Rapid and Programmable Protein Mutagenesis Using Plasmid Recombineering

Sean A. Higgins[1], Sorel Ouonkap[1], David F. Savage[1,2,*]

[1] Department of Molecular and Cell Biology, UC Berkeley, Berkeley, CA, 94720, USA

[2] Department of Chemistry, UC Berkeley, Berkeley, CA, 94720, USA

* To whom correspondence should be addressed. Email: savage@berkeley.edu.

Address: 2151 Berkeley Way, Berkeley, CA 94720; (510) 643-7847

ABSTRACT

Comprehensive and programmable protein mutagenesis is critical for understanding structure-function relationships and improving protein function. However, current techniques enabling comprehensive protein mutagenesis are based on PCR and require *in vitro* reactions involving specialized protocols and reagents. This has complicated efforts to rapidly and reliably produce desired comprehensive protein libraries. Here we demonstrate that plasmid recombineering is a simple and robust *in vivo* method for the generation of protein mutants for both comprehensive library generation as well as programmable targeting of sequence space. Using the fluorescent protein iLOV as a model target, we build a complete mutagenesis library and find it to be specific and unbiased, detecting 99.8% of our intended mutations. We then develop a thermostability screen and utilize our comprehensive mutation data to rapidly construct a targeted and multiplexed library that identifies significantly improved variants, thus demonstrating rapid protein engineering in a simple one-pot protocol.

INTRODUCTION

Directed mutagenesis of a desired protein is an important technique both for understanding structure-function relationships as well as improving protein function for research, biotechnology, and medical applications. For example, techniques like deep mutational scanning, where every position in a protein is mutated to all possible amino acids, can be applied to understand key variants associated with disease[1], while targeted mutagenesis of proteins such as Green Fluorescent Protein (GFP) have expanded our capacity to visualize many biological processes[2]. The ability to generate comprehensive mutation libraries and programmed libraries focused on specific locations or amino acids is crucial to these applications.

In order to address these needs, many polymerase chain reaction (PCR) based approaches have been developed. Firnberg and Ostermeier have built libraries composed almost entirely of single mutations using specialized protocols based on uracil-containing template DNA[3], while Melnikov and Mikkelsen constructed a comprehensive library by splitting one gene into many different regions small enough to be synthesized on a programmable microarray, followed by multiplexed *in vitro* recombination[4]. Belsare and Lewis have demonstrated targeted, combinatorial library construction using alternating cycles of fragment and joining PCR[5]. Nevertheless, all these techniques suffer from the requirement of multiple complex *in vitro* reactions that are labor-intensive and require specialized protocols or reagents.

An alternative approach would be to incorporate synthetic oligonucleotides *in vivo* directly into a gene of interest in a programmable fashion. In *E. coli*, oligonucleotides introduced into the cell via electroporation can recombine with the genome or resident plasmids with the help of the lambda phage protein Beta, in a process termed recombineering[6]. Mechanistically, it is thought that Beta-bound oligonucleotides anneal to the replication fork of replicating deoxyribonucleic acid (DNA) and are subsequently incorporated into the daughter strand, thus directly encoding mutations into a new DNA molecule[7]. Recombineering is therefore a compelling method for genetic manipulation. Cheap and easily obtained standard oligonucleotides are the only varying input and the protocol - mixing oligonucleotides in one-pot reactions - is straightforward.

This process was shown to be capable of mutating the *E. coli* genome for rapid metabolic engineering in a process termed Multiplexed Automated Genome Engineering (MAGE), which used multiple rounds of recombineering to increase the penetrance of mutations[8]. Other work has demonstrated that thousands of pooled, barcoded oligonucleotides can be used, in parallel, to modify the expression of > 95% of *E. coli* genes and map their effect on fitness[9]. More recent studies have combined recombineering with the programmable DNA nuclease Cas9, as a means of enforcing mutational penetrance, to mutate tens of thousands of loci in parallel with high efficiency[10].

Despite its success in genome engineering, recombineering of plasmids is relatively uncharacterized[11]. Plasmid recombineering (PR) is of particular interest in protein mutagenesis as plasmids are easily shuttled between different strains and

74 organisms for cloning and screening. Notably, recombineering strains achieve
75 enhanced mutation efficiency by knocking out mismatch repair and possess a higher
76 genome-wide mutation rate[12], which can complicate screens or selections sensitive
77 to suppressor mutations. The use of plasmids, however, uncouples protein variation
78 from any background mutation in the genome. Thomason et al. have previously
79 demonstrated that PR is capable of generating mutations, insertions, and deletions
80 with efficiencies comparable to genomic recombineering. We reasoned that the
81 principles of MAGE – a one-pot reaction and multiple mutation rounds – would be
82 applicable to PR as well.

83 To benchmark comprehensive PR for protein engineering we sought to
84 measure the efficiency, bias, and overall performance of saturation mutagenesis on
85 the small protein iLOV. iLOV is a 110 residue protein derived from the Light,
86 Oxygen, Voltage (LOV) domains of the *A. thaliana* phototropin 2 protein[13]. The
87 native LOV domain binds flavin mononucleotide (FMN) and uses this co-factor as a
88 photosensor to direct downstream signal transduction. Mutational analysis has
89 revealed that a cysteine to alanine substitution in the FMN binding site interrupts
90 the native photocycle and instead dramatically increases the protein's fluorescent
91 properties. iLOV is an ideal candidate for further engineering because fluorescent
92 proteins that don't require molecular oxygen for chromophore maturation are a
93 desirable alternative to green fluorescent protein. In previous experiments, DNA
94 shuffling was used to isolate iLOV, a variant that has six amino acid mutations
95 relative to the wild-type (WT) phototropin 2 LOV2 sequence and an improved
96 fluorescence quantum yield of 0.44[13]. Additional approaches to engineer further

97 improved iLOV variants have also relied on error-prone PCR and DNA shuffling,
98 missing much of the possible sequence-space[14]. Due to the potential utility of iLOV
99 and its comparatively limited engineering relative to other fluorescent proteins, we
100 hypothesized iLOV could serve as an excellent model system for exploring the utility
101 of PR. Finally, the gene length of iLOV is exceptionally short (330 bp) and analysis of
102 iLOV libraries is suited to deep sequencing. Current paired-end sequencing covers
103 the entirety of the open reading frame and can accurately identify all mutations to a
104 single sequenced plasmid. This provides insight into the mechanisms and utility of
105 recombineering.

