## Title
Dissecting enzyme function with microfluidic-based deep mutational scanning



## Journal
Proceedings of the National Academy of Sciences of the United States of America, 112(23)

## ISSN
0027-8424

## Authors
Romero, Philip A
Tran, Tuan M
Abate, Adam R

## Publication Date
2015-06-09

## DOI
10.1073/pnas.1422285112

Peer reviewed

# Dissecting enzyme function with microfluidic-based deep mutational scanning

Philip A. Romero, Tuan M. Tran, and Adam R. Abate[1]

Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, CA 94158

Edited by David Baker, University of Washington, Seattle, WA, and approved May 6, 2015 (received for review November 21, 2014)

**Natural enzymes are incredibly proficient catalysts, but engineering them to have new or improved functions is challenging due to the complexity of how an enzyme's sequence relates to its biochemical properties. Here, we present an ultrahigh-throughput method for mapping enzyme sequence–function relationships that combines droplet microfluidic screening with next-generation DNA sequencing. We apply our method to map the activity of millions of glycosidase sequence variants. Microfluidic-based deep mutational scanning provides a comprehensive and unbiased view of the enzyme function landscape. The mapping displays expected patterns of mutational tolerance and a strong correspondence to sequence variation within the enzyme family, but also reveals previously unreported sites that are crucial for glycosidase function. We modified the screening protocol to include a high-temperature incubation step, and the resulting thermotolerance landscape allowed the discovery of mutations that enhance enzyme thermostability. Droplet microfluidics provides a general platform for enzyme screening that, when combined with DNA-sequencing technologies, enables high-throughput mapping of enzyme sequence space.**

protein engineering | droplet-based microfluidics | high-throughput DNA sequencing

Enzymes are powerful biological catalysts capable of remarkably accelerating the rates of chemical transformations (1). The molecular bases of these rate accelerations are often complex, using multiple steps, multiple catalytic mechanisms, and relying on numerous molecular interactions, in addition to those provided by the main catalytic groups. This complexity imposes a significant barrier to understanding how an enzyme's sequence impacts its function and, thus, on our ability to rationally design biocatalysts with new or enhanced functions (2–4).

Comprehensive mappings of sequence–function relationships can be used to dissect the molecular basis of protein function in an unbiased manner (5). Growth selections or in vitro binding screens can be combined with next-generation DNA sequencing to generate detailed mappings between a protein's sequence and its biochemical properties, such as binding affinity, enzymatic activity, and stability (6–9). This deep mutational scanning approach has been used to study the structure of the protein fitness landscape, discover new functional sites, improve molecular energy functions, and identify beneficial combinations of mutations for protein engineering. However, these methods rely on functional assays coupled to cell growth or protein binding, severely limiting the types of proteins that can be analyzed. For example, most enzymes of biological or industrial relevance cannot be analyzed using existing methods because they do not catalyze a reaction that can be directly coupled to cell growth. Experimental advances are needed to broaden the applicability of deep mutational scanning to the diverse palette of functions performed by enzymes.

In this paper, we present a general method for mapping protein sequence–function relationships that greatly expands the scope of biochemical functions that can be analyzed. Ultrahigh-throughput droplet-based microfluidic screening enables us to characterize the chemical activities of millions of enzyme variants. By sorting the variants based on chemical activity and performing next-generation DNA sequencing of sorted and unsorted libraries, we obtain a detailed mapping of how changes to enzyme sequence impact chemical function. We demonstrate this method using a glycosidase enzyme important in the deconstruction of biomass into fermentable sugars for biofuel production. Through comprehensive mutagenesis and functional characterization of this enzyme, we were able, with minimal bias, to discover residues crucial to function and identify mutations that enhance its activity at elevated temperatures. This approach can be applied to any enzyme whose chemical activity can be measured with a fluorogenic assay in microfluidic droplets (10–13). Our method extends the applicability of deep mutational scanning to a wide range of protein functions and reaction conditions not accessible by other high-throughput methods.

## Results

**High-Throughput Sequence–Function Mapping.** Protein sequence space is vast and an enzyme's functional properties may depend on hundreds to thousands of molecular interactions, most of which will have never been characterized. Systematically exploring this space thus necessitates methods capable of characterizing massive numbers of sequence variants. We have developed a general method for performing millions of sequence–function measurements on an enzyme (Fig. 1*A*). A library of enzyme variants is expressed in *Escherichia coli*, and single cells are encapsulated in microfluidic droplets containing lysis reagents and a fluorogenic enzyme substrate (Fig. S1*A*). Upon lysis, the expressed enzyme variant is released into the droplet, allowing it to interact with the substrate. The surrounding oil acts as a barrier that keeps reagents contained within the droplets, preventing product molecules

### Significance

As powerful biological catalysts, enzymes can solve challenging problems that range from the industrial production of chemicals to the treatment of human disease. The ability to design new enzymes with tailor-made chemical functions would have a far-reaching impact. However, this important capability has been limited by our cursory understanding of enzyme catalysis. Here, we report a method that uses unbiased empirical analysis to dissect the molecular basis of enzyme function. By comprehensively mapping how changes in an enzyme's amino acid sequence affect its activity, we obtain a detailed view of the interactions that shape the enzyme function landscape. Large, unbiased analyses of enzyme function allow the discovery of new biochemical mechanisms that will improve our ability to engineer custom biocatalysts.

