A comprehensive, high-resolution map of a gene's fitness landscape

Elad Firnberg[1], Jason W. Labonte[1], Jeffrey J. Gray[1], and Marc Ostermeier[1]

[1]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218 USA

Supplemental Materials and Methods
Supplemental References

SUPPLEMENTAL MATERIALS AND METHODS

Description of the band-pass selection system.

The band-pass genetic selection for β-lactamase activity can select for *E. coli* cells exhibiting any desired level of β-lactamase activity (Sohka, et al. 2009) (Fig. S1). The two plasmids that comprise the system and the interactions of their components are shown in Supplementary Fig. 1A. In this system, cells with too little β-lactamase activity (relative to the amount of β-lactam antibiotic such as ampicillin) cannot grow due to the β-lactam's inhibitory effect on cell wall synthesis. However, the β-lactam also serves to confer resistance to the antibiotic tetracycline (Tet) via induction of the *ampC* promoter. Cells with too much β-lactamase activity (relative to the amount of β-lactam antibiotic) cannot grow in the presence of Tet since rapid degradation of the β-lactam prevents sufficient induction of TetC expression. Thus cells challenged to grow in the presence of Tet and Amp will grow only if the amount of β-lactamase activity is balanced between having enough activity to degrade the β-lactam to allow cell wall synthesis but not too much to prevent induction of TetC (Fig. S1B). Since increased β-lactam concentration shifts the amount of β-lactamase activity necessary for growth to higher levels, our cells can be used to select for any level of β-lactamase activity (from the absence of activity to full activity) simply by adding the necessary concentration of the β-lactam. A more in depth characterization of this system is provided in Sokha et al. (Sohka, et al. 2009) including the linear relationship between a cell's minimum inhibitory concentration (MIC) of Amp in the absence of Tet and the concentration of Amp in the presence of Tet that provides the best growth for that cell. This linear relationship holds over three orders of magnitude of Amp concentration. In the experiments reported here, gene fitness corresponds to the Amp concentration providing the best growth in the presence of Tet, which is roughly 25% of the MIC_{Amp}. Although in our system *TEM-1* is under the IPTG

inducible *tac* promoter instead of its native promoter, the level of Amp resistance is similar to that of its native promoter (only 4-fold higher) and very little protein aggregates (Figure S15B). Thus, the expression of TEM-1 from the *tac* promoter is comparable that from its native promoter and does not have overexpression artifacts.

Gene fitness determination.

The method for measuring fitness makes use of the band-pass genetic selection system (Fig. S1) and is schematically depicted in Fig. S2. Comprehensive codon mutagenesis library CCM2 comprises three separately constructed libraries, one for each third of the gene (Firnberg and Ostermeier 2012). Collectively, they are designed to contain all possible single codon substitutions in the *TEM-1* gene. Each library was plated on LB-agar plates containing 50 μg/ml spectinomycin, 50 μg/ml chloramphenicol, 300 μM IPTG, 20 μg/ml tetracycline and 13 different ampicillin concentrations (2-fold increments of ampicillin ranging from 0.25 μg/ml to 1024 μg/ml) at a cell density of 1700 CFU/cm^2. Plates were incubated at 37°C for 20 hrs. Colonies were recovered from the 39 plates with LB broth, and plasmid DNA isolated using the Qiagen QIAprep Spin Miniprep kit (27106). The plasmid DNA was linearized by restriction endonuclease digestion with SphI and purified using the Zymo DNA Clean & Concentrator kit. Because the plasmid miniprep also contained the band-pass plasmid, pTS42, the concentration of the library plasmid, pSkunk3-CCM2, was determined by running a sample of the linearized DNA on an ethidium bromide agarose gel and analyzing the band intensities of the two respective linearized plasmids. The mass ratio of pTS42:pSkunk3-CCM2 was determined to be ~12.5.

