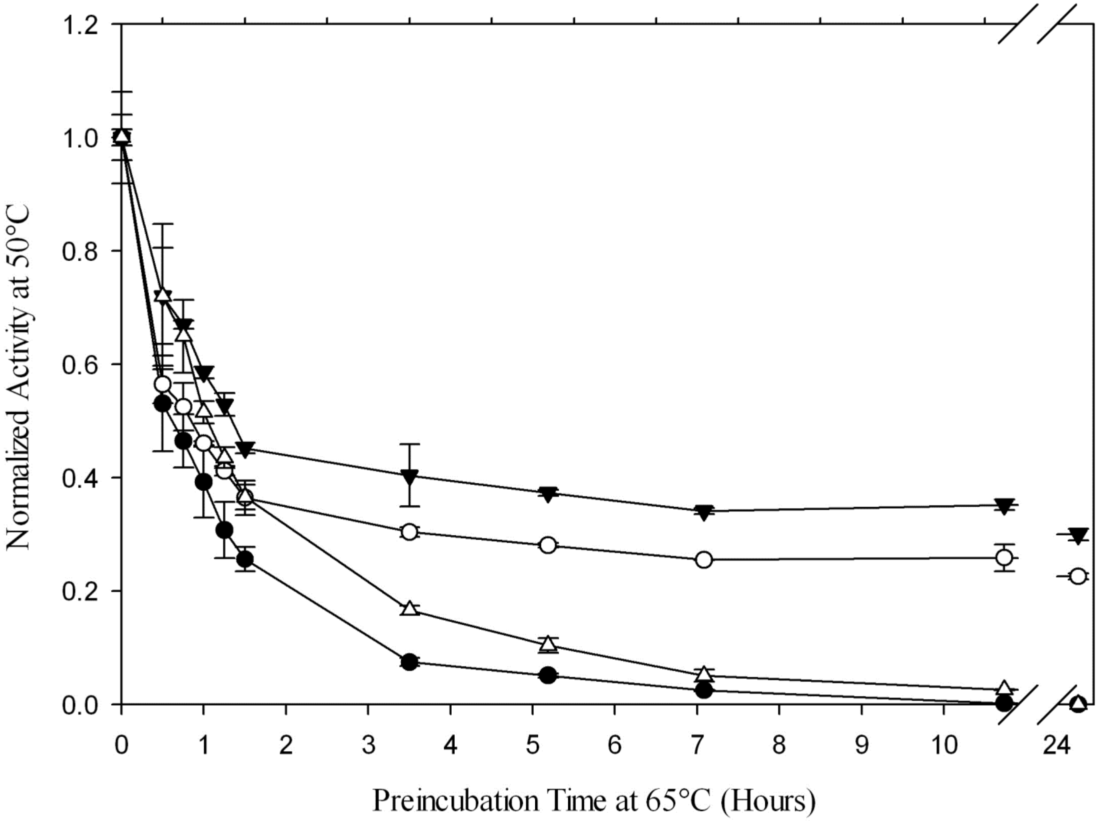

1.2
1.0
0.8
0.6
0.4
0.2
0.0
Normalized Activity at 50°C
0
1
2
3
4
5
6
7
8
9
10
24
Preincubation Time at 65°C (Hours)