Author contributions: P.A.R. and A.R.A. designed research; P.A.R. performed research; T.M.T. contributed new reagents/analytic tools; P.A.R. and A.R.A. analyzed data; and P.A.R. and A.R.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: adam.abate@ucsf.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1422285112/-/DCSupplemental.

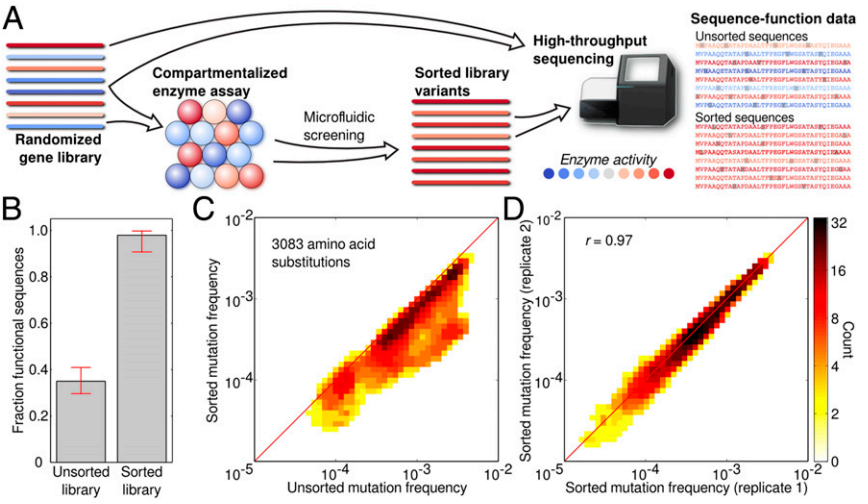



**Fig. 1.** High-throughput sequence–function mapping. (*A*) A conceptual overview of the sequence–function mapping protocol. Individual members of a randomized gene library are assayed in aqueous microdroplets, and microfluidic screening is used to sort out the active variants. The unsorted and sorted variant pools are then analyzed using high-throughput DNA sequencing. The resulting sequence–function dataset is used to understand the functional impact of mutations. (*B*) Droplet-based microfluidic screening recovers functional sequences from the initial random mutagenesis library. Individual clones from the unsorted and sorted libraries were tested in a plate-based assay and were considered functional if their end-point activity was greater than 50% of Bgl3's. Initially, only 35% of the library was functional, but after screening the fraction of functional sequences increased to 98%. Error bars represent the 90% binomial proportion confidence interval. (*C*) The frequency of 3,083 amino acid substitutions in the unsorted and sorted libraries. A large fraction of mutations decrease in frequency after sorting, suggesting they are deleterious to Bgl3 function. (*D*) Reproducibility of the sequence–function mapping protocol. Two independent experimental replicates show very good agreement in amino acid frequencies.

generated by one variant from mixing with those of another in a different droplet. Droplets that contain efficient variants thus rapidly accumulate fluorescent product, whereas those with inactive variants remain dim. The DNA sequences of the active variants are then recovered using a high-throughput microfluidic sorter to recover the bright droplets (14). The sorter can analyze more than 100 enzyme variants per second, reaching 1 million in just a few hours. The sorted and unsorted gene libraries are then processed using next-generation DNA sequencing and statistical analysis.

As a demonstration of the generality and power of our sequence–function mapping method, we used it to analyze Bgl3, a β-glucosidase enzyme from *Streptomyces* sp. We chose Bgl3 because it catalyzes an important step in the deconstruction of biomass into fermentable sugars, it is a remarkably proficient catalyst ($k_{cat}/k_{uncat} \sim 10^{16}$), its structure has been solved to high resolution, and it has a simple fluorogenic assay. To enable accurate sorting of active from inactive variants, we developed an emulsion-based β-glucosidase assay that showed excellent discrimination between wild-type (WT) Bgl3 and an inactive mutant (Fig. S1 *B–D*). We used error-prone PCR to generate a Bgl3 mutant library with an average of 3.8 amino acid substitutions per gene. We screened this library for a total of 23 h (four separate runs), analyzing over 10 million variants, 3.4 million of which contained measurable enzymatic activity and were recovered via microfluidic sorting (Fig. S1*E*). To confirm enrichment of functional sequences within the sorted population, we tested a random sampling of mutants in a plate assay before and after sorting (Fig. 1*B*). Before sorting, ~35% of variants were found to be functional, the remainder inactive due, presumably, to deleterious point mutations. After sorting, the fraction of functional sequences increased to 98%. The sorted sequences had an average of 2.0 amino acid substitutions per gene, approximately one-half that of the unsorted library.

We processed the unsorted and sorted gene libraries using the Nextera XT sequencing library prep kit, sequenced using an Illumina MiSeq, version 3, 2 × 300 run, and mapped the sequence reads to the *bgl3* gene using Bowtie2. The DNA sequencing showed good coverage across the entire *bgl3* gene for both the unsorted and sorted libraries (Fig. S2*A*). The Bgl3 construct has 500 amino acid positions and therefore a total of 10,000 (500 × 20) possible amino acid substitutions including nonsense mutations. After applying sequencing quality filters, there were sufficient statistics to quantify the frequency of 3,083 (31%) of these amino acid substitutions. The remaining 6,917 substitutions were difficult to access because they require two or three nucleotide mutations within a single codon, which is a rare occurrence in libraries generated via error-prone PCR (Fig. S2*B*).