PCR amplicons of each of the 39 sub-libraries were created using Titanium Lib-A "A" and "B" fusion primers that included a 10-base MID barcode identifying the sub-library from which the DNA originated. In order to minimize the rate of recombination or

"PCR-jumping" between DNA template strands in the PCR amplification, experiments were performed to find the minimum amount of template DNA and minimum number of PCR cycles necessary to obtain sufficient PCR product. Each 25 µl PCR reaction had 22.4 pg (5 x 10^6 molecules pSkunk3-CCM2) linearized template DNA, 0.5 µM each barcoded primer, 200 µM each dNTP, 1X HF Phusion buffer, and 0.5 units Phusion high-fidelity DNA polymerase. Cycler conditions were 98°C for 30 sec, 25 cycles of 98°C for 10 sec, 61.9°C for 15 sec, 72°C for 3 min, and then 72°C for 5 min. PCR product DNA concentration for each reaction was determined using the Quant-iT Picogreen dsDNA Assay kit (P7589). The barcoded amplicons were then mixed together in molar proportion to the number of colonies that grew on their respective sub-library selection plate. A total of 2.4 µg of the amplicon mixture was electrophoresed on a TAE 0.7% agarose gel and then gel purified using the QIAquick Gel Extraction Kit (28706). Further purification was performed using the Agencourt AMPure XP PCR Purification kit (A63880) to remove short DNA fragments, primers, and primer dimers. The final purified amplicon DNA concentration was determined using picogreen and diluted to 10^9 molecules/µl in 1X TE and then further diluted to 10^7 molecules/µl in DI water. 454 sequencing was performed by Tufts University Core Facility on a Roche 454 GS FLX+ instrument.

We created custom MATLAB scripts to analyze the raw 454 sequencing reads. From a combination of five full or partial plate runs, we obtained a total of 1,325,979 reads. We aligned the reads to the template sequence, sorted the reads by barcode, and mapped all codon substitution mutations, ignoring indels (which are a common sequencing error). Reads were filtered out if the average base call quality score was less than 30, the read did not span the entire mutagenesis region, or if there was more than one codon mutation per read. We obtained 772,296 reads that passed our filtering requirements and had a single codon substitution.

We tabulated the number of sequencing counts for each allele in each sub-library (see Data S1). Since the distribution of growth as a function of Amp is roughly symmetric when plotted as the $\log_2$(Amp concentration) (Sohka, et al. 2009), we determined the unnormalized fitness *f* of allele *i* as

$$f_i = \frac{\sum_{p=1}^{13} c_{i,p} \log_2\left(a_p\right)}{\sum_{p=1}^{13} c_{i,p}} \qquad \text{Equation 2}$$

in which $c_{i,p}$ is the number of counts of allele *i* on sub-library plate *p* in the deep sequencing data and a_p is the concentration of Amp on sub-library plate *p* in μg/ml . Counts of a particular allele can be expected to appear on 3-4 adjacent sub-library plates (Sohka, et al. 2009), corresponding to an 8-16 fold window of Amp concentration. In a minority of cases however, counts were observed in more than four sub-libraries or counts were observed to cluster in non-adjacent sub-libraries. Two phenomena can account for this: sequencing errors and the presence of an unintended, fitness-altering mutation outside the sequencing region. The frequency of sequencing errors is low relative to that of mutations, since 87% of the library members contain a mutation, and sequencing errors should occur with equal frequency among all sub-libraries. We implemented several measures to facilitate appropriate fitness assignment. First, an allele was assigned a fitness value only if its sequence was observed at least five times. Second, the window of the four adjacent sub-libraries with the highest combined sequencing counts was identified and only these values were included in the fitness calculation. Third, if an allele presented with multiple clusters of counts, the cluster corresponding to the fitness closest to the average fitness of the other synonymous codons was selected. For alleles with mutations in the start codon, knowledge about

alternative *E. coli* start codons was used to assign fitness. For alleles with nonsense mutations with multiple clusters of counts, the lowest cluster was used to assign fitness. For the fitness of missense mutations, we combined the sequencing counts of the corresponding synonymous codons (Data S2) and recalculated the fitness using the same methods as above for codon substitutions.