Figure 3-11

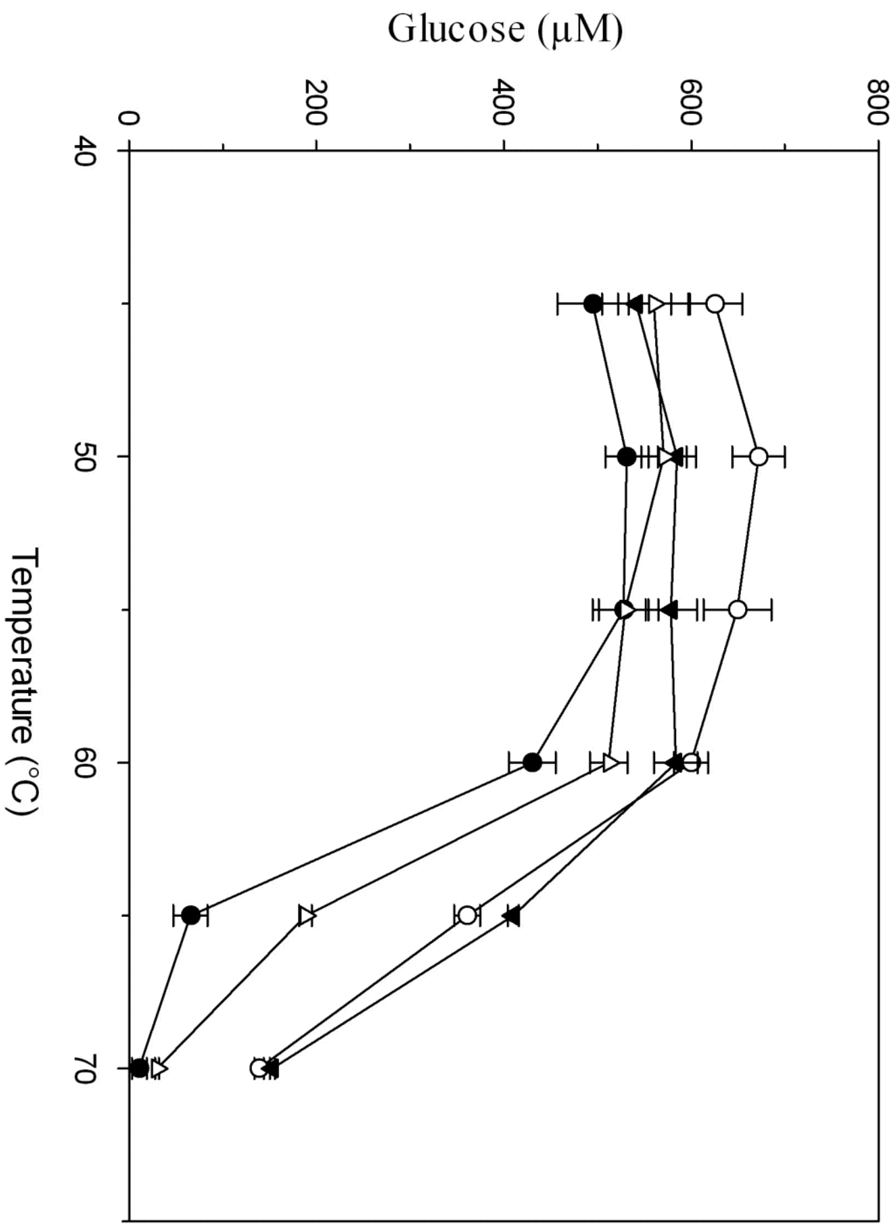


Figure 3-12

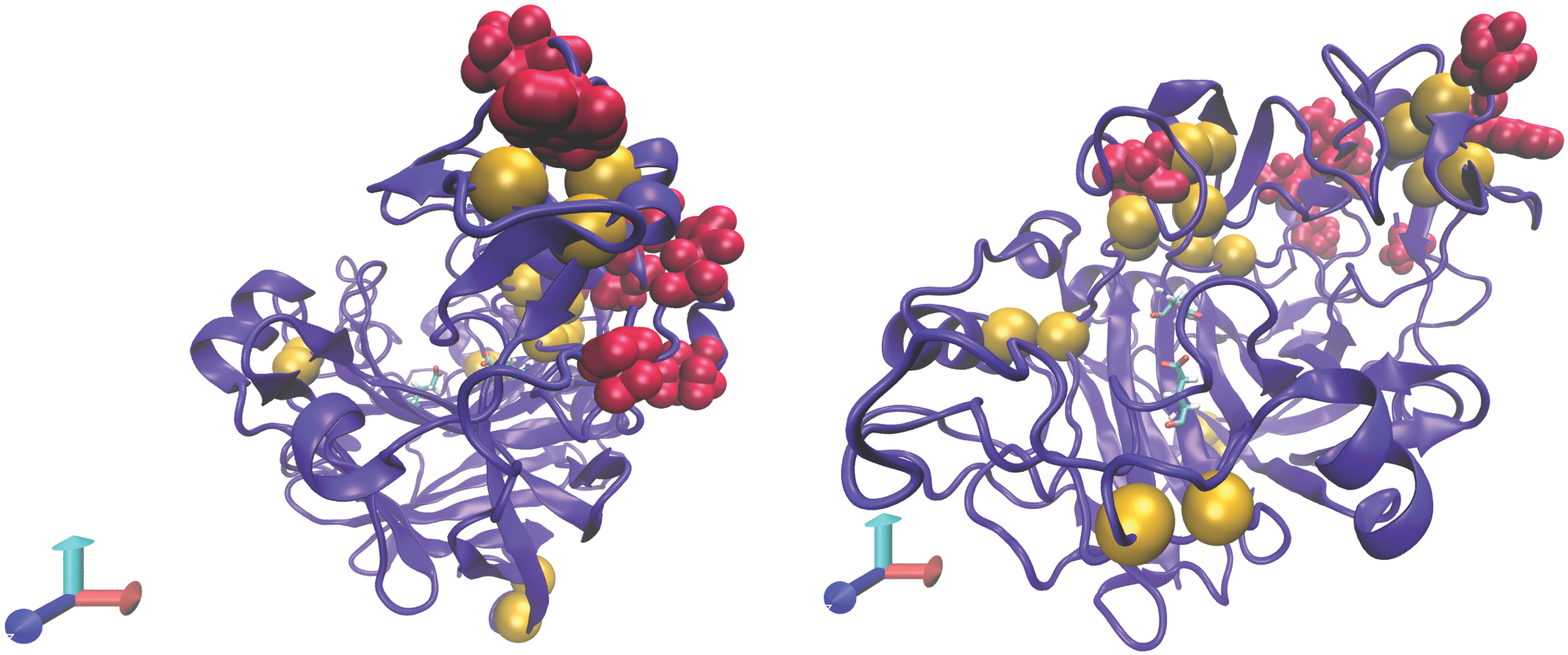


Figure 3-13

Figure 3-1. Biased Clique Shuffling Library Design Schematics. Classically, parental templates are mixed in equimolar proportions for DNA shuffling. Here, the diversity of the library is tuned via a biasing of the template proportions. Further, second generation libraries are enriched with improved sequences from the first generation library.

Figure 3-2. High-throughput screen of 50% TeCel7A biased library in 3mM cellobiose. The same initial biased library described earlier in this chapter that was screened for improved thermostability at 65°C was screened for reduced product inhibition. The screen was carried out in sealed 384-well plates incubated under water. Only sixteen chimeras yielded final fluorescence values higher than the TeCel7A control. Reaction Conditions: 50°C, 50mM sodium acetate pH 4.85, 836µM MuLac.

Figure 3-3. Comparison of inhibited and uninhibited initial rates on MuLac. The initial rate of the parent TeCel7A and four mutants were measured in the presence and absence of 250µM cellobiose. G8 and J22 had improved specific activities on MuLac, but were not any more tolerant to cellobiose. J4 and J14 were less inhibited by cellobiose than TeCel7A. Reaction conditions: 0.03µM enzyme, 837µM MuLac, 50°C, 50mM sodium acetate pH 4.85.

Figure 3-4. Inhibition curve of J14 compared to that of TeCel7A and *T. longibrachiatum* Cel7A. Activity of these Cel7A was measured at varying cellobiose concentrations, normalized to the value at zero cellobiose concentration, and plotted here. The J14 mutant is less inhibited than either TeCel7A or *T. longibrachiatum Cel7A.*

Figure 3-5. Example high-throughput data for the second generation screen. Shown here are the raw data for TeCel7A and mutant 5A10 at four different cellobiose concentrations. Note that, although TeCel7A has higher activity at low cellobiose concentrations, 5A10 released more fluorescence at 2mM cellobiose concentration. In addition, the relative loss in activity observed in 5A10 at 2mM cellobiose compared to zero cellobiose is less than that observed for TeCel7A.

Figure 3-6. Calculated rates from high-throughput data versus cellobiose concentration. Shown here are rate data at different cellobiose concentrations as calculated from the plot in Figure 3-5. A similar plot exists for each of the sequences tested in high-throughput. Observe how the 5A10 mutant is less inhibited by cellobiose than the parent TeCel7A under these conditions.

Figure 3-7. Assay of 5A10 and TeCel7A supernatant without robotics. To confirm that 5A10 was less inhibited, supernatant from the wells containing 5A10 and TeCel7A was

assayed by hand. As in the high-throughput data, the 5A10 mutant retains a higher percentage of its uninhibited activity when in the presence of cellobiose than does TeCel7A.

Figure 3-8. Schematic of Molecular Biology Steps in Biased Clique Shuffling. Agarose gels showing 1.8 kbp amplification (A), 200–75 bp digestion (B), and 1.8 kbp reassembly (C) of 11 GH7 genes in order of appearance in Table 2-1. In each gel, the ladder on the left is 1 kb GeneRulerTM Plus from Thermo Scientific.

Figure 3-9. A subset of high-throughput data from the screen at 65°C. Twelve chimeras from the biased library that remained active after 35 h at 65°C. From top to bottom: 1G21, 2E10, 2I13, 1I7, 2E4, 2C17, 2J14, 2M19, 2K8, 2J9, 2K15, 2J17. Dotted line corresponds to Te Cel7A. Assay conditions: 837µM MuL, 50 mM sodium acetate pH 4.8, 1:35 dilution of yeast culture supernatant.

Figure 3-10. Temperature-activity profile of TeCel7a and mutants on MuLac. Methylumbelliferone released after 10min of MuL hydrolysis at different temperatures (open circle 1G21; open triangle 2E10; closed triangle 2I13; and closed circle Te). Assay conditions: 50mM sodium acetate pH 5, 837 µM MuL, 0.03 µM enzyme. The temperature was controlled using a thermocycler.

Figure 3-11. Residual activity after heating at 65°C. Residual enzyme activity at 50°C after incubating at 65°C for different times (open circle 1G21; open triangle 2E10; closed triangle 2I13; and closed circle Te). The 65°C incubation contained 0.3 µM enzyme in sodium acetate pH 5. Assay conditions: 0.03 µM enzyme, 50mM sodium acetate pH 5, 837 µM MuL, 50°C, 10 min. The temperature was controlled using a thermocycler.