The effect of an amino acid substitution can be estimated by how much its frequency changes in response to functional screening. A majority of mutations decreased in frequency in the sorted library, suggesting they are deleterious to the enzyme's function (Fig. 1*C*). This observation is consistent with other studies analyzing the effects of random mutations on protein function (15–18). To further evaluate the method, we tested the reproducibility of the mapping by comparing amino acid frequencies from two independent sorting experiments (Fig. 1*D*). These datasets show excellent agreement ($r = 0.97$) across all 3,083 point mutations. Our microfluidic sequence–function mapping method was further validated on a panel of Bgl3 variants with known enzyme activities (Fig. S3).

**Site-Specific Mutational Tolerance.** Data from millions of functional sequence variants can be used to identify residues important for enzyme function. Residues that cannot be mutated to other amino acids are likely to play a specific role required for enzyme activity. The degree to which a site can tolerate amino acid change is thus an indicator of its functional importance. The relative entropy (RE) can be used to score a residue's mutational tolerance, because it quantifies how much the amino acid probability distribution changes between the unsorted and sorted libraries (Fig. 2*A*). A site whose distribution shifts significantly from random has high relative entropy, implying that a specific amino acid must reside at that position for the enzyme to remain functional.

The mutational tolerance of a site should be related to its position in the protein's 3D structure, because this determines the other residues with which it interacts. To investigate the relationship between enzyme structure and mutational tolerance, we mapped the relative entropy of each position onto the Bgl3 crystal structure (Fig. 2*B*). As expected, the catalytic nucleophile (E383) and general acid/base (E178) are both highly intolerant to mutation, falling at the 99th and 95th percentiles, respectively. We also expect core residues to be less tolerant to mutation than surface residues because the protein core tends to be well packed, forming many interresidue interactions. To support this, the α-helices that compose the TIM-barrel wall display an alternating pattern, where the interior helix face is less tolerant to mutation than the exterior face (Fig. 2*B*). Overall, buried residues are less tolerant to mutation than solvent-exposed residues (Fig. 2*C*).