106 Here we demonstrate that PR is capable of constructing both comprehensive
107 protein libraries and targeted mutagenesis libraries focusing on a small section of
108 sequence space. We built a complete mutagenesis library of iLOV and found it to be
109 specific and unbiased, detecting 99.8 % of our intended mutations. We explored this
110 fitness landscape in the context of thermostability using a plated-based screen that
111 allowed us to identify many desirable thermostabilizing mutations. To demonstrate
112 the iterative and programmable nature of our platform, we designed and built a
113 multiplexed library focused on these mutational hotspots and isolated significantly
114 more stable variants. In total, this work demonstrates that plasmid recombineering
115 is a rapid and robust method for the generation of protein mutants for both
116 unbiased, comprehensive libraries and programmable targeting of specific regions
117 in sequence space.

118

119

RESULTS AND DISCUSSION

A one-day, one-pot reaction generates a comprehensive mutation library

To generate a comprehensive mutation library of iLOV, oligonucleotides were designed to target the gene within a high copy ColE1 plasmid (Supp. Figure 1). The target plasmid contained a promoterless iLOV coding region to prevent growth biases between mutants possessing different fitness. 10 ng of target plasmid was mixed with an equimolar mixture of 110 recombineering oligonucleotides, one primer for each codon in iLOV (Figure 1A). These oligonucleotides were 60 bp long and complementary to the lagging strand, which was previously demonstrated to be more efficient than targeting the leading strand[15]. Oligos contained a centrally located NNM mutation codon, (where N = A/C/G/T and M = A/C) which encodes all amino acids except methionine and tryptophan. Tryptophan, in particular, is known to quench flavin fluorescence in flavoproteins[16] and was excluded from the library. Although modified oligonucleotides have been shown to enhance recombineering efficiency, e.g. phosphorothioation, standard oligonucleotides were used to minimize cost and complexity[8].

Initial experiments confirmed that PR can be used to generate diverse libraries in a programmable fashion. The plasmid and oligonucleotide mixture was first electroporated into the recombineering strain EcNR2[8], grown overnight and miniprepped. Deep sequencing of the recovered library, hereafter termed the Round 1 library, revealed substantial mutagenesis compared to the non-recombineered plasmid (Supp. Figure 2). Further analysis revealed that, while the majority of reads were WT iLOV sequence, 29% of reads contained a single codon mutation. These

mutations covered every position in the protein and nearly all targeted amino acid conversions were observed (Figure 1B). 1867 out of the possible 1870 single residue mutations were detected in the Round 1 library (169972 reads passing quality threshold).

Further founds of PR were used to increase the penetrance of mutations. Although the Round 1 library covered targeted mutations comprehensively, it was roughly 61% WT iLOV sequence. Furthermore, the programmable nature of PR allows the construction of more targeted libraries with combinatorial multiplexing of high fitness mutations, such as would be useful in directed evolution (Figure 1C). Four additional rounds of recombineering were performed to reduce the WT fraction of the library and investigate the distribution of variants with 2+ mutations. The number of reads with codon mutations increased substantially with further rounds of recombineering (Figure 1D). In the Round 5 library, single mutations were the most common (33%), with the remainder composed of WT sequences (26%) and sequences containing 2+ mutations (41%).

Recombineering libraries can be finely controlled to alter the composition of mutations

In the absence of prior information, the ideal mutagenesis technique would produce every mutant targeted with equal frequency. Because each variant is initially present in the same amount, a uniform library requires the least amount of screening (or selection) in order to isolate improved mutants. A non-uniform method, in contrast, might produce a highly variable distribution of mutants and

166 requires more screening or selection in order to fully explore sequence space. We
167 therefore analyzed our sequencing data to characterize the uniformity and sequence
168 preferences of PR libraries.

169 We detected two distinct types of bias: positional bias and mismatch bias. In
170 positional bias, oligonucleotides targeted to different positions in the coding
171 sequence incorporate with characteristically different efficiencies (regardless of the
172 mutation produced at that position). In mismatch bias, oligonucleotides that are
173 more similar to the WT sequence (e.g. differing only at the first base of the targeted
174 codon) incorporated with high efficiency while more divergent oligos (e.g. different
175 at all three positions) incorporated with lower efficiency. Positional bias is evident
176 when comparing the frequencies of single codon mutations across all 110 codons in
177 iLOV (Figure 2A). Despite the nearly 60 bp of homology between the oligonucleotide
178 and the template plasmid, oligos targeting adjacent codons can exhibit a 2-3 fold
179 different incorporation. The largest such discrepancy is found at codon 42, which is
180 mutated 5.7 times more frequently than codon 41. Variation in efficiency has been
181 observed in other recombineering studies and was somewhat correlated with the
182 oligonucleotide binding energy[8]. While our data was not clearly correlated with
183 binding energy (Supp. Figure 3), a biological replicate revealed replicable positional
184 bias (Supp. Figure 4), suggesting the presence of an underlying physical mechanism
185 for positional bias.

186 Comprehensive mutagenesis applications, such as deep mutational scanning,
187 ideally begin with uniformly distributed mutants in a naïve library. We
188 hypothesized that the positional bias observed in the Round 1 library – constructed

using equimolar mutagenic oligonucleotides – could be corrected by altering the mixture of oligonucleotides used in the electroporation step. A second library was therefore constructed by normalizing the concentration of each oligonucleotide in the library according to the frequency of mutations obtained at the corresponding position in Round 1. This library was constructed using PR, and the first 40 amino acids were deep sequenced. The normalized library was indeed more uniform, with a largest adjacent codon discrepancy of 2.1 fold, compared to 3.3 fold in the equimolar replication library (Figure 2B).

In order to understand the importance of this normalization in a quantitative fashion, we performed a simulation to evaluate the pragmatic impact of oligonucleotide normalization on effective library size – the number of samples that must be taken from a library to achieve a desired representation of the library diversity. In the ideal case, i.e. mutants are found in the library with equal frequency, if one wishes to sample, say, 95 % of the diversity contained within the library, we must screen roughly three times the number of distinct members. That is, the effective library size is threefold the targeted library size. Our equimolar and normalized libraries are not uniformly distributed, but random sampling from our sequenced data *in silico* allows calculation of the effective library size. Our simulation revealed a substantial improvement in sampling efficiency, from a mean effective library size of 139 in the original library to 107 in the normalized library, which is quite close to that of an ideal uniformly distributed library (Supp. Figure 5).