We normalized all fitnesses by the fitness of wildtype as follows:

$$w_i = \frac{2^{f_i}}{2^{f_{WT}}} \qquad \text{Equation 3}$$

This result is a normalized fitness w_i that is 1.0 for wildtype *TEM-1*, > 1.0 for beneficial mutations and between 0 and 1.0 for deleterious mutations. We determined the fitness of wildtype *TEM-1* (f_{WT}) using Eq. 2 using the counts of all alleles with a synonymous substitution in *TEM-1*, since the fitness of these varied very little. As a check, we compared this value to the fitness determined by Equation 2 using the counts of all sequencing reads that lacked a mutation. The two values differed by only 2.5%. Gene fitness values are tabulated in Data S1. Protein fitness values were found by averaging the gene fitness values of synonymous genes and are tabulated in Data S2.

We determined an upper limit on the error in our fitness measurements. We assumed (solely for the purpose of this error determination) that synonymous mutations have no fitness effect. We compared an allele's gene fitness to the mean of all alleles with a synonymous mutation at the same position (i.e. the protein fitness) and expressed this difference as a percentage of the mean. The distribution of values for this 'percent difference in fitness' did not vary with fitness, indicating our fitness measurements are equally accurate at low and high fitness values (Fig. S3A,B), unlike in growth competition experiments. As expected, the width of the distribution narrowed with the number of times the allele was observed in the deep sequencing results (the 'allele count') (Fig. S3C,D). We used the standard deviation of this distribution as a function of allele count

as an estimate of the error in gene fitness. This error is an upper limit since synonymous mutations can have fitness effects (Plotkin and Kudla 2011).

Prediction of protein thermodynamic stability.

PyRosetta v3.4.0 r55307 (Chaudhury, et al. 2010) was used to compute the difference in score (in Rosetta Energy Units, REU) between the mature structures (lacking the signal sequence) of each amino acid mutant and wild type TEM-1 (Protein Data Bank identifier 1XPB (Fonze, et al. 1995))). The score is designed to capture the change in thermodynamic stability caused by the mutation ($\Delta\Delta G$) (Das and Baker 2008). First, all side chains were repacked (sampling from the 2010 Dunbrack rotamer library (Shapovalov and Dunbrack 2011)) and minimized for the wild type structure using the talaris2013 scoring function. Next, each missense mutation was introduced and all residues within a 10 Å distance of the mutated residue's center were repacked followed by a linear minimization of the backbone and all side chains. This procedure was performed 50 times, and the predicted $\Delta\Delta G$ is the average of the three lowest scoring structures. The PyRosetta script used has been included as part of the standard PyRosetta installation (http://www.pyrosetta.org/dow) and can be found in the apps directory of the PyRosetta installation as delta_score_per_mutation.py. PoPMuSiC predictions of $\Delta\Delta G$ (Fig S12B) were determine online at http://babylone.ulb.ac.be/popmusic (Dehouck, et al. 2011).

Preparation of samples for protein abundance and total catalytic activity assays.

For analysis of the sub-libraries, the three libraries of CCM2 were combined and plated on 13 different Amp concentrations as in the gene fitness measurement experiment described above. Individual clones were plated on the same Amp concentration upon which they were found (i.e. 8 or 16 μg/ml) at the same colony density at which the