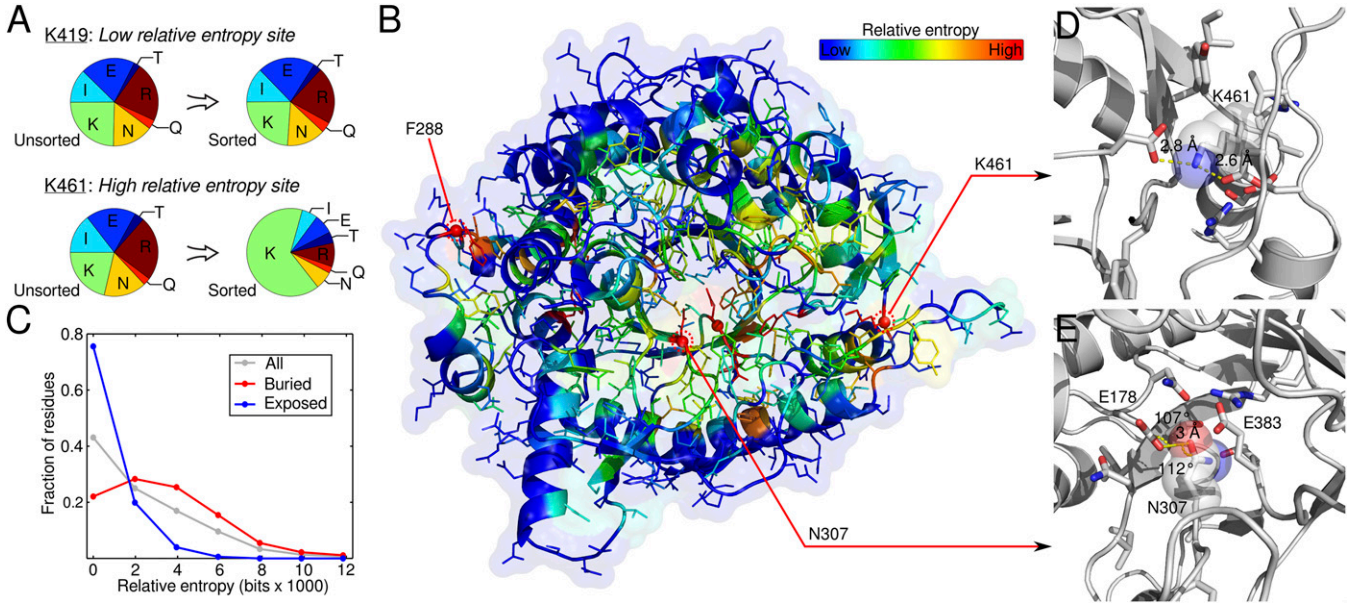



**Fig. 2.** Analysis of site-specific mutational tolerance. (*A*) Relative entropy (RE) describes how much the amino acid probability distribution changes in response to functional screening. The amino acid distribution of mutated codons is shown for a low RE site and a high RE site. Only synonymous substitutions are shown for the WT amino acid. The low RE site (K419) shows little change between the unsorted and sorted libraries, suggesting this position can tolerate substitutions to other amino acids. In contrast, the high RE site (K461) shows a strong shift back to the WT residue. (*B*) Structural patterns of mutational tolerance. The relative entropy of each site was mapped onto the Bgl3 crystal structure (Protein Data Bank ID code 1GNX). Sites with the highest relative entropies (≥99th percentile) have a red sphere at their α carbon. As expected, known functional sites, such as the catalytic residues, are highly intolerant to mutation. The analysis also reveals previously unannotated positions that are intolerant to mutation and may therefore play an important role in Bgl3 function. Three of these sites (F288, N307, and K461) are labeled in the figure. (*C*) The mutational tolerance of a position depends on its solvent exposure. The distribution of relative entropies for all positions is shown in gray. Buried residues [relative surface area (RSA) < 0.2] tend to have higher relative entropies and are therefore less tolerant to mutations than solvent-exposed residues (RSA ≥ 0.2). (*D*) Detailed view of K461 in Bgl3 structure. K461 (transparent spheres) forms salt bridges with two nearby aspartic acid residues. The short interatomic distances and their networked nature, suggests these interactions are strong and may be important for the structural stability of the enzyme. (*E*) Detailed view of N307 in Bgl3 structure. N307 (transparent spheres) is located directly between the enzyme's nucleophile (E383) and the general acid/base (E178). Based on the distance and angles of the residues, N307 appears to hydrogen bond with E178, which may be important for perturbing the $pK_a$ of that group and, thus, the catalytic mechanism of the enzyme.

The analysis of mutational tolerance reveals sites that play an important functional role, several of which have never been described in the literature. For example, lysine 461 has the highest relative entropy of any residue (100th percentile), although, oddly, it is far from the active site (Fig. 2*B*). Targeted mutagenesis shows no other amino acid can be accepted at this location, validating the mutational tolerance findings (Fig. S4*C*). In the crystal structure, K461 is involved in networked salt bridges with two aspartic acid residues (Fig. 2*D*). The short distance of these interactions indicates they are strong and suggests that K461 may be important for the structural stability of the enzyme. Indeed, substitutions at this position significantly decrease the enzyme's soluble expression (Fig. S4*C*).

Asparagine 307 is another residue with high relative entropy (99th percentile) that, again, has not been described previously. N307 is located in the enzyme's active site and appears to be hydrogen bonding with the general acid/base E178 in the crystal structure (Fig. 2*E*). Targeted mutagenesis at this position also shows no other amino acid is tolerated, again validating the results of the mutational tolerance map obtained with our approach (Fig. S4*B*). Unlike K461, substitutions at N307 demolish enzyme activity but have minimal influence on soluble expression, suggesting N307's role in the enzyme's catalytic mechanism. We hypothesize that N307 may act to shift the $pK_a$ of the general acid/base, which is crucial for the $pK_a$-cycling mechanism of most retaining glycosidases (19). These results demonstrate the power of comprehensive and unbiased sequence–function mapping for investigating enzyme function and identifying important residues.

**Comparison with the Natural Sequence Record.** Bgl3 is a member of glycoside hydrolase family 1 (GH1), a large enzyme family accepting a broad range of glycosylated substrates (20, 21). The sequences within the GH1 family typically differ by hundreds of mutations, providing a diverse sampling of the sequence space explored by natural evolution. By contrast, our experimental sequence–function mapping densely samples the local space of sequences within a few mutations of Bgl3. Comparing the global versus local view of sequence space may provide insight into the evolutionary constraints imposed on members of the GH1 family.

To investigate how our results compare with the natural sequence record, we used a large GH1 multiple sequence alignment to calculate a relative entropy sequence conservation score (22, 23). Bgl3's mutational tolerance shows a strong correspondence with the observed GH1 sequence conservation. Gene-scale patterns can be visualized by taking a moving average (five-site window) of the relative entropy and sequence conservation scores across sequence positions (Fig. 3*A*). The experimental mutational tolerance and GH1 conservation are strikingly similar, and their patterns tend to correspond with secondary structure elements. Overall, the experimental relative entropy and the sequence conservation score display a strong, statistically significant correlation ($r = 0.59$, $P < 1E\text{-}45$; Fig. 3*B*), suggesting that most sites important for Bgl3 function are also important throughout the GH1 family.

There are, however, unexpected and interesting exceptions to the correspondence between Bgl3's mutational tolerance and GH1 sequence conservation. The most extreme is position 288, which is highly intolerant to mutation in Bgl3 (99th percentile for RE) but has little conservation in the GH1 alignment (11th