The presence of mismatch bias in the library demonstrates that recombineering favors incorporation of oligos that are more similar to the WT

template sequence. Our oligonucleotides all contained one, two, or three nucleotide mismatches relative to WT iLOV. After computationally enumerating all possible oligonucleotides for each codon in iLOV, we expected that 14% of oligonucleotides would contain one mismatch, 43% two, and 43% three. In contrast, oligonucleotide-template pairs with two or three mismatches were observed at only 31% and 22% respectively, while single nucleotide mismatches were overrepresented by more than threefold at 47% (Figure 2C).

While a single round of PR generates many single mutants, additional rounds shift the distribution to increasing mutation numbers. The Round 5 library, for example, is 25% double mutants. These mutations are well represented, with 5817 out of a possible 5940 locations detected (Figure 2D). However, it is clear that the same positional bias seen in the Round 1 library is preserved in subsequent rounds. This is to be expected if recombineering events are statistically independent, and can be seen by the correlation of double-mutation hotspots with single mutation positional bias (Supp. Figure 6). This data suggests that double mutations in a normalized library will be far more uniformly distributed.

The double mutation data also indicates that some recombineering events are not perfectly independent. A plot of the pairwise distance between all detected double mutants reveals an uneven distribution in frequency (Figure 2E). Specifically, double mutations are less likely to be within 30 bp (i.e. 10 amino acids) of one another. This effect becomes more pronounced with additional rounds of recombineering (Supp. Figure 7). In the Round 5 library, this 'zone of exclusion' is significant enough that double mutations a few amino acids apart are nearly four

times less frequent than double mutations with a much larger separation (e.g. three amino acid gap versus 10 amino acid gap). We hypothesize that this bias is due to the mechanism of recombineering which requires oligonucleotides to anneal to a complementary locus via homology arms flanking the NNM mutagenic codon. In this mechanism, sequential oligonucleotides would 'overwrite' previous mutations due to incorporation of the most recent oligonucleotide's homology arms, which extend 30 bp on either side of the central NNM. Another potential mechanism that disfavors incorporation of nearby double mutants is mutational reduction of oligo incorporation efficiency. In this hypothetical mechanism, mutations generated in early rounds of PR could reduce the homology and, therefore, incorporation efficiency of oligos in subsequent PR rounds.

Screening the iLOV library identifies mutations conferring thermostability

Previous screening and structural work indicates that a well-packed binding site for the FMN fluorophore may lead to improved photochemical properties of iLOV, such as photostability, by limiting the dynamics of the FMN chromophore and its ability to dissipate energy following excitation[14]. Additionally, searches for improved LOV-based fluorescent reporters have turned up homologous variants such as CreiLOV that, while brighter (50% greater quantum yield), exhibit substantial toxicity upon expression[17]. More generally, Tawfik and colleagues have theorized that thermostable proteins serve as more fruitful starting points for engineering and directed evolution[18]. In this view, variants with greater thermostability can better tolerate mutations, increasing the likelihood of observing

mutations that improve protein function in a manner unrelated to thermostability. We thus investigated the thermostability of our library, which contained nearly every single amino acid mutation and a small fraction of possible double mutations.

We created a plate-based assay to screen for thermostabilized variants in the Round 5 library. The library was plated at high colony density onto standard LB-Agar, grown overnight, then incubated at 60 °C for two hours (Figure 3A). This treatment completely abrogates the fluorescence of WT iLOV as well as nearly all mutants (Figure 3B). Some colonies remained fluorescent, however, and these were recovered and expressed in 96-well plate format. Cultures expressing these library members were lysed, clarified, and analyzed for thermostability. Fluorescence measurements were taken every 0.5° C during a temperature ramp from 25° C to 95° C and used to calculate the melting temperature (T_m) for each clone, the temperature at which 50% of maximal fluorescence is retained (Figure 3C). All 93 assayed clones demonstrated a substantial increase in thermostability relative to iLOV (Figure 3D). Improvements in T_m ranged from two to nearly ten degrees C.

Recombineering based multiplexing allows rapid and robust directed evolution

Many methods now exist for generating comprehensive single mutant libraries. However, most such methods are incapable of building targeted libraries for exploring the effect of mutations at sites of interest highlighted by previous mutational or structural studies. Because PR is capable of both targeted and comprehensive mutagenesis, we hypothesized that PR could fulfill these

281 requirements in a cost-effective and straightforward one-pot, one-transformation
282 protocol.

283 To demonstrate the rapid multiplexing capability of PR, we designed a
284 second library containing the top 25 most frequent thermostabilizing mutations
285 from the initial plate screen (Figure 4A). Several of these mutations consisted of
286 alternative amino acids at the same position, and in these cases a different
287 oligonucleotide was designed for each. The encoding oligonucleotides were
288 designed such that homology arms would stop short of neighboring mutations so as
289 not to overwrite them (Supp. Figure 8).

290 This multiplexed library resulted in striking improvements to
291 thermostability, with the best variants having T_m values nearly 20° C greater than
292 iLOV (Figure 4B). Isolating and re-cloning these variants verified that despite the
293 significant increase in T_ms, the shape of their melt curves was not significantly
294 different from that of iLOV, even for the most thermostable mutant (Figure 4C). It
295 was noted during the screening process that one variant in particular, hereafter
296 referred to as thermostable LOV (tLOV), seemed to produce abnormally bright
297 lysate under high expression conditions. tLOV and iLOV were expressed and
298 purified in parallel and their absorbance and emission were characterized *in vitro*.
299 To accurately quantify relative quantum yield, the emission curves were normalized
300 for absorption at 450 nm and integrated. tLOV was found to be approximately 10%
301 brighter (Supp. Figure 9), and sequencing revealed the presence of four mutations
302 scattered throughout the protein, all introduced by PR (Figure 4D). Notably, none of
303 these locations are directly within the FMN binding pocket, and their contribution to

304 thermostability or quantum yield improvement are not obvious. Thus, it would have
305 been difficult to rationally design tLOV. Moreover, a targeted mutagenesis technique
306 is required for isolating quadruple mutants reliably: iLOV is a small protein, but a
307 quadruple mutant library would contain > 10^{13} variants, well beyond our current
308 screening capacities. Finally, thermostability was verified by differential scanning
309 calorimetry of iLOV vs tLOV (Supp. Figure 10).