Figure 3-12. Temperature-activity profile of TeCel7A and mutants on Avicel. Glucose released after 22 h of Avicel hydrolysis at different temperatures (open circle 1G21; open triangle 2E10; closed triangle 2I13; and closed circle Te). Assay composition: 4 g/L Avicel, 0.57 µM enzyme, 50mM sodium acetate pH 5. The temperature was controlled using a thermocycler.

Figure 3-13. Structure of TeCel7A. Te Cel7A structure highlighting regional bias of 10 mutation sites in the 1G21, 2I13, and 2E10 chimeras. Disulfide bonds are highlighted as gold spheres. Residues where mutations occurred in one or more of the three chimeras are shown in red. Created with VMD from PDB 1Q9H (Grassick et al., 2004; Humphrey et al., 1996).

Chapter 4

The Importance of Pyroglutamate in Cel7A

The commercialization of lignocellulosic biofuels relies in part on the ability to engineer cellulase enzymes to have properties compatible with practical processing conditions. The cellulase Cel7A has been a common engineering target because it is present in very high concentrations in commercial cellulase cocktails. Significant effort has thus been focused on its recombinant expression. In particular, the yeast *Saccharomyces cerevisiae* has often been used both in the engineering and basic study of Cel7A. However, the expression titer and extent of glycosylation of Cel7A expressed in *S. cerevisiae* vary widely for Cel7A genes from different organisms, and the recombinant enzymes tend to be less active and less stable than their native counterparts. These observations motivate further study of recombinant expression of Cel7A in *S. cerevisiae*. Here, we compare the properties of Cel7A from *Talaromyces emersonii* expressed in both *S. cerevisiae* and the filamentous fungus *Neurospora crassa*. The Cel7A expressed in *N. crassa* had a higher melting temperature (by 10°C) and higher specific activity (2-fold at 65°C) than the Cel7A expressed in *S. cerevisiae*. We examined several post-translational modifications and found that the underlying cause of this disparity was the lack of N-terminal glutamine cyclization in the Cel7A expressed in *S. cerevisiae*. Treating the enzyme *in vitro* with glutaminyl cyclase improved the properties of Cel7A expressed in *S. cerevisiae* to match those of Cel7A expressed in *N. crassa*.

Commercial enzymatic hydrolysis of lignocellulosic biomass to fermentable sugars for biofuel production relies in part on the ability to engineer cellulase enzymes to be compatible with industrial process conditions. Fungal cellulase Cel7A in particular has been of interest due to its dominant presence in commercial fungal cellulase mixtures (Phillips et al., 2011). Cel7A tends to be strongly inhibited by its reaction products, is only moderately stable, and has a narrow pH range for activity (Voutilainen et al., 2008). Efforts to study and engineer this enzyme, in both industrial and academic laboratories, have relied largely on heterologous expression hosts (Cherry et al., 2009; Dana et al., 2012; Day et al., 2004; Goedegebuur et al., 2006; Heinzelman et al., 2010; Komor et al., 2012; Voutilainen et al., 2010; Voutilainen et al., 2013).

The properties of heterologously-expressed Cel7A in *Saccharomyces cerevisiae* are unpredictable. For example, Smith and coworkers report an optimum temperature (T_{opt}) of 40°C for *Trichoderma reesei* Cel7A heterologously expressed in *S. cerevisiae*, while it is well known that the T_{opt} of the native enzyme is over 50°C (Baker et al., 1992; Smith et al., 2013). Reports of T_{opt} for recombinant Cel7A deviate widely even when deliberate attempts are made to use identical expression conditions (Heinzelman et al., 2010). Voutilainen and colleagues reported that Cel7A from *Talaromyces emersonii* (TeCel7A) expressed in *S. cerevisiae* maintains 100% of its optimal activity at 65°C while other reports indicate that optimal activity is reduced by more than half at 65°C (Dana et al., 2012; Heinzelman et al., 2010; Komor et al., 2012; Voutilainen et al., 2010).

An additional complication is that expression levels of Cel7A in *S. cerevisiae* are generally quite low and depend on the primary sequence of the particular Cel7A. Moreover, Cel7A is often hyperglycosylated, running as a high molecular weight smear upwards of 200 kDa on an SDS PAGE gel (the native enzyme runs at about 66kDa) (Ilmén et al., 2011; Penttilä et al., 1988). While particular blocks of primary sequence have been implicated (Heinzelman et al., 2010), and the unfolded protein response has been demonstrated to play a role (Ilmén et al., 2011), it remains unclear why there is such diversity in expression level and extent of hyperglycosylation among Cel7A sequences.

To address these questions, we expressed TeCel7A in the cellulolytic filamentous fungus *Neurospora crassa* and measured activity of the purified enzyme as a function of temperature (**Figure 4-1**). Previously, we employed *S. cerevisiae* to express the TeCel7A catalytic domain appended to a linker and carbohydrate binding module (CBM) (Dana et al., 2012). We observed that this enzyme expressed in *N. crassa* (Nc-TeCel7A) had more than twice the activity of that from *S. cerevisiae* (Sc-TeCel7A). Furthermore, the T_{opt} of Nc-TeCel7A was higher than that of Sc-TeCel7A by 10°C. Taken together, these inconsistencies, along with those present in the literature, pointed to distinct post-translational forms of Cel7A. We set out to understand what these forms were by comparing Nc-TeCel7A and Sc-TeCel7A.

To facilitate its purification, Sc-TeCel7A has typically been produced with a C-terminal 6x histidine tag appended to the CBM. In the current work, the fungal Nc-TeCel7A was produced without this tag. To explore the effect of the his_6 tag, it was removed from Sc-TeCel7A, and its removal resulted in a substantial increase in activity of the enzyme toward the cellulosic substrate Avicel (Figure 4-1). The T_{opt} in the 15-hour assay also increased, suggesting an increase in stabilizing interactions between Sc-TeCel7A and cellulose. However, this alone was not enough to recover the specific activity or T_{opt} observed with Nc-TeCel7A.

We also considered the potential impact of improper disulfide bond formation in Sc-TeCel7A. As described previously, the *S. cerevisiae* strain employed in the present work carries an upregulated protein-disulfide-isomerase (PDI) phenotype that improves the expression titer of Sc-TeCel7A (Dana et al., 2012; Robinson et al., 1994). Reaction with Ellman's reagent was used to test for free thiols, but none were detected. Therefore, there was no evidence to suggest that unpaired disulfide bonds were responsible for the decrease in activity and stability of Sc-TeCel7A.

Understanding the effect of glycosylation on Cel7A function has been an area of intensive study (Adney et al., 2009; Beckham et al., 2012; Gao et al., 2012; Jeoh et al., 2008; Stals et al., 2004; Taylor et al., 2012). Researchers have revealed precise glycan structures and modeled how strategically-placed glycans can mediate interactions between enzyme and cellulose. The strain of *S. cerevisiae* employed in this study limits hyperglycosylation owing to a PMR1 deletion (Dana et al., 2012). It was hypothesized that Sc-TeCel7A could have had errant glycosylation patterning that caused a decrease in stability and/or activity. To test for this possibility, we treated Sc-TeCel7A with α1,2/ α1,3-mannosidase and the endoglycosidase EndoH. No increase in T_{opt} was observed; however, the specific activity was improved by up to 25% (Figure 4-1, dashed lines). This increase was more pronounced for the Sc-TeCel7A enzymes that did not contain C-terminal his_6 tags.