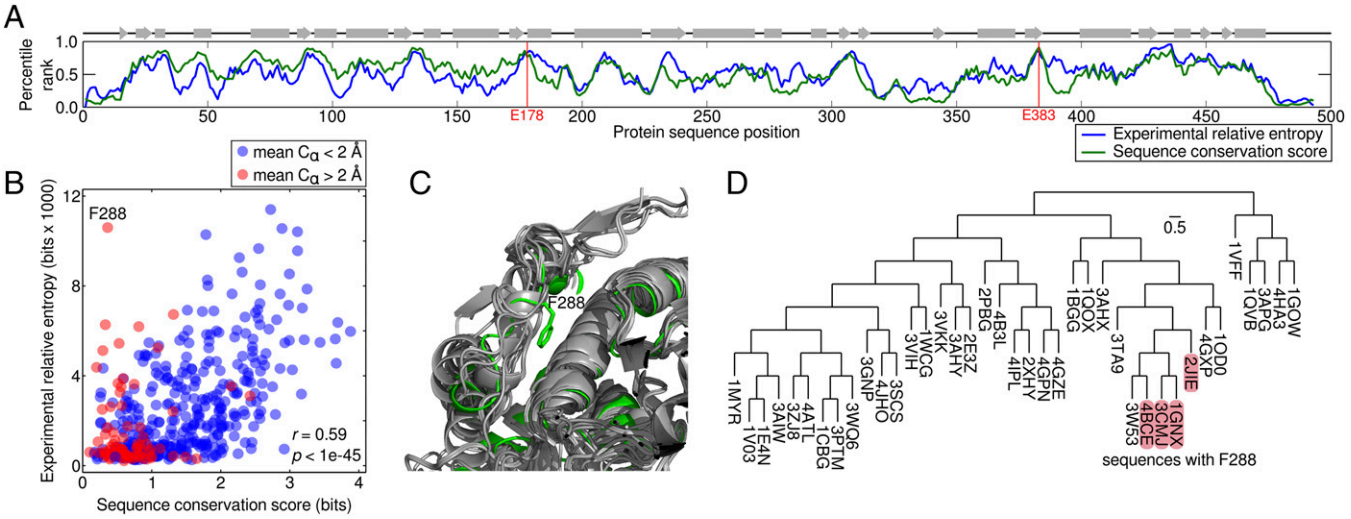



**Fig. 3.** Comparison with natural sequence variation. (*A*) Large-scale patterns of Bgl3's mutational tolerance and the observed GH1 sequence conservation. A moving average (five-site window) of the experimental relative entropy and sequence conservation scores is plotted over sequence positions. Percentile ranks are used to plot the two scores on the same axis. The overall patterns of Bgl3 mutational tolerance and GH1 conservation are very similar and tend to correspond with secondary structure elements (displayed across the top). (*B*) The relationship between a site's mutational tolerance and sequence conservation. A scatter plot of the experimental relative entropy and sequence conservation scores displays a strong correlation ($r = 0.59$; $P < 1E\text{-}45$), indicating that sites important for Bgl3 function are also important throughout the GH1 family. Outlying sites, such as F288, can be explained by structural diversity within the enzyme family. Structural diversity (mean $C_\alpha$ displacement) was quantified by aligning all related structures to Bgl3, calculating each structure's $C_\alpha$ displacement from Bgl3 at each position, and averaging over all structures. Positions with a high experimental relative entropy, but low sequence conservation score (top, left corner) tend to come from regions with more structural diversity (red points). (*C*) Structural diversity may explain outlying sites. Position 288 is highly intolerant to mutation in Bgl3 (99th percentile for RE) but has little conservation in the GH1 alignment (11th percentile for sequence conservation score). An alignment of GH1 structures reveals that position 288 occurs in a structurally diverse loop. We hypothesize that F288 is important for Bgl3 function, but its interactions are not conserved throughout the GH1 family. (*D*) Sequence–function mapping provides a local view of sequence space. A phylogenetic tree of GH1 structures shows the few sequences that do contain F288 are closely related.

percentile for sequence conservation). Targeted mutagenesis at this location again validates the sequence–function mapping results, confirming that Bgl3 can only tolerate 21% of all amino acid substitutions at position 288 (Fig. S4*A*). The fact that other GH1 members can accept most amino acids at position 288 suggests that Bgl3 evolution may be constrained by mutational epistasis at this site.

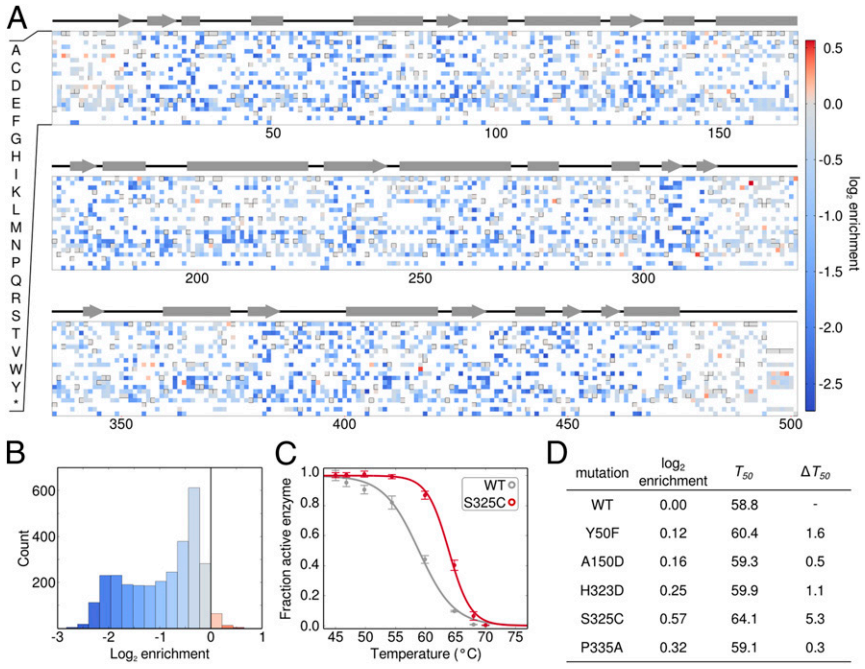
A closer look at GH1 structures reveals that position 288 occurs within a loop region displaying high diversity in the family (Fig. 3*C*). In fact, the most outlying positions (high experimental RE and low sequence conservation) occur in regions with high structural variation within the GH1 family (Fig. 3*B*, red points). We hypothesize that, through the course of natural evolution, Bgl3 may have evolved unique structural motifs that constrain its mutational tolerance relative to the GH1 family. We expect closely related sequences to also share these motifs and therefore to have similar residue preferences. Indeed, the phylogenetic tree of GH1 structures shows the few members that do contain F288 are closely related (Fig. 3*D*). Similar mutational idiosyncrasies may exist in all family members, but their conservation patterns become obscured when observing the entire family alignment.