310 An ideal method for generating comprehensive protein libraries would be
311 simple and robust, enabling both complete and targeted mutagenesis without a
312 change in reagents. In this study, we demonstrate that PR can be effectively used for
313 both comprehensive and programmable mutagenesis of the fluorescent protein
314 iLOV, using methods that are generalizable to any gene of interest.

315 Previous recombineering studies have successfully built libraries in the
316 genome, demonstrating specificity and fine control of mutational composition.
317 Wang et al.[8] used multiple rounds of recombineering to mutagenize six consecutive
318 nucleotides using 90 bp oligonucleotides and found a 75% mutation rate after five
319 rounds. We find quite similar behavior in the sequencing analysis of the iLOV
320 recombineered plasmid libraries constructed here. Our Round 5 library consisted of
321 74% mutant variants, and the mutations were well distributed in sequence space.
322 The modest positional bias could be substantially ameliorated by altering the ratios
323 of oligonucleotides added to the electroporation mixture. This resulted in a nearly
324 uniform distribution of mutations such that the effective library size was almost
325 ideal (Supp. Figure 5). This approach could easily be used to accommodate more
326 complex libraries with weighted mutation frequencies at various locations. The

correlation of replicate libraries (Supp. Figure 4) strongly suggests an underlying physical mechanism for differing oligonucleotide recombineering efficiencies, and it will be important to understand this effect in order to predict efficiencies rather than rely on empirical data as used here. Regardless of these modifications, the reagent cost and experimental effort remain minimal in all cases – standard 60 bp oligonucleotides, a one-pot reaction, and simple cycles of electroporation, growth, and plasmid isolation.

Additional biases resulting from the mechanism of recombineering were detected and while their magnitude was smaller, their effect on library size and screening can be significant. Previous work has found that recombineering efficiency drops sharply with the size of the modification made[8]. Here, too, single nucleotide mutations were observed to be more common than double and triple nucleotide mutations. This effect alters the distribution of codons present in the final library, with the template codons determining this frequency shift. Notably, mismatch bias is intrinsic to the annealing of oligonucleotides and is likely present for *in vitro* methods as well.

The 'zone of exclusion' around a first mutation generates another mode of bias. A second mutation is less likely to appear inside this zone than outside of it. This effect became more pronounced with increasing rounds of recombineering. Is it likely this results from oligonucleotides' homology arms overwriting earlier mutations during later rounds of PR. In other words, we hypothesized that homology arms are capable of introducing revertant mutations. We thus predict that the length of homology arms would impact the length of the 'zone of exclusion.' As

some amount of homology is absolutely required for recombineering, no form of recombineering is suitable for efficiently generating adjacent mutations from different oligonucleotides. This limitation can be overcome by using single oligonucleotides encoding sequential mutations but at the cost of reduced efficiency due to mismatch bias. Again, if this effect fundamentally stems from the annealing of oligonucleotides then it is likely present for *in vitro* methods as well.

In vitro methods represent the most powerful alternatives for generating comprehensive mutation libraries. The chief advantage of PCR-based mutagenesis is that it can generate a library composed almost entirely of single mutations. Ostermeier and colleagues accomplished this by utilizing specialized protocols based on uracil-containing template DNA[3]. Such methods have been used to generate nearly comprehensive libraries of genes for exploring the entirety of the fitness landscape[19]. The final proportion of wild-type sequences in this work is comparable to that of the Round 5 library (~25%). In contrast, the remainder of their library is composed almost entirely of single mutations only, while our Round 5 library contained 41% 2+ mutation variants. These two approaches thus generate very different mutation profiles, and choice of one or another will depend on the goals of any particular experiment.

In directed protein evolution, iterative rounds of mutagenesis can be used to multiplex fitness-improving mutations. PCR-based protocols[3], direct gene synthesis[4], and some other recombineering techniques[10] excel at generating libraries composed of single mutations. However, in many applications, 2+ mutations are desired at non-contiguous locations. Recent work has developed a

PCR-based method to accomplish this goal *in vitro*[5], and we hypothesized that PR was well suited to serve as a complementary approach *in vivo*, doing away with cloning altogether. To this end, we comprehensively explored the iLOV single mutation sequence-space for thermostability, selected the fitness enhancing mutations, and demonstrated the utility of PR for advanced protein engineering by multiplexing many different single and double mutations at discontinuous sites across iLOV in a second library. The ability to select and easily mutate numerous specific and non-contiguous locations across a protein is highly useful for a variety of techniques that utilize experimental or phylogenetic data to computationally predict and enhance enzymes[20], explore epistatic interactions[21], or even scan SNPs in human proteins for disease prediction[1].

iLOV engineering has been relatively limited in comparison to other fluorescent proteins[13]. Because the domain has been taken out of its natural structural context, we hypothesized that its thermal stability could be increased. Consistent with this idea, many mutations were found to improve the thermal stability of iLOV up to a robust, 10° C increase in T_m. These improvements were then stacked by multiplexing the 25 most common mutations from the first screen. As a measure of convenience, the same one-pot, single electroporation protocol was used, and the library was designed, built, and tested in one week. One particular variant among the thermostabilized pool, tLOV, was found to be ~10% brighter that iLOV *in vitro*. This result is consistent with previous work demonstrating that constraining the FMN fluorophore can improve the photochemical properties of iLOV[14]. It would be interesting to perform comprehensive mutagenesis of the

396 thermostabilized iLOV mutants in search of further improvements to the protein's
397 brightness or red/blue spectral shifting, as increased thermostability has been
398 hypothesized to permit greater exploration of function-altering mutations[18].

399 In summary, we have demonstrated that PR retains many of the ideal
400 properties of genome recombineering, including specificity and programmability.
401 We found that PR was suited for the construction of both comprehensive and
402 targeted libraries, and that the simplicity of the protocol led to rapid and reliable
403 screening experiments. In particular, PR is suitable for cycles of iterative design,
404 construction, and sampling of genetic libraries, requiring no specialized reagents or
405 protocols. We developed a thermostability screen of the fluorescent protein iLOV
406 and used the resulting mutation data to rapidly construct a multiplexed library that
407 identified significantly improved variants, including the first enhancement to the
408 protein's brightness since its development.