Having investigated the possible effects of disulfide bonds, the C-terminal his_6 tag, and glycosylation, the relevance of the N-terminal pyroglutamate commonly observed in crystal structures of Cel7A (Figure 4-2) was also examined. Glutaminyl cyclase catalyzes the cyclization of N-terminal glutamine and glutamate residues to pyroglutamate, as depicted in Figure 4-3 (Schilling et al., 2008a). Similar to an N-terminal proline, N-terminal pyroglutamate possesses no charge, which allows the protein to have the unusual property of possessing a hydrophobic terminus. This permits the N-terminal residue to tuck into hydrophobic compartments within the protein (Figure 4-2**)**. Despite a recent expansion in glutaminyl cyclase literature following the discovery of its link to Alzheimer's Disease (Schilling et al., 2008b), little is known about its presence and function in yeasts and filamentous fungi. We hypothesized that if the N-terminal glutamine was not cyclized in Sc-TeCel7A, the charged amino terminus would disrupt the protein structure by favoring interactions with water, resulting in a void space in the hydrophobic region it would otherwise occupy. The solvent-exposed terminus would be accessible *in vitro* to exogenously added glutaminyl cyclase. To test this hypothesis, Sc-TeCel7A was treated with human glutaminyl cyclase.

The activity and stability of the cyclase-treated enzyme were significantly improved. The melting temperature increased by nearly 10°C (Figure 4-4, Table 4-1), the specific activity doubled at 65°C, and the T_{opt} increased by 5°C (Figure 4-1). In addition, the glutaminyl-cyclized Sc-TeCel7A enzyme was virtually indistinguishable from Nc-TeCel7A in T_m, T_{opt}, and specific activity. This indicated that TeCel7A requires an N-terminal pyroglutamate for optimal function, and that the primary cause for the differences in activity and optimal temperatures of Sc-TeCel7A and Nc-TeCel7A was the lack of glutamine cyclization in the former. The absence of glutaminyl cyclase activity in the *S. cerevisiae* secretory pathway has not been previously noted, but it appears to be important for full function of Cel7A.

Sc-TeCel7A samples that were exposed to sodium phosphate, pH 7.5, for several days during nickel affinity chromatography purification and then stored in sodium acetate, pH 5, at 4°C for 5 months were analyzed using differential scanning calorimetry. Two peaks were identifiable: one that coincided with the T_m of fresh Sc-TeCel7A and one that coincided with glutaminyl cyclase treated Sc-TeCel7A (Figure 4-4). This indicates that the N-terminal glutamine spontaneously cyclized at a non-neglible rate during enzyme storage and could explain discrepancies previously reported in Sc-TeCel7A properties. Spontaneous cyclization of N-terminal glutamine on the timescale of several days has previously been observed in monoclonal antibodies and was found to be dependent on a number of parameters, including buffer composition, pH, and temperature (Dick et al., 2007).

Previously, we applied the Biased Clique Shuffling technique to evolve Sc-TeCel7A for improved activity at higher temperatures and noted that the mutations tended towards one region of the tertiary structure (Dana et al., 2012). Interestingly, this region encompasses the N-terminal glutamine, suggesting that the mutations present in the thermostable mutants may have been compensatory in nature; that is, they may have partially rescued the stability of the non-cyclized form of the enzyme.

Researchers have spent considerable effort developing expression of Cel7A in *S. cerevisiae* and have demonstrated varying extents of activation of the unfolded protein response depending on the specific Cel7A enzyme expressed (Ilmén et al., 2011). Cel7A enzymes may vary in their tolerance to a charged N-terminus, and those that cannot fold efficiently trigger the unfolded protein response in *S. cerevisiae*. The absence of N-terminal cyclization may be the underlying cause of the varying activities, glycosylation patterns, and expression titers reported with heterologous Cel7A expression in *S. cerevisiae*.

Using an *S. cerevisiae* expression system that possesses glutaminyl cyclase activity toward Cel7A may improve expression titers and reduce variability in glycosylation patterning, and should improve specific activity of Cel7A produced in yeast. We are currently examining the expression of not only Cel7A, but also of Cel7B and other proteins that natively contain N-terminal pyroglutamate. This protein modification pervades the fungal secretome, appearing in xylanases, amylases, and cellobiose dehydrogenases, among other proteins, and may explain differences in activites prevalent in the literature.

Materials and Methods

Glutaminyl cyclase treatment

Lyophilized human glutaminyl cyclase recombinantly expressed in insect cells with his_6 tag purchased from Sino Biological, China (13752-H07B) was reconstituted to 0.2 mg/mL with 100µL water and diluted 1:10 into the 36 hour cyclase reaction at 30°C with 80µM Sc-TeCel7A in 50mM sodium phosphate pH 7.4.

Expression of Sc-TeCel7A

T. emersonii Cel7A appended to a linker and CBM with C-terminal his_6 tag was expressed and purified from *S. cerevisiae* as described previously (Dana et al., 2012). Sc-TeCel7A without his_6 tag was purified using anion exchange chromatography: GE MonoQ 10/100 GL column, running buffer 50mM sodium phosphate pH7, elution buffer same with 1M NaCl, gradient 0-50% in 40 minutes at 4mL/min. Samples were then buffer exchanged into 50mM sodium acetate pH 5.

Glycosidase treatment of Sc-TeCel7A

Sc-TeCel7A was treated with $EndoH_f$ (NEB P0703S) and α1,2/α1,3 mannosidase (NEB P0729S) following manufacturer's recommended conditions and then purified using anion exchange chromatography.

Cloning, expression, and purification of Nc-TeCel7A

The gene encoding Nc-TeCel7A was synthesized separately using an *N. crassa* codon bias and was cloned in pCSR1:GPD vector, which directs gene integration to the *csr-1* locus. Gene expression is promoted by the constitutive *Myceliophthora* *thermophila gpdA* promoter (personal communication by Dr. T. Starr & Prof. N. L. Glass, Energy Biosciences Institute, UC Berkeley, CA). The *N. crassa* wild type (WT) strain (FGSC2489) was obtained from the Fungal Genetics Stock Center (FGSC, University of Missouri, Kansas City, Missouri, USA). Strains were pre-cultured on Vogel's medium (VM) agar supplemented with 2% sucrose for 3 days in the dark at 30°C then shifted to 25°C constant light for 5 additional days to obtain conidia for transformations and growth experiments. Transformation by electroporation for construct integration into the *csr-1* locus was performed as previously described (Bardiya and Shiu, 2007). Positive transformants were verified by PCR genotyping of the *csr-1* locus using the following genotyping primer set: 5'-CCGCGGTAGTCGTTGTTGGAAG-3' and 5'-GTACATCAAGGCGAACCTACGTCC-3'. Positive transformants were screened for heterologous protein expression in 24-well plate format. Briefly, 3mL of VM supplemented with 3% glucose was inoculated with 10^6 conidia per mL and grown at 25°C and 200rpm for 44hr. Supernatants were assayed for protein concentration and cellulase activity. The strains with the highest specific activity were chosen for larger-scale expression. One-liter of 3% glucose VM in 2L Erlenmeyer flasks was inoculated at 10^6 conidia per mL and grown at 25°C, 200rpm for 44hr. Culture broths were filtered through glass microfiber filters (934-AH, Whatman) followed by 0.22 µm PES filters (Corning). Nc-TeCel7A was precipitated from the culture supernatant with 75% ammonium sulfate saturation at 4°C. The supernatant was spun down and the precipitated proteins were resuspended in 50mM sodium phosphate pH 7 buffer and desalted using HiPrep 26/10 desalting column with 50mM sodium phosphate pH 7 as the running buffer. The protein was then loaded on a MonoQ column and eluted using sodium phosphate pH 7 buffer containing 1M NaCl (gradient: 0-50% in 40 minutes at 4mL/min) before final buffer exchange into 50mM sodium acetate pH5 for storage.