These results highlight how sequence–function mapping provides a detailed local view of sequence space, whereas large multiple-sequence alignments provide a global perspective. A local sequence space mapping is important for applications such as protein engineering or the prediction of disease-associated mutations, because they focus on the mutational properties of the specific family member under investigation.

**High-Temperature Screening Enriches for Stabilizing Mutations.** Previous work in enzyme sequence–function mapping has used in vivo assays that couple an enzyme's function to cellular growth (7, 24–26). These in vivo selections are limited not only in the types of enzyme functions that can be analyzed, but also by the range of experimental conditions compatible with the intracellular environment. An advantage of droplet-based microfluidics is the ability to precisely control screening conditions, such as time, temperature, and concentration. Screening under altered conditions allows for enrichment of variants with enhanced unnatural properties.

To investigate this capability, we modified the microfluidic screening protocol to include a heat challenge directly after droplet formation (Fig. S5). We hypothesized that this should enrich for mutations that increase Bgl3's thermostability. We screened a total of 10 million enzyme variants, 2 million (20%) of which were determined to remain active and recovered via sorting. In this experiment, the heat challenge inactivated approximately one-half of the variants active in the original room temperature screen.

To observe the effects of the heat challenge on the functional space of enzyme sequences, we plotted the enrichment value for every observed amino acid substitution along the length of the enzyme (Fig. 4*A*). Overall, most mutations (97%) decreased in frequency (blue), but a small number showed positive enrichment values (red, Fig. 4*B*). The mutation with the greatest enrichment was S325C, located in an unresolved loop of the Bgl3 structure. This mutant was constructed and characterized and, indeed, found to yield a 5.3 °C increase in thermostability (Fig. 4*C*). We believe S325C is involved in a disulfide bond because performing the thermostability measurements in the presence of the reducing agent DTT abolishes the stability enhancement (Fig. S7). Identifying single mutations with such dramatic stability improvements is very difficult using other protein engineering methods. Other substitutions with positive enrichment values also increase the enzyme's thermostability (Fig. 4*D* and Fig. S8). This simple protocol allows the identification of thermostabilizing mutations and can be adapted to enrich for a variety of additional properties by screening under different conditions.




| mutation | $\log_2$ enrichment | $T_{50}$ | $\Delta T_{50}$ |
|---|---|---|---|
| WT | 0.00 | 58.8 | - |
| Y50F | 0.12 | 60.4 | 1.6 |
| A150D | 0.16 | 59.3 | 0.5 |
| H323D | 0.25 | 59.9 | 1.1 |
| S325C | 0.57 | 64.1 | 5.3 |
| P335A | 0.32 | 59.1 | 0.3 |

**Fig. 4.** Identification of stabilizing point mutations. (*A*) High-temperature screening enriches for stabilizing mutations. The enrichment value of 2,956 amino acid substitutions plotted over sequence positions. Amino acids that were not observed are colored as white and the WT residue is colored gray with a box around it. (*B*) The overall distribution of enrichment values. Only 3% of substitutions have a positive enrichment value. (*C*) Thermal inactivation curves for WT Bgl3 and the mutant with the highest enrichment value. S325C increases the $T_{50}$ of the enzyme by 5.3 °C. (*D*) Enriched mutations confer enhanced thermostability. A panel of five mutations was chosen based on their enrichment value and its reproducibility over experimental replicates. The enriched mutations were experimentally characterized, and all showed moderate-to-large increases in thermostability. The magnitudes of the stability increases depend on the assay conditions and tend to be lower when tested under conditions different from the screen (Fig. S6).

## Discussion

Deep mutational scanning is a powerful tool for exploring the molecular basis of protein function (7, 15, 25, 26). However, restrictions on functional assays have limited its general applicability, particularly for enzymes. We have presented a method for characterizing millions of enzyme variants by compartmentalizing reactions in aqueous microdroplets. The assays use an optical readout and can therefore be readily adapted to the numerous classes of enzymes with fluorescence-based activity assays.

Our experimental protocol enabled the analysis of over 1 million Bgl3 variants, and we used the resulting sequence–function map to evaluate the enzyme's tolerance to mutation. This unbiased analysis discovered sites within the enzyme that cannot tolerate mutations and are therefore likely to play an important role in Bgl3 function. Alternately, sites with a high tolerance to mutation are important for protein evolution and engineering because they can accept diversification while still maintaining catalytic function; this provides the protein engineer with flexibility in enhancing certain properties while maintaining others. The sequence–function mapping approach provides a local view of protein sequence space that can identify important interactions overlooked by large alignments of homologous sequences.