409

MATERIALS AND METHODS

Strains and Media

Strain EcNR2 (Addgene ID: 26931)[8] was used for generating PR libraries in plasmid pSAH031 (Addgene ID: 90330). For thermostability screening and protein expression, iLOV libraries were cloned into pTKEI-Dest (Addgene ID: 79784)[22] using Golden Gate cloning[23] with restriction enzyme BsmbI (NEB) and transformed into either Tuner (Novagen) or XJ b Autolysis *E. coli* (Zymo Research). Unless otherwise stated, strains were grown in standard LB (Teknova) supplemented with kanamycin (Fisher) at 60 µg/mL.

Recombineering Library Construction

Libraries were constructed using a modified protocol from Wang 2011[24]. Briefly, 110 oligonucleotides (Supp. Table 1) or 25 thermostabilizing oligonucleotides (Supp. Table 2) were mixed and diluted in water. A final volume of 50 µL of 2 µM oligonucleotides, plus 10 ng of pSAH031, was electroporated into 1 mL of induced and washed EcNR2 using a 1 mm electroporation cuvette (BioRad GenePulser). A Harvard Apparatus ECM 630 Electroporation System was used with settings 1800 kV, 200 Ω, 25 µF. Three replicate electroporations were performed, then individually allowed to recover at 30° C for 2 hr in 1 mL of SOC (Teknova) without antibiotic. LB and kanamycin was then added to 6 mL final volume and grown overnight. Cultures were miniprepped (QIAprep Spin Miniprep Kit) and monomer plasmids were isolated by agarose gel electrophoresis and gel extraction (QIAquick Gel Extraction Kit) to remove multimer plasmids[11]. The three replicates were then combined, completing a round of PR.

Library Sequencing and Analysis

The iLOV open reading frame was amplified from PR libraries by PCR to add indices and priming sequences for deep sequencing (Supp. Table 3). PCR products were sequenced on Illumina platform sequencers (MiSeq and HiSeq) through the Berkeley Genomics Sequencing Laboratory. Sequencing data were analyzed with a custom MATLAB pipeline. Briefly, reads were filtered to remove those that were of low quality, frameshifted, or did not exactly match the annealing portion of the amplifying primers. An internal control (Supp. Figure 2) verified the low error rate of PCR and Illumina sequencing relative to PR. Finally, full-length reads were compared with the iLOV target sequence for mutation analysis.

In silico Effective Library Size Simulation

Simulations were performed using a custom MATLAB script. The simulations focused on amino acid positions 5 – 40, for which we possess comprehensive frequency data from the equimolar oligonucleotide and normalized PR experiments. Successive sampling, with replacement, of these 36 positions was performed, with each position's likelihood commensurate with the observed mutation frequency. Once 34 out of 36 positions had been observed in the simulation, which we arbitrarily define as 'well-sampled' (94.4% of the library diversity), the total number of samples taken was recorded as the effective library size. This process was then repeated 10^4 times for each library to generate a distribution of effective library sizes (Supp. Figure 5).

Thermostability Screening

455 Colonies were screened on 10 cm dishes containing standard LB-agar with
456 kanamycin. Tuner cells expressing the library were found to be fluorescent in the
457 absence of induction after 24 hours. Plates were incubated at 60° C for 2 hours, after
458 which the vast majority of colonies were no longer fluorescent. Approximately
459 500,000 colonies were screened, of which 244 remained fluorescent. These colonies
460 were pooled, miniprepped, and deep sequenced to identify protein mutations. This
461 DNA was also transformed into XJb cells to recover individual variants. Colonies
462 were then grown overnight in 96-well deep well plates with 1 mL of LB + kanamycin
463 supplemented with 100 µM Isopropyl β-D-1-thiogalactopyranoside (IPTG) and 3
464 mM arabinose at 37° C. Cultures were frozen and thawed to lyse the cells, then
465 clarified by centrifugation. Supernatant was analyzed in a StepOnePlus™ Real-Time
466 PCR System (Applied Biosystems) to estimate a T_m for each variant.

467 **Protein Expression and *in vitro* Characterization**

468 iLOV variants were expressed and lysed in XJb cells as above but in 100 mL
469 volume. Excess FMN (Sigma) was added to the lysate to ensure all proteins
470 possessed ligand. iLOV variants were then purified by nickel affinity
471 chromatography using HisPur™ Ni-NTA Resin (Thermo Scientific). Variants were
472 filtered using Vivaspin 6 3,000 molecular weight cut-off (Sartorius) with phosphate
473 buffered saline (PBS) pH 7.4 (Gibco). Variants were further purified by size
474 exclusion chromatography using a NGC Chromatography System (Bio-Rad). Purified
475 proteins were stored at 4° C in PBS. For quantum yield determination, emission
476 between 460 nm and 600 nm was measured in a FluoroLog Spectrophotometer
477 (Horiba), and 450 nm absorbance was measured in an Infinity M1000 PRO

478 monochromator (Tecan). Emission curves were integrated in MATLAB and
479 normalized for absorbance. Thermostability measurements were made in a Nano
480 differential scanning calorimeter (TA Instruments).
481

FIGURES WITH LEGENDS

Figure 1.