Avicel Assay

96-well PCR plates were filled with 85µL of 11.76 g/L Avicel in 50mM sodium acetate pH5 and preheated to the set temperature. Then, 15µL of 1.33µM enzyme was added to initiate the reaction and mixed thoroughly. Each reaction was carried out in duplicate wells. Reactions were stopped by placement on ice after 15 hours and spun down at 4°C. Supernatants were pipetted off and mixed with beta-glucosidase from almonds (Sigma G0395), which was thoroughly buffer exchanged to remove sugars, and allowed to react for 3 hours at 37°C. Controls were included to ensure complete hydrolysis of the cellobiose by β-glucosidase. Glucose was measured based on the amplex red assay for peroxide formation using glucose oxidase combined with peroxidase, as described previously (Dana et al., 2012; Kim et al., 2010).

Differential Scanning Calorimetry

TeCel7A (0.8 mg/mL) in 50mM sodium acetate pH 5 was loaded into the sample capillary of a TA Instruments NanoDSC while buffer alone was loaded into the reference capillary. After equilibration, a temperature schedule from 25°C to 85°C was employed using a ramp rate of 1°C/min. Raw data are presented and T_m corresponds to the apex of each peak.

Cellulase Preparation		T_m (°C)
N. crassa	**TeCel7A**	74.7
S. cerevisiae	**TeCel7A**	64.2
S. cerevisiae	**TeCel7A**	73.9

Table 4-1. Melting temperatures (T_m) of TeCel7A preparations measured by differential scanning calorimetry. The melting temperature of TeCel7A expressed in *S. cerevisiae* was about 10 degrees celsius less than that TeCel7A expressed in *N. crassa*. Treating the TeCel7A expressed in *S. cerevisiae* with glutaminyl cyclase caused a 10 degree increase in melting temperature to reach a value very near that of TeCel7A expressed in *N. crassa*.

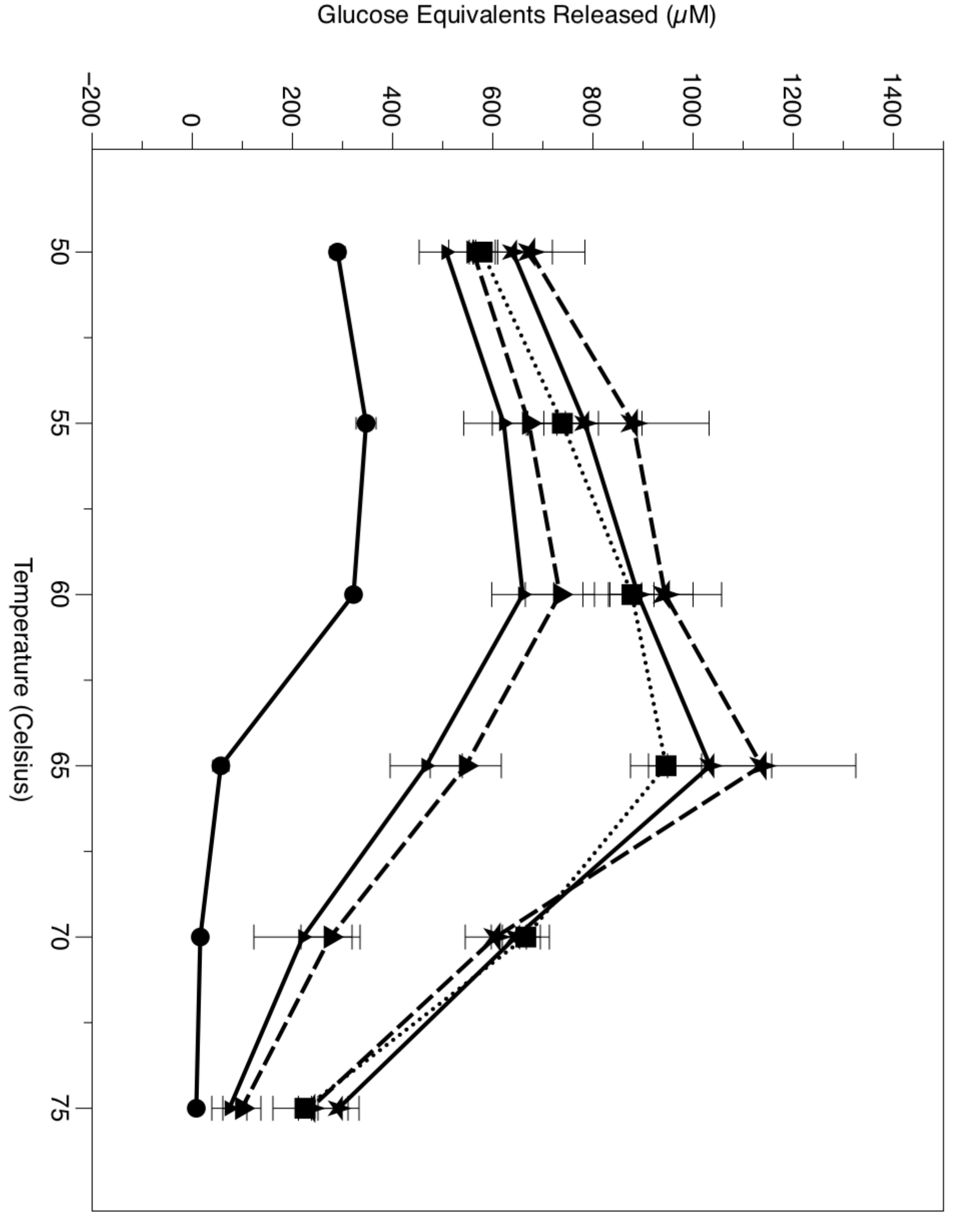

Glucose Equivalents Released (μM)
−200
0
200
400
600
800
1000
1200
1400
Temperature (Celsius)
50
55
60
65
70
75