original library grew (i.e. the numbers of the colonies on the plates were the same). Colonies were recovered by sweeping with LB broth containing 15 v/v% glycerol and 2 w/v% glucose. Aliquots comprising cultures with equal cell density were pelleted and the supernatant removed. After freezing at -80 ^{o}C for 10 min, the cells were lysed in 250 µL BugBuster Protein Extraction Reagent (Novagen 70584-3) containing ≥0.25 U/µL benzonase nuclease (Sigma E1014-25KU), 3 units/µL rLysozyme (Novagen 71110), and 1% protease inhibitor cocktail (Sigma P8849). Samples were incubated with gentle shaking at 4^{o}C for 30 min, then centrifuged at 14,000 rpm for 30 min at 4°C. The supernatant was recovered as the soluble protein fraction and the pellet resuspended in 250 µL 8M urea as the insoluble fraction. Samples were aliquoted and stored at -80°C. Total protein concentration of each lysate sample was measured using the DC protein assay (Bio-Rad 500-0111) with a BSA standard.

Protein abundance.

Western blots for each sub-library were performed with 10 µg total protein of each lysate sample. A standard curve was prepared by diluting the lysate from sub-library 13 in 2-fold increments, with increasing additions of control lysate from cells not expressing any *TEM-1* allele to maintain a constant amount of 10 µg total protein. SDS-PAGE gels (Novex NP0323BOX) were electrophoresed for 45 min at 190 V, and transferred to a PVDF membrane (Bio-Rad #162-0177) for 30 min at 15 V. The membrane was blocked with a 4% milk solution in PBST (1X PBS, 0.05% tween 20) for 1 hr at room temperature with shaking. The primary anti-TEM-1 mouse monoclonal antibody (Thermo MA1-20370, 500-fold dilution in blocking solution) was incubated for 2 hrs at room temperature or overnight at 4^{o}C with shaking. The membrane was then washed three times with PBST for 5 min with shaking. The secondary goat anti-mouse antibody (Bio-Rad 170-5047, 20,000-fold dilution in blocking buffer) was incubated for 1 hr with shaking at RT. The

membrane was then washed three times with PBST for 5 min with shaking, and a final wash with 1X PBS. A total of 1 ml of chemiluminescence detection reagent (Bio-Rad #170-5070) was then applied to the membrane, incubated for 1 min, and then imaged on a Bio-Rad Gel Doc XR system, recording exposures at 5 sec interval. Representative westerns are shown in Supplementary Fig. 13. Quantity One 1-D analysis software (Bio-Rad) was used to quantify the band intensity for BLA sub-library or clone samples. For each western blot, the image of longest exposure before any band reached detector saturation was selected for quantification. The local adjusted volume parameter was used as the measure of band intensity because it subtracts the local background around each band. Intensity was then converted to relative protein concentration using the standard curve correlation with the same corresponding exposure time.

Total catalytic activity.

Catalytic activity of the sub-libraries and clones was determined by measuring nitrocefin hydrolysis rates. A solution of 50 μM nitrocefin in 10 mM phosphate buffer pH 7.4 was incubated in a 96-well plate (BD Falcon 353072) at 37^{o}C for 2 min. Then 1.0 to 2.5 μL of the soluble fraction of the lysate was added to a final volume of 200 μL and mixed by pipetting up and down briefly. The rate of hydrolysis was measured as the initial slope of absorbance as a function of time as measured at 486 nm at 37°C in a SpectraMax Plus 384 Microplate reader. The initial rate was normalized by the total amount of protein added for each sample.

Theoretical calculation of total catalytic activity vs. fitness.

Allele fitness results from the ability to hydrolyze Amp at the concentration of Amp present on the plate. Our experimental measure of total catalytic activity uses the initial rate of nitrocefin hydrolysis at 50 μM nitrocefin. These will not correlate precisely 1:1