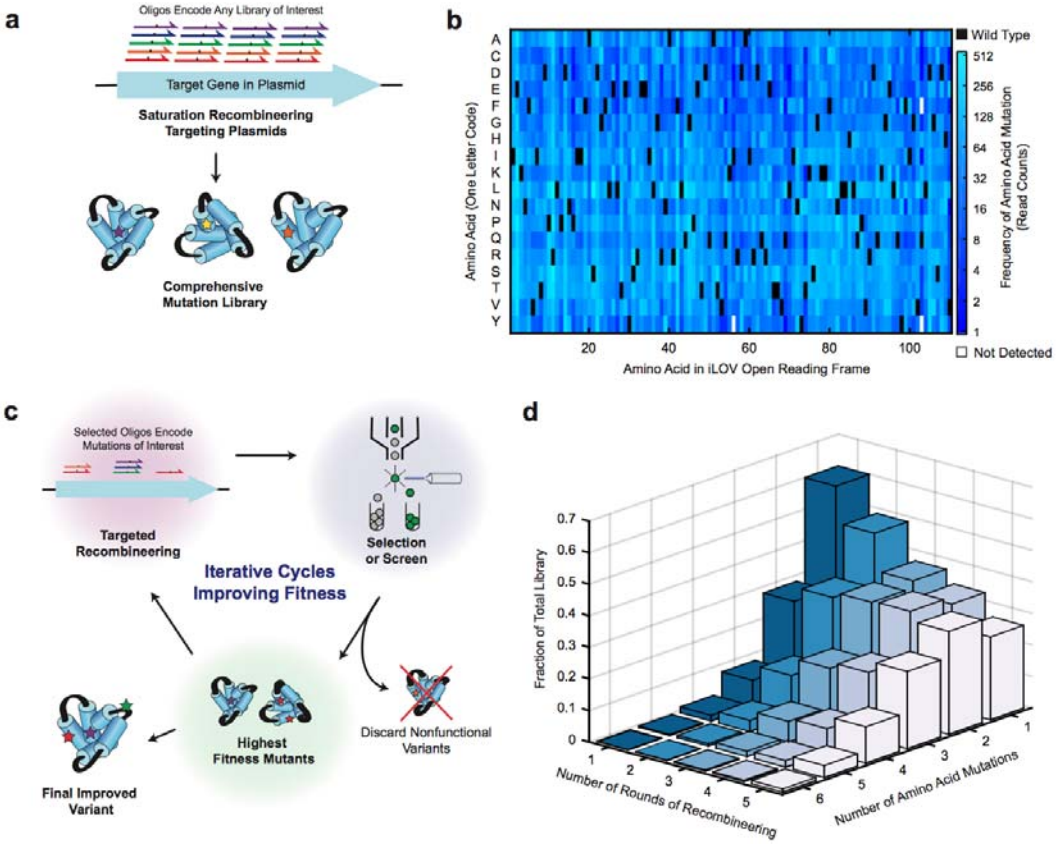


Figure 1. Plasmid Recombineering (PR) of iLOV generates a specific and comprehensive mutation library. (A) Cartoon of comprehensive PR accomplished using synthetic oligonucleotides tiled across the target gene. (B) Frequency of single amino acid mutations mapped by residue and location in the Round 1 library. Black indicates the WT iLOV residue. White indicates no detected reads. (C) Cartoon of programmable PR targeting a specific sequence space. Sampling of the highest fitness mutants can inform subsequent library design, with recombineering oligos specifically targeting mutations of interest. (D) Mutational distribution of the library after additional rounds of recombineering.

Figure 2.

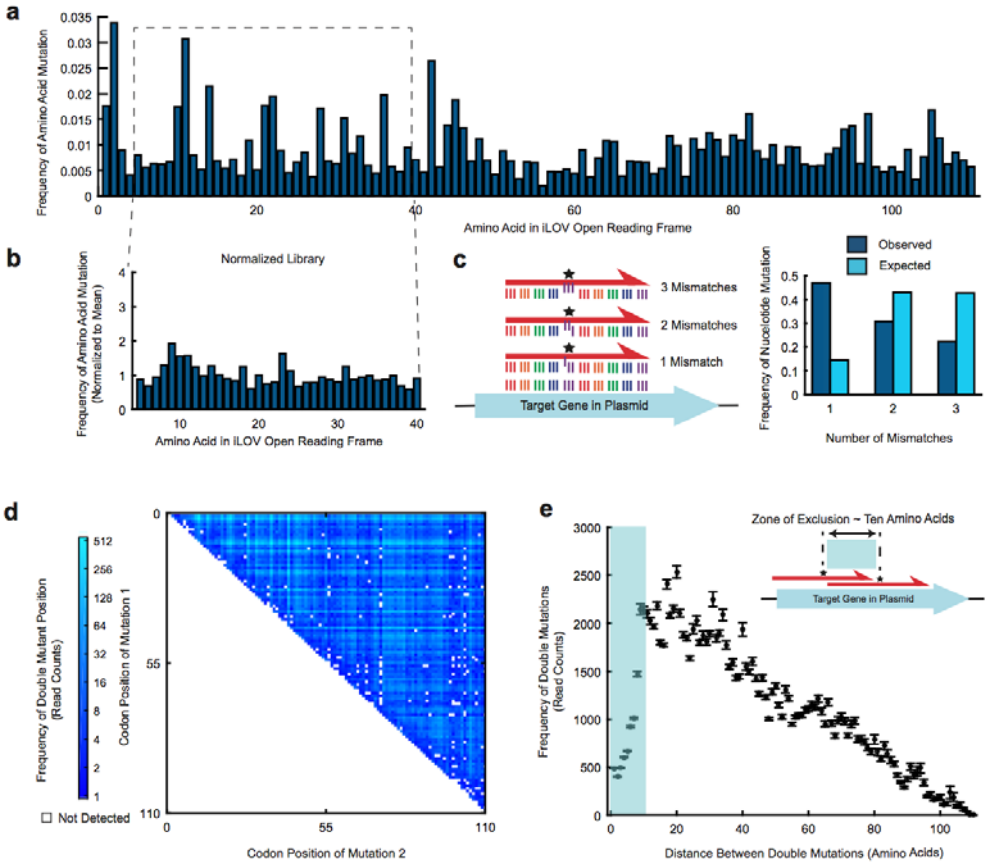
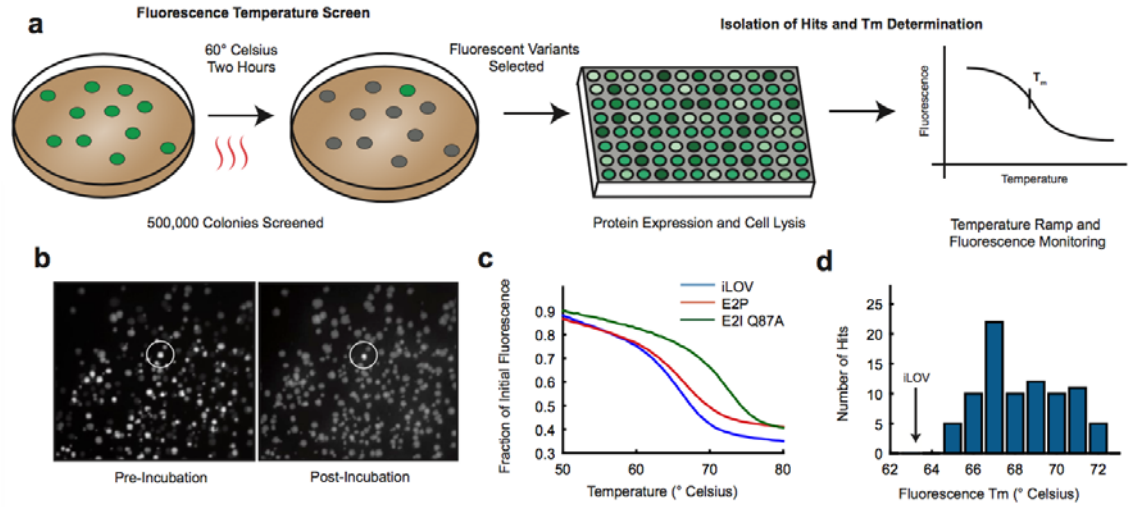