Figure 4-1

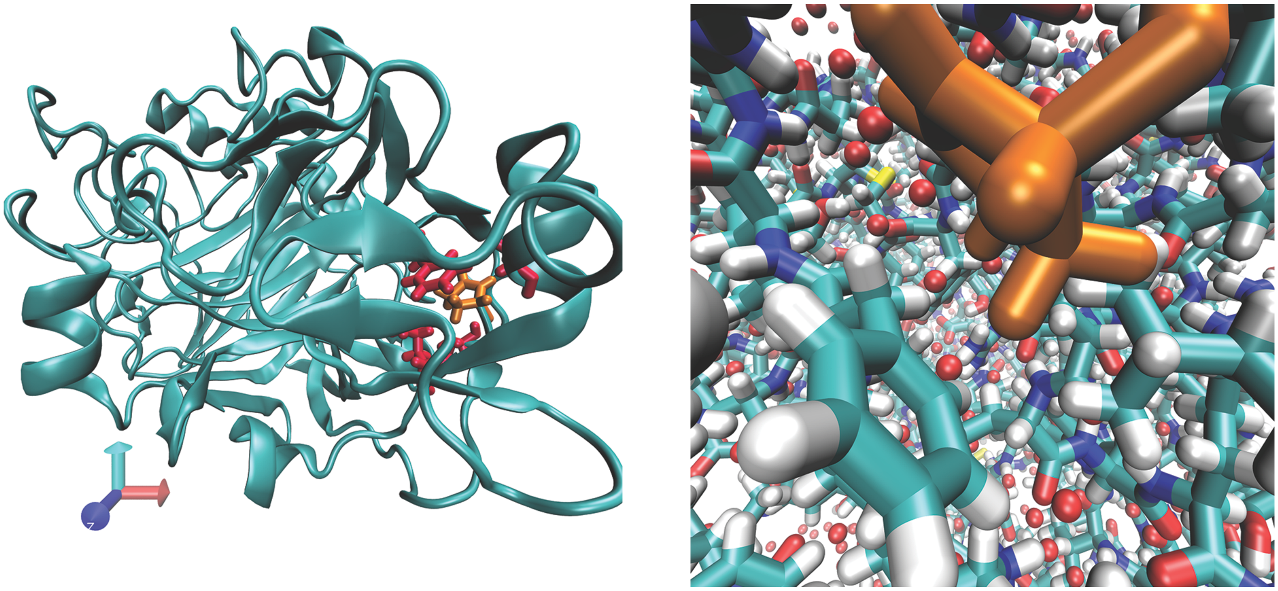


Figure 4-2

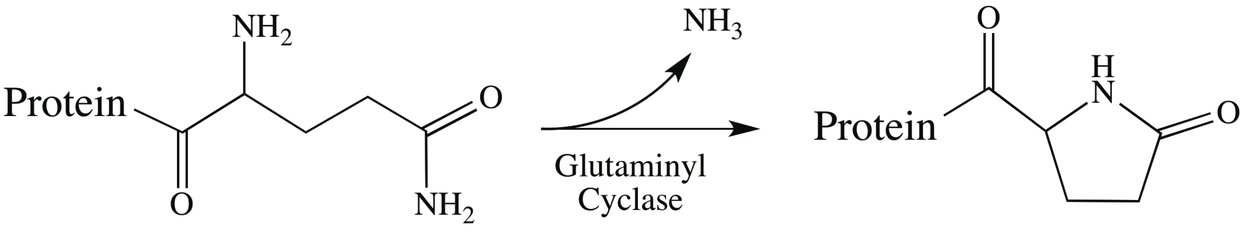


Figure 4-3

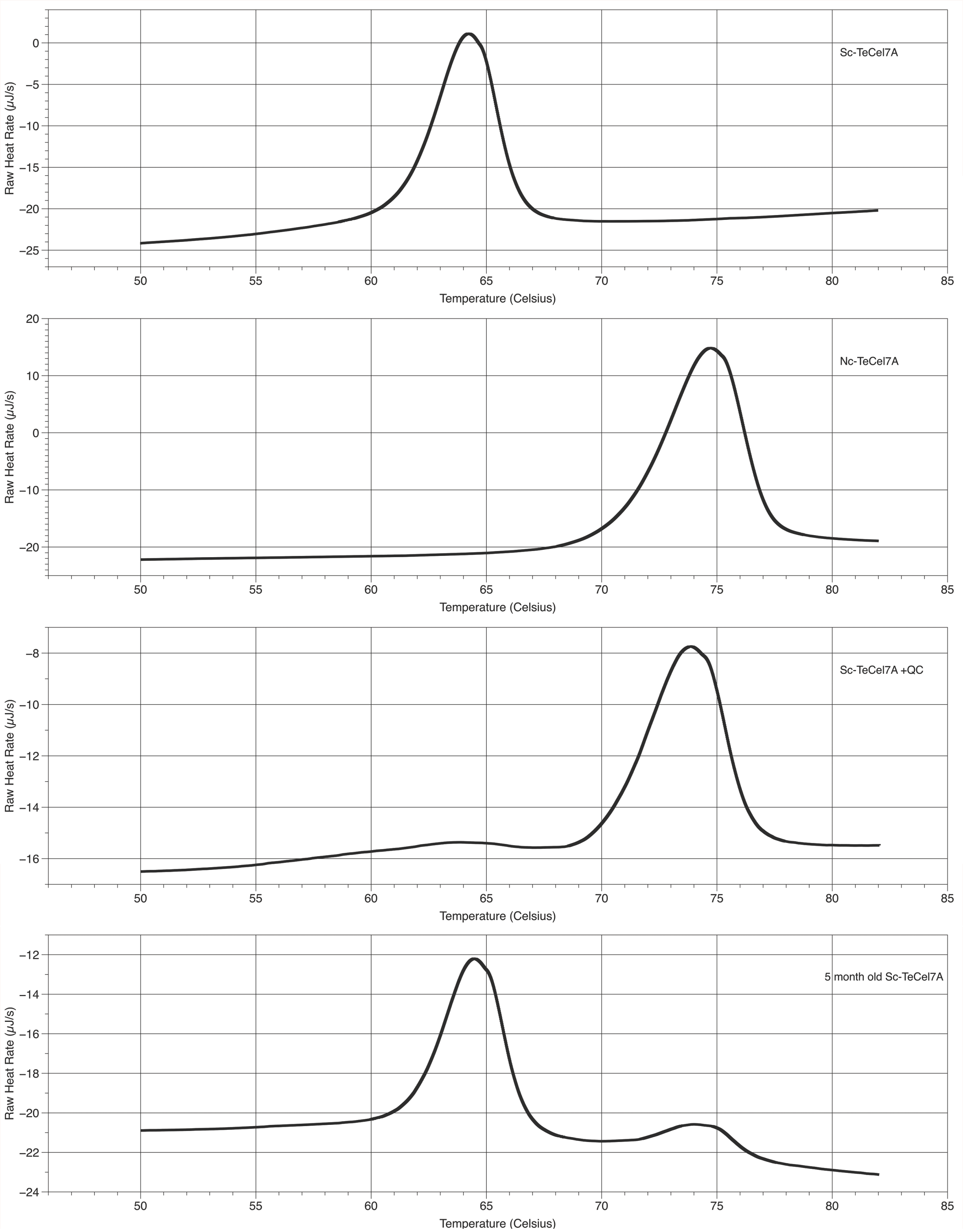

Sc-TeCel7A
Raw Heat Rate (μJ/s)
0
−5
−10
−15
−20
−25
50
55
60
65
70
75
80
85
Temperature (Celsius)
Nc-TeCel7A
Raw Heat Rate (μJ/s)
20
10
0
−10
−20
50
55
60
65
70
75
80
85
Temperature (Celsius)
Sc-TeCel7A +QC
Raw Heat Rate (μJ/s)
−8
−10
−12
−14
−16
50
55
60
65
70
75
80
85
Temperature (Celsius)
5 month old Sc-TeCel7A
Raw Heat Rate (μJ/s)
−12
−14
−16
−18
−20
−22
−24
50
55
60
65
70
75
80
85
Temperature (Celsius)

Figure 4-4

Figure 4-1. Activity of TeCel7A enzymes against Avicel. Reactions were carried out in duplicate by incubating TeCel7A (0.2 µM) with Avicel (10 g/L) in 100 uL for 15 hours in a thermocycler at various temperatures. The total glucose equivalents released were then measured two or three times for each duplicate reaction. Glucose equivalents released are reported as the average of these 4-6 values with error bars representing one standard deviation above and below the average. Circles: Sc-TeCel7A with his_6 tag. Triangles: Sc-TeCel7A with no his_6 tag. Dotted line with squares: Nc-TeCel7A with no his_6 tag. Stars: Sc-TeCel7A with no his_6 tag treated with glutaminyl cyclase. Dashed lines: deglycosylated by treatment with α1,2/α1,3 mannosidase and $EndoH_f$.