Droplet-based microfluidic screening provides a flexible platform for assaying enzyme activity over a broad range of reaction conditions (10–13). We adapted our screening protocol to include a heat challenge and enriched for mutations that increase the enzyme's thermostability. An alternative approach for identifying stabilizing mutations from high-throughput sequence–function data was recently developed that involved scoring a residue's ability to rescue the deleterious effects of other mutations (27). However, the droplet-based screening approach is extremely versatile and could theoretically be used to identify variants with enhanced properties including increased $k_{cat}$ (reduced reaction time), decreased $K_m$ (reduced substrate concentration), increased tolerance to biomass pretreatments (increased ionic liquid concentration), and reduced product inhibition (increased glucose concentration). Systematically mapping multiple enzyme properties will allow us to evaluate the trade-offs between properties and enable multiobjective protein engineering.

Experimentally mapping protein sequence space requires high-throughput library synthesis, screening, and sequencing, any of which could be a bottleneck. From this work, we found library construction and sequencing to be more limiting than microfluidic screening. Our random mutagenesis library contained 6 million unique variants (colony-forming units), and the transformation efficiency limited the size of this library. The microfluidic sorter analyzed over 10 million enzyme variants in 23 h, and the throughput of more recent sorter designs is more than an order of magnitude faster (28)—enabling the screening of libraries beyond $10^8$ variants. Although Illumina DNA sequencers can provide a large number of sequencing reads, read length is currently limited to ~600 bp, about one-third of the *bgl3* gene. A number of new methods to generate longer read lengths have recently been developed (29, 30) and would allow a pairwise analysis by correlating the effects of mutations at distant sequence positions.

Our method relies on a microfluidic droplet sorter that requires specialized instrumentation not typically found in a biochemistry laboratory. However, an alternative to screening enzyme variants in water-in-oil droplets is to screen using water-in-oil-in-water double emulsions (31). Double-emulsion droplets also provide microcompartments with which to test individual enzyme variants but can be generated using commercially available microfluidic systems (Dolomite Microfluidics) and sorted using standard cell sorters (32). This should provide an easily adoptable and widely available solution for implementing our sequence–function mapping method.

Our method could potentially be applied to a large number of different enzyme classes. In addition to glycosidases, emulsion-based methods have been used to screen DNA/RNA polymerases, oxidoreductases, sulfatases, peroxidases, esterases, proteases, and even ribozymes (10, 11, 33–37). The greatest challenge with emulsion-based screening is finding a fluorescent assay for one's particular enzyme of interest. It is important to note that some small-molecule dyes readily exchange between emulsion droplets and limit the ability to resolve functional differences (38).

The ability to rationally engineer enzymes will have a far-reaching impact on areas that range from medicine and agriculture to environmental protection and industrial chemistry. However, enzyme function involves an extraordinarily complex balance of numerous physical interactions, which has limited the design of

tailor-made enzymes. Large sequence–function datasets will provide an increasingly detailed view of the determinants of enzyme function. When combined with methods from statistics and machine learning, protein design rules can be extracted and applied in an automated manner (39). Given the rapid pace of advances in high-throughput experimentation, data-driven protein engineering may be able to outpace more traditional physics-based methods.

## Materials and Methods

All microfluidic devices were fabricated in-house using standard soft lithography techniques (Fig. S9). Photomasks were used to pattern layers of photoresist (SU-8 3025) on a silicon wafer, and uncured polydimethylsiloxane (PDMS) (11:1 polymer–to–cross-linker ratio) was poured over the mold. The PDMS was cured at 80 °C for 1 h, extracted from the mold with a scalpel, and access holes were punched using a 0.75-mm biopsy core. The devices were then bonded to glass slides after a plasma surface treatment. The device channels were made hydrophobic by flushing with Aquapel (Pittsburgh Glass Works) and then baking for an additional 10 min at 80 °C. Microfluidic fluorescence measurements were performed using a custom-built fluorimeter (Fig. S10).

**ACKNOWLEDGMENTS.** We thank R. A. Heins for providing the *bgl3* gene and useful feedback. We acknowledge J. Fraser, P. Babbit, and T. Kortemme for helpful discussions and feedback on the manuscript. P.A.R. is supported by the National Institute of General Medical Sciences of the NIH under Award F32GM107107, the University of California President's Postdoctoral Fellowship Program, and the Burroughs Wellcome Fund Postdoctoral Enrichment Program. T.M.T. is supported by the National Science Foundation Graduate Research Fellowship under Grant 1144247. This work was funded by a National Science Foundation CAREER Award (DBI-1253293), the NIH New Innovator Award (AR068129-01) and an R21 (HG007233-01), the Defense Advanced Research Projects Agency Living Foundries Program (HR0011-12-C-0065), a Research Award from the California Institute for Quantitative Biosciences, and the Bridging the Gap Award from the Rogers Family Foundation.