Figure 2. The recombineering mechanism produces moderate bias within the library and can be manipulated by varying the delivered oligonucleotide concentration. (A) Frequency of single amino acid mutations across iLOV in the Round 1 library. (B) Biological replicate of Round 1 library with normalized oligonucleotide concentrations. (C) Observed vs. expected distribution of one, two, or three nucleotide mismatches among single amino acid mutations. The 95% confidence intervals for these measurements are all smaller than +/- 0.0025 (normal approximation to the binomial distribution). (D) Frequency of double mutations in the Round 5 library. White = not detected. Highly represented rows and columns arise from positional bias (Supp. Figure 6). (E) Absolute read counts in the Round 5

506 library as a function of the pairwise distance between double mutations. Note the
507 decreasing linear trend arises from fewer possibilities of double mutations with
508 increasing distance. Error bars, standard deviation.
509
510
511

512 **Figure 3.**



513

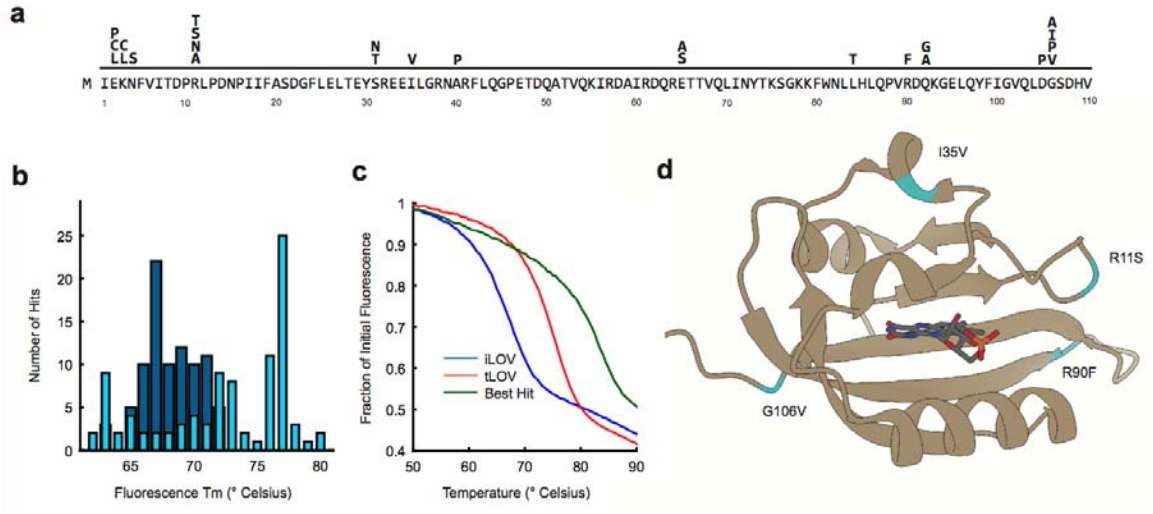
514 Figure 3. A plate-based thermostability screen identifies mutations that improve
515 iLOV fluorescence at elevated temperatures. (A) Cartoon of the thermostability
516 screen assay and subsequent hit validation procedure. (B) Representative
517 fluorescent images of library colonies before and after temperature challenge. (C)
518 Representative fluorescent thermal melt curves of lysate for two thermostable hits.
519 (D) Histogram of T_ms for 93 thermostabilized iLOV variants.

520

521

522

523 **Figure 4.**



524

525 Figure 4. Multiplexing thermostabilizing mutations rapidly identifies doubly
526 improved iLOV variants. (A) 25 thermostabilizing mutations mapped to the iLOV
527 protein sequence. (B) Histogram of initial thermostable hits (Blue) superimposed on
528 hits from the multiplexed library (Cyan). (C) Representative melt curves of two
529 thermostabilized variants compared to iLOV. (D) Crystal structure of iLOV (PDB
530 4EES) indicating the location of four mutations found in tLOV: R11S, I35V, R90F,
531 G106V.

532

SUPPORTING INFORMATION

The Supporting Information is available free of charge on the ACS Publications website at DOI:

Figure S1: Map of iLOV PR target plasmid. Figure S2: Internal control for sequence analysis pipeline. Figure S3: Recombineering efficiency correlation with binding energy or hairpin formation. Figure S4: Recombineering efficiency correlation between replicates. Figure S5: Simulation of effective library size *in silico*. Figure S6: Positional bias correlation with double mutations. Figure S7: 'Zone of Exclusion' for Rounds 1,3, and 5. Figure S8: Design of multiplexed library oligonucleotides. Figure S9: Emission scan of purified iLOV and tLOV. Figure S10: Thermostability of iLOV and tLOV via differential scanning calorimetry. Table S1: PR oligonucleotides for comprehensive iLOV library. Table S2: PR oligonucleotides for multiplexed thermostability library. Table S3: PCR primers used for Illumina sequencing. Table S4: Nucleotide and amino acid sequences of iLOV and tLOV.

AUTHOR INFORMATION

Corresponding Author

Email: savage@berkeley.edu. Address: 2151 Berkeley Way, Berkeley, CA 94704; (510) 643-7847

Author Contributions

S.A.H. and D.F.S designed the research. S.A.H. and S.O. performed the experiments. S.A.H. performed the computational analysis and analyzed the data. S.A.H. and D.F.S wrote the paper. Reagents described in this work are available on Addgene (https://www.addgene.org/David_Savage/).

Notes

The authors declare no competing financial interest.

560 **ACKNOWLEDGMENTS**

561 We thank H. Wang (Columbia) for providing the *E. coli* strain EcNR2. We
562 would like to thank A. Flamholz and B. Oakes for productive discussions and
563 readings of the manuscript, and C. Cassidy-Amstutz for assistance with *in vitro*
564 biochemistry. This work was supported by NIH New Innovator award
565 1DP2EB018658-01 to D.F.S., and S.A.H. was supported by NIH Training Grant
566 5T32GM066698-10 and Agilent.