owing to the differences in catalytic constants for hydrolysis of the two substrates and the fact that nitrocefin hydrolysis is measured at a set nitrocefin concentration but fitness is evaluated at different Amp concentrations. However, we can predict the form of the relationship. We assumed Michaelis-Menten kinetics and the following values for k_{cat} and K_m for Amp (1187 s^{-1} and 42.7 μM) and nitrocefin (917 s^{-1} and 128 μM), which are average values for TEM-1 of that measured in several previous studies. We first calculated the reduction of total Amp hydrolysis activity expected for the expected mean fitness of the sub-library. We then assumed that this reduction in catalytic activity comes from equal percent reductions in the k_{cat} and K_m. We then assumed that this same percent reduction in the catalytic constants will occur for nitrocefin hydrolysis. We then calculated the predicted relative initial rate of nitrocefin hydrolysis at 50 μM nitrocefin using the Michaelis-Menten equation. We performed these calculations (a) assuming the protein abundance was the same for all fitness values and (b) using the experimentally observed measurements of protein abundance shown in Fig. 5B. The results are presented as Fig. S14.

SUPPLEMENTAL REFERENCES

Bershtein S, Goldin K, Tawfik DS. 2008. Intense neutral drifts yield robust and evolvable consensus proteins. J Mol Biol 379: 1029-1044.

Brown NG, Pennington JM, Huang W, Ayvaz T, Palzkill T. 2010. Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum TEM beta-lactamases. J Mol Biol 404: 832-846.

Chaudhury S, Lyskov S, Gray JJ. 2010. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 26: 689-691.

Das R, Baker D. 2008. Macromolecular modeling with rosetta. Annu Rev Biochem 77: 363-382.

Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. 2011. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics 12: 151.

Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, Muzny D, Gibbs RA, Palzkill T. 2012. Deep sequencing of systematic combinatorial libraries reveals beta-lactamase sequence constraints at high resolution. J Mol Biol 424: 150-167.

Dietz H, Pfeifle D, Wiedemann B. 1997. The signal molecule for beta-lactamase induction in Enterobacter cloacae is the anhydromuramyl-pentapeptide. Antimicrob Agents Chemother 41: 2113-2120.

Firnberg E, Ostermeier M. 2012. PFunkel: efficient, expansive, user-defined mutagenesis. PLoS ONE 7: e52031.

Fonze E, Charlier P, To'th Y, Vermeire M, Raquet X, Dubus A, Frere JM. 1995. TEM1 beta-lactamase structure solved by molecular replacement and refined structure of the S235A mutant. Acta Crystallogr D Biol Crystallogr 51: 682-694.

Guillaume G, Vanhove M, Lamotte-Brasseur J, Ledent P, Jamin M, Joris B, Frere JM. 1997. Site-directed mutagenesis of glutamate 166 in two beta-lactamases. Kinetic and molecular modeling studies. J Biol Chem 272: 5438-5444.

Kather I, Jakob RP, Dobbek H, Schmid FX. 2008. Increased folding stability of TEM-1 beta-lactamase by in vitro selection. J Mol Biol 383: 238-251.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12: 32-42.

Potapov V, Cohen M, Schreiber G. 2009. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22: 553-560.

Raquet X, Vanhove M, Lamotte-Brasseur J, Goussard S, Courvalin P, Frere JM. 1995. Stability of TEM beta-lactamase mutants hydrolyzing third generation cephalosporins. Proteins 23: 63-72.

Shapovalov MV, Dunbrack RL, Jr. 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19: 844-858.

Sohka T, Heins RA, Phelan RM, Greisler JM, Townsend CA, Ostermeier M. 2009. An externally-tunable bacterial band-pass filter. Proc Natl Acad Sci USA 106: 10135-10140.

Valtonen SJ, Kurittu JS, Karp MT. 2002. A luminescent Escherichia coli biosensor for the high throughput detection of beta-lactams. J Biomol Screen 7: 127-134.

Wang X, Minasov G, Shoichet BK. 2002a. Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. J Mol Biol 320: 85-95.

Wang X, Minasov G, Shoichet BK. 2002b. Noncovalent interaction energies in covalent complexes: TEM-1 beta-lactamase and beta-lactams. Proteins 47: 86-96.