1. Wolfenden R, Snider MJ (2001) The depth of chemical time and the power of enzymes as catalysts. *Acc Chem Res* 34(12):938–945.
2. Baker D (2010) An exciting but challenging road ahead for computational enzyme design. *Protein Sci* 19(10):1817–1819.
3. Lassila JK, Baker D, Herschlag D (2010) Origins of catalysis by computationally designed retroaldolase enzymes. *Proc Natl Acad Sci USA* 107(11):4937–4942.
4. Frushicheva MP, Cao J, Chu ZT, Warshel A (2010) Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. *Proc Natl Acad Sci USA* 107(39):16869–16874.
5. Fowler DM, Fields S (2014) Deep mutational scanning: A new style of protein science. *Nat Methods* 11(8):801–807.
6. Fowler DM, et al. (2010) High-resolution mapping of protein sequence-function relationships. *Nat Methods* 7(9):741–746.
7. Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape. *Proc Natl Acad Sci USA* 108(19):7896–7901.
8. Whitehead TA, et al. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol* 30(6):543–548.
9. McLaughlin RN, Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R (2012) The spatial architecture of protein function and adaptation. *Nature* 491(7422):138–142.
10. Agresti JJ, et al. (2010) Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc Natl Acad Sci USA* 107(9):4004–4009.
11. Kintses B, et al. (2012) Picoliter cell lysate assays in microfluidic droplet compartments for directed enzyme evolution. *Chem Biol* 19(8):1001–1009.
12. Granieri L, Baret JC, Griffiths AD, Merten CA (2010) High-throughput screening of enzymes by retroviral display using droplet-based microfluidics. *Chem Biol* 17(3):229–235.
13. Fallah-Araghi A, Baret J-C, Ryckelynck M, Griffiths AD (2012) A completely in vitro ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab Chip* 12(5):882–891.
14. Fidalgo LM, et al. (2008) From microdroplets to microfluidics: Selective emulsion separation in microfluidic devices. *Angew Chem Int Ed Engl* 47(11):2042–2045.
15. Jacquier H, et al. (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc Natl Acad Sci USA* 110(32):13067–13072.
16. Guo HH, Choe J, Loeb LA (2004) Protein tolerance to random amino acid change. *Proc Natl Acad Sci USA* 101(25):9205–9210.
17. Bloom JD, et al. (2005) Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA* 102(3):606–611.
18. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444(7121):929–932.
19. Zechel DL, Withers SG (2000) Glycosidase mechanisms: Anatomy of a finely tuned catalyst. *Acc Chem Res* 33(1):11–18.
20. Davies G, Henrissat B (1995) Structures and mechanisms of glycosyl hydrolases. *Structure* 3(9):853–859.
21. Marana SR (2006) Molecular basis of substrate specificity in family 1 glycoside hydrolases. *IUBMB Life* 58(2):63–73.
22. Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: Evolutionary units of three-dimensional structure. *Cell* 138(4):774–786.
23. Sullivan BJ, et al. (2012) Stabilizing proteins from sequence statistics: The interplay of conservation and correlation in triosephosphate isomerase stability. *J Mol Biol* 420(4-5):384–399.
24. Adkar BV, et al. (2012) Protein model discrimination using mutational sensitivity derived from deep sequencing. *Structure* 20(2):371–381.
25. Wu NC, et al. (2013) Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. *J Virol* 87(2):1193–1199.
26. Wagenaar TR, et al. (2014) Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell Melanoma Res* 27(1):124–133.
27. Araya CL, et al. (2012) A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci USA* 109(42):16858–16863.
28. Sciambi A, Abate AR (2015) Accurate microfluidic sorting of droplets at 30 kHz. *Lab Chip* 15(1):47–51.
29. Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods* 7(2):119–122.
30. Lundin S, et al. (2013) Hierarchical molecular tagging to resolve long continuous sequences by massively parallel sequencing. *Sci Rep* 3:1186.
31. Aharoni A, Griffiths AD, Tawfik DS (2005) High-throughput screens and selections of enzyme-encoding genes. *Curr Opin Chem Biol* 9(2):210–216.
32. Lim SW, Abate AR (2013) Ultrahigh-throughput sorting of microfluidic drops with flow cytometry. *Lab Chip* 13(23):4563–4572.
33. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci USA* 98(8):4552–4557.
34. Beneyton T, Coldren F, Baret J-C, Griffiths AD, Taly V (2014) CotA laccase: High-throughput manipulation and analysis of recombinant enzyme libraries expressed in *E. coli* using droplet-based microfluidics. *Analyst (Lond)* 139(13):3314–3323.
35. Ma F, Xie Y, Huang C, Feng Y, Yang G (2014) An improved single cell ultrahigh throughput screening method based on in vitro compartmentalization. *PLoS One* 9(2):e89785.
36. Tu R, Martinez R, Prodanovic R, Klein M, Schwaneberg U (2011) A flow cytometry-based screening system for directed evolution of proteases. *J Biomol Screen* 16(3):285–294.
37. Ryckelynck M, et al. (2015) Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. *RNA* 21(3):458–469.
38. Skhiri Y, et al. (2012) Dynamics of molecular transport by surfactants in emulsions. *Soft Matter* 8(41):10618–10627.
39. Romero PA, Krause A, Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci USA* 110(3):E193–E201.
40. Bloom JD, et al. (2007) Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol* 5:29.
41. Quan J, Tian J (2011) Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nat Protoc* 6(2):242–251.
42. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
43. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42(Database issue):D490–D495.
44. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
45. Dereeper A, et al. (2008) Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 36(Web Server issue):W465–W469.