567

REFERENCES

(1) Majithia, A. R., Tsuda, B., Agostini, M., Gnanapradeepan, K., Rice, R., Peloso, G., Patel, K. A., Zhang, X., Broekema, M. F., Patterson, N., Duby, M., Sharpe, T., Kalkhoven, E., Rosen, E. D., Barroso, I., Ellard, S., Kathiresan, S., O'Rahilly, S., Chatterjee, K., Florez, J. C., Mikkelsen, T., Savage, D. B., and Altshuler, D. (2016) Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet. 48*, 1570–1575.

(2) Heim, R., Cubitt, a B., and Tsien, R. Y. (1995) Improved green fluorescence. *Nature 373*, 663–664.

(3) Firnberg, E., and Ostermeier, M. (2012) PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLoS One 7*, e52031.

(4) Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. (2014) Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res. 42*, e112.

(5) Belsare, K. D., Andorfer, M. C., Cardenas, F., Chael, J. R., Park, H. J., and Lewis, J. C. (2016) A Simple Combinatorial Codon Mutagenesis Method for Targeted Protein Engineering. *ACS Synth. Biol. 6*, 416–420.

(6) Copeland, N. G., Jenkins, N. a, and Court, D. L. (2001) Recombineering: a powerful new tool for mouse functional genomics. *Nat. Rev. Genet. 2*, 769–779.

(7) Mosberg, J. a., Lajoie, M. J., and Church, G. M. (2010) Lambda red recombineering in Escherichia coli occurs through a fully single-stranded intermediate. *Genetics 186*, 791–799.

(8) Wang, H. H., Isaacs, F. J., Carr, P. a, Sun, Z. Z., Xu, G., Forest, C. R., and Church, G. M.

591 (2009) Programming cells by multiplex genome engineering and accelerated
592 evolution. *Nature 460*, 894–898.
593 (9) Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. a, and Gill, R. T.
594 (2010) Rapid profiling of a microbial genome using mixtures of barcoded
595 oligonucleotides. *Nat. Biotechnol. 28*, 856–862.
596 (10) Garst, A. D., Bassalo, M. C., Pines, G., Lynch, S. a, Halweg-Edwards, A. L., Liu, R.,
597 Liang, L., Wang, Z., Zeitoun, R., Alexander, W. G., and Gill, R. T. (2016) Genome-wide
598 mapping of mutations at single-nucleotide resolution for protein, metabolic and
599 genome engineering. *Nat. Biotechnol. 35*, 48–55.
600 (11) Thomason, L. C., Constantino, N., Shaw, D. V., and Court, D. L. (2007) Multicopy
601 plasmid modification with phage λ red recombineering. *Plasmid 58*, 148–158.
602 (12) Turrientes, M. C., Baquero, F., Levin, B. R., Martinez, J. L., Ripoll, A., Gonzalez-
603 Alba, J. M., Tobes, R., Manrique, M., Baquero, M. R., Rodriguez-Dominguez, M. J.,
604 Canton, R., and Galan, J. C. (2013) Normal Mutation Rate Variants Arise in a Mutator
605 (Mut S) Escherichia coli Population. *PLoS One 8*, e72963.
606 (13) Chapman, S., Faulkner, C., Kaiserli, E., Garcia-Mata, C., Savenkov, E. I., Roberts, A.
607 G., Oparka, K. J., and Christie, J. M. (2008) The photoreversible fluorescent protein
608 iLOV outperforms GFP as a reporter of plant virus infection. *Proc. Natl. Acad. Sci. U.*
609 *S. A. 105*, 20038–20043.
610 (14) Christie, J. M., Hitomi, K., Arvai, A. S., Hartfield, K. a., Mettlen, M., Pratt, A. J.,
611 Tainer, J. a., and Getzoff, E. D. (2012) Structural tuning of the fluorescent protein
612 iLOV for improved photostability. *J. Biol. Chem. 287*, 22295–22304.
613 (15) Lim, S. I., Min, B. E., and Jung, G. Y. (2008) Lagging Strand-Biased Initiation of

614 Red Recombination by Linear Double-Stranded DNAs. *J. Mol. Biol. 384*, 1098–1105.

615 (16) Callis, P. R., and Liu, T. (2006) Short range photoinduced electron transfer in

616 proteins: QM-MM simulations of tryptophan and flavin fluorescence quenching in

617 proteins. *Chem. Phys. 326*, 230–239.

618 (17) Mukherjee, A., Weyant, K. B., Agrawal, U., Walker, J., Cann, I. K. O., and

619 Schroeder, C. M. (2015) Engineering and characterization of new LOV-based

620 fluorescent proteins from chlamydomonas reinhardtii and vaucheria frigida. *ACS*

621 *Synth. Biol. 4*, 371–377.

622 (18) Tokuriki, N., and Tawfik, D. S. (2009) Stability effects of mutations and protein

623 evolvability. *Curr. Opin. Struct. Biol. 19*, 596–604.

624 (19) Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014) A

625 comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol. 31*,

626 1581–1592.

627 (20) Heinzelman, P., Snow, C. D., Wu, I., Nguyen, C., Villalobos, A., Govindarajan, S.,

628 Minshull, J., and Arnold, F. H. (2009) A family of thermostable fungal cellulases

629 created by structure-guided recombination. *Proc. Natl. Acad. Sci. U. S. A. 106*, 5610–

630 5615.

631 (21) Olson, C. A., Wu, N. C., and Sun, R. (2014) A comprehensive biophysical

632 description of pairwise epistasis throughout an entire protein domain. *Curr. Biol. 24*,

633 2643–2651.

634 (22) Nadler, D. C., Morgan, S.-A., Flamholz, A., Kortright, K. E., and Savage, D. F.

635 (2016) Rapid construction of metabolite biosensors using domain-insertion

636 profiling. *Nat. Commun. 7*, 12266.

637 (23) Engler, C., Kandzia, R., and Marillonnet, S. (2008) A one pot, one step, precision

638 cloning method with high throughput capability. *PLoS One 3*, e3647.

639 (24) Wang, H., and Church, G. (2011) Multiplexed genome engineering and

640 genotyping methods applications for synthetic biology and metabolic engineering.

641 *Methods Enzym. 498*, 409–426.

642