Figure 4-2. Crystal structure of TeCel7A catalytic domain. Left: The pyroglutamate, highlighted in orange, is secluded from the solvent in a tightly packed hydrophobic region. Surrounding hydrophobic leucine, phenylalanine, and two alanine residues are shown in red. Pyroglutamate is within 3 angstroms of phenylalanine and 2.5 angstroms of leucine. **Right:** Water, depicted here as red spheres, is scarce in the microenvironment of pyroglutamate. The presence of glutamate in its open and charged configuration would presumably disrupt the Van der Waal's and/or hydrophobic interactions that contribute to the stability of the fold. Created with VMD (Humphrey et al., 1996) and PDB ID 3PFJ.

Figure 4-3. N-terminal glutamine cyclization. Glutaminyl cyclase catalyzes the cyclization of the N-terminal glutamine residue, preventing the N-terminal amino group from becoming charged at pH 5.

Figure 4-4. Differential scanning calorimetry data measuring T_m values of TeCel7A. These data indicate that Sc-TeCel7A lacks the N-terminal glutamine modification, but slowly undergoes spontaneous cyclization. Note that the T_m of the glutaminyl cyclase-treated TeCel7A from *S. cerevisiae* is almost identical to that of the *N. crassa* TeCel7A. Top: Sc-TeCel7A with his_6 tag. Second: Nc-TeCel7A without his_6 tag. Third: Sc-TeCel7A with his_6 tag treated with glutaminyl cyclase. Fourth: Sc-TeCel7A with his_6 tag after long-term storage at 4°C.

References

Adney WS, Jeoh T, Beckham GT, Chou Y-C, Baker JO, Michener W, Brunecky R, Himmel ME. 2009. Probing the role of N-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis. Cellulose 16:699–709.

Antebi A, Fink GR. 1992. The yeast Ca(2+)-ATPase homologue, PMR1, is required for normal Golgi function and localizes in a novel Golgi-like distribution. Mol. Biol. Cell 3:633–654.

Baker JO, Tatsumoto K, Grohmann K, Woodward J, Wichert JM, Shoemaker SP, Himmel ME. 1992. Thermal denaturation ofTrichoderma reesei cellulases studied by differential scanning calorimetry and tryptophan fluorescence. Appl. Biochem. Biotechnol. 34-35:217–231.

Bardiya N, Shiu PKT. 2007. Cyclosporin A-resistance based gene placement system for Neurospora crassa. Fungal Genet. Biol. FG B 44:307–314.

Beckham GT, Dai Z, Matthews JF, Momany M, Payne CM, Adney WS, Baker SE, Himmel ME. 2012. Harnessing glycosylation to improve cellulase activity. Curr. Opin. Biotechnol. 23:338–345.

Bommarius AS, Blum JK, Abrahamson MJ. 2011. Status of protein engineering for biocatalysts: how to design an industrially useful biocatalyst. Curr. Opin. Chem. Biol. 15:194–200.

Bryksin AV, Matsumura I. 2010. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. BioTechniques 48:463–465.

Bu L, Beckham GT, Shirts MR, Nimlos MR, Adney WS, Himmel ME, Crowley MF. 2011. Probing carbohydrate product expulsion from a processive cellulase with multiple absolute binding free energy methods. J. Biol. Chem. 286:18161–18169.

Cherry J, Harris P, Jones A, Teter S, Ward C, Yi J. 2009. Variants of glycoside hydrolases. WO2005030926 A3.

Cirino PC, Mayer KM, Umeno D. Generating Mutant Libraries Using Error-Prone PCR. In: . Dir. Evol. Libr. Creat. New Jersey: Humana Press, Vol. 231, pp. 3–10. http://www.springerprotocols.com/Abstract/doi/10.1385/1-59259-395-X:3.

Dana CM, Saija P, Kal SM, Bryan MB, Blanch HW, Clark DS. 2012. Biased clique shuffling reveals stabilizing mutations in cellulase Cel7A. Biotechnol. Bioeng. 109:2710–2719.

Day AG, Goedegebuur F, Gualfetti P, Mitchinson C, Neefe P, Sandgren M, Shaw A, Stahlberg J. 2004. Novel variant hyprocrea jecorina cbh1 cellulases. WO2004016760 A2.

Dick LW, Kim C, Qiu D, Cheng K-C. 2007. Determination of the origin of the N-terminal pyro-glutamate variation in monoclonal antibodies using model peptides. Biotechnol. Bioeng. 97:544–553.

Gao L, Gao F, Wang L, Geng C, Chi L, Zhao J, Qu Y. 2012. N-Glycoform Diversity of Cellobiohydrolase I from Penicillium decumbens and Synergism of Nonhydrolytic Glycoform in Cellulose Degradation. J. Biol. Chem. 287:15906–15915.

Ghose TK. 1987. Measurement of cellulase activities. Pure Appl. Chem. 59:257–268.

Goedegebuur F, Gualfetti P, Mitchinson C, Neefe P. 2006. Novel cbh1 homologs and variant cbh1 cellulases. WO2005028636 A3.

Graham JE, Clark ME, Nadler DC, Huffer S, Chokhawala HA, Rowland SE, Blanch HW, Clark DS, Robb FT. 2011. Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment. Nat. Commun. 2:375.

Gueldener U, Heinisch J, Koehler GJ, Voss D, Hegemann JH. 2002. A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. Nucleic Acids Res. 30:e23.

Den Haan R, Mcbride JE, Grange DCL, Lynd LR, Van Zyl WH. 2007. Functional expression of cellobiohydrolases in Saccharomyces cerevisiae towards one-step conversion of cellulose to ethanol. Enzyme Microb. Technol. 40:1291–1299.

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. PEDS 23:871–880.

Hill K, Boone C, Goebl M, Puccia R, Sdicu AM, Bussey H. 1992. Yeast KRE2 defines a new gene family encoding probable secretory proteins, and is required for the correct N-glycosylation of proteins. Genetics 130:273–283.

Ilmén M, den Haan R, Brevnova E, McBride J, Wiswall E, Froehlich A, Koivula A, Voutilainen SP, Siika-Aho M, la Grange DC, Thorngren N, Ahlgren S, Mellon M, Deleault K, Rajgarhia V, van Zyl WH, Penttilä M. 2011. High level secretion of cellobiohydrolases by Saccharomyces cerevisiae. Biotechnol. Biofuels 4:30.

Jeoh T, Michener W, Himmel ME, Decker SR, Adney WS. 2008. Implications of cellobiohydrolase glycosylation for use in biomass conversion. Biotechnol. Biofuels 1:10.

Kim T-W, Chokhawala HA, Nadler DC, Nadler D, Blanch HW, Clark DS. 2010. Binding modules alter the activity of chimeric cellulases: Effects of biomass pretreatment and enzyme source. Biotechnol. Bioeng. 107:601–611.

Komor RS, Romero PA, Xie CB, Arnold FH. 2012. Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Eng. Des. Sel. PEDS 25:827–833.

Labbé S, Thiele DJ. 1999. Copper ion inducible and repressible promoter systems in yeast. Methods Enzymol. 306:145–153.

Lehmann M, Pasamontes L, Lassen SF, Wyss M. 2000. The consensus concept for thermostability engineering of proteins. Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol. 1543:408–415.

Luce RD, Perry AD. 1949. A method of matrix analysis of group structure. Psychometrika 14:95–116.

Lutz S, Patrick WM. 2004. Novel methods for directed evolution of enzymes: quality, not quantity. Curr. Opin. Biotechnol. 15:291–297.

Moore GL, Maranas CD, Gutshall KR, Brenchley JE. 2000. Modeling and optimization of DNA recombination. Comput. Chem. Eng. 24:693–699.

Nakayama K, Nagasu T, Shimma Y, Kuromitsu J, Jigami Y. 1992. OCH1 encodes a novel membrane bound mannosyltransferase: outer chain elongation of asparagine-linked oligosaccharides. EMBO J. 11:2511–2519.

Neylon C. 2004. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. Nucleic Acids Res. 32:1448–1459.

Östergård PRJ. 2002. A fast algorithm for the maximum clique problem. Discrete Appl. Math. 120:197–207.

Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH. 2006. Structure-guided recombination creates an artificial family of cytochromes P450. PLoS Biol. 4:e112.

Penttilä ME, André L, Lehtovaara P, Bailey M, Teeri TT, Knowles JK. 1988. Efficient secretion of two fungal cellobiohydrolases by Saccharomyces cerevisiae. Gene 63:103–112.

Peralta-Yahya P, Carter BT, Lin H, Tao H, Cornish VW. 2008. High-Throughput Selection for Cellulase Catalysts Using Chemical Complementation. J. Am. Chem. Soc. 130:17446–17452.

Phillips CM, Iavarone AT, Marletta MA. 2011. Quantitative proteomic approach for cellulose degradation by Neurospora crassa. J. Proteome Res. 10:4177–4185.

Qin Y, Wei X, Song X, Qu Y. 2008. Engineering endoglucanase II from Trichoderma reesei to improve the catalytic efficiency at a higher pH optimum. J. Biotechnol. 135:190–195.

Rakestraw JA, Sazinsky SL, Piatesi A, Antipov E, Wittrup KD. 2009. Directed evolution of a secretory leader for the improved expression of heterologous proteins and full-length antibodies in Saccharomyces cerevisiae. Biotechnol. Bioeng. 103:1192–1201.

Rasila TS, Pajunen MI, Savilahti H. 2009. Critical evaluation of random mutagenesis by error-prone polymerase chain reaction protocols, Escherichia coli mutator strain, and hydroxylamine treatment. Anal. Biochem. 388:71–80.

Robinson AS, Hines V, Wittrup KD. 1994. Protein disulfide isomerase overexpression increases secretion of foreign proteins in Saccharomyces cerevisiae. Biotechnol. Nat. Publ. Co. 12:381–384.

Romero PA, Arnold FH. 2009. Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. 10:866–876.

Ruttersmith LD, Daniel RM. 1991. Thermostable cellobiohydrolase from the thermophilic eubacterium Thermotoga sp. strain FjSS3-B.1. Purification and properties. Biochem. J. 277 (Pt 3):887–890.

Schilling S, Wasternack C, Demuth H-U. 2008a. Glutaminyl cyclases from animals and plants: a case of functionally convergent protein evolution. Biol. Chem. 389:983–991.

Schilling S, Zeitschel U, Hoffmann T, Heiser U, Francke M, Kehlen A, Holzer M, Hutter-Paier B, Prokesch M, Windisch M, Jagla W, Schlenzig D, Lindner C, Rudolph T, Reuter G, Cynis H, Montag D, Demuth H-U, Rossner S. 2008b. Glutaminyl cyclase inhibition attenuates pyroglutamate Abeta and Alzheimer's disease-like pathology. Nat. Med. 14:1106–1111.

Smith MA, Bedbrook CN, Wu T, Arnold FH. 2013. Hypocrea jecorina Cellobiohydrolase I Stabilizing Mutations Identified Using Noncontiguous Recombination. ACS Synth. Biol.

Stals I, Sandra K, Geysens S, Contreras R, Beeumen JV, Claeyssens M. 2004. Factors influencing glycosylation of Trichoderma reesei cellulases. I: Postsecretorial changes of the O- and N-glycosylation pattern of Cel7A. Glycobiology 14:713–724.

Stemmer WP. 1994. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc. Natl. Acad. Sci. 91:10747–10751.

Suominen PL, Mäntylä AL, Karhunen T, Hakola S, Nevalainen H. 1993. High frequency one-step gene replacement in Trichoderma reesei. II. Effects of deletions of individual cellulase genes. Mol. Gen. Genet. MGG 241:523–530.

Taylor CB, Talib MF, McCabe C, Bu L, Adney WS, Himmel ME, Crowley MF, Beckham GT. 2012. Computational Investigation of Glycosylation Effects on a Family 1 Carbohydrate-binding Module. J. Biol. Chem. 287:3147–3155.

Tian C, Beeson WT, Iavarone AT, Sun J, Marletta MA, Cate JHD, Glass NL. 2009. Systems analysis of plant cell wall degradation by the model filamentous fungus Neurospora crassa. Proc. Natl. Acad. Sci.pnas.0906810106.

Voutilainen SP, Murray PG, Tuohy MG, Koivula A. 2010. Expression of Talaromyces emersonii cellobiohydrolase Cel7A in Saccharomyces cerevisiae and rational mutagenesis to improve its thermostability and activity. Protein Eng. Des. Sel. PEDS 23:69–79.

Voutilainen SP, Nurmi-Rantala S, Penttilä M, Koivula A. 2013. Engineering chimeric thermostable GH7 cellobiohydrolases in Saccharomyces cerevisiae. Appl. Microbiol. Biotechnol.

Voutilainen SP, Puranen T, Siika-Aho M, Lappalainen A, Alapuranen M, Kallio J, Hooman S, Viikari L, Vehmaanperä J, Koivula A. 2008. Cloning, expression, and characterization of novel thermostable family 7 cellobiohydrolases. Biotechnol. Bioeng. 101:515–528.

Wong TS, Roccatano D, Zacharias M, Schwaneberg U. 2006. A statistical analysis of random mutagenesis methods used for directed protein evolution. J. Mol. Biol. 355:858–871.

Yip CL, Welch SK, Klebl F, Gilbert T, Seidel P, Grant FJ, O'Hara PJ, MacKay VL. 1994. Cloning and analysis of the Saccharomyces cerevisiae MNN9 and MNN1 genes required for complex glycosylation of secreted proteins. Proc. Natl. Acad. Sci. U. S. A. 91:2723–2727.