ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

Speeding up Directed Evolution: Combining the Advantages of Solid-Phase Combinatorial Gene Synthesis with Statistically Guided Reduction of Screening Effort

Sabrina Hoebenreich,[†,‡,∥] Felipe E. Zilly,[†,∥] Carlos G. Acevedo-Rocha,[†,‡] Matías Zilly,[§] and Manfred T. Reetz*[,†,‡]

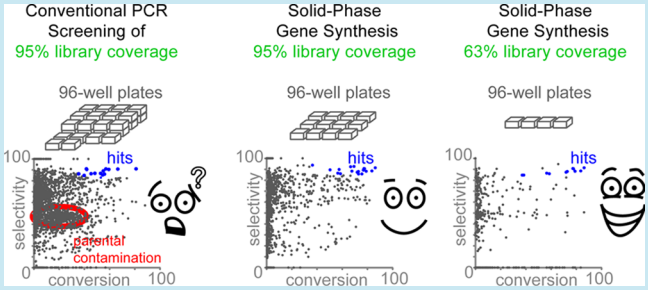
[†]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

[‡]Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Straße, 35032 Marburg, Germany

[§]Fakultät für Physik, Universität Duisburg-Essen, Lotharstraße 1, 47048 Duisburg, Germany

Supporting Information

ABSTRACT: Efficient and economic methods in directed evolution at the protein, metabolic, and genome level are needed for biocatalyst development and the success of synthetic biology. In contrast to random strategies, semirational approaches such as saturation mutagenesis explore the sequence space in a focused manner. Although several combinatorial libraries based on saturation mutagenesis have been reported using solid-phase gene synthesis, direct comparison with traditional PCR-based methods is currently lacking. In this work, we compare combinatorial protein libraries created in-house via PCR versus those generated by commercial solid-phase gene synthesis. Using descriptive statistics and probabilistic distributions on amino acid occurrence frequencies, the quality of the libraries was assessed and compared, revealing that the outsourced libraries are characterized by less bias and outliers than the PCR-based ones. Afterward, we screened all libraries following a traditional algorithm for almost complete library coverage and compared this approach with an emergent statistical concept suggesting screening a lower portion of the protein sequence space. Upon analyzing the biocatalytic landscapes and best hits of all combinatorial libraries, we show that the screening effort could have been reduced in all cases by more than 50%, while still finding at least one of the best mutants.



KEYWORDS: *protein engineering, metabolic engineering, genome engineering, screening effort, biocatalysis, cytochrome P450*

Chemical gene synthesis has become an essential tool in biotechnology and synthetic biology in particular for the manipulation of proteins,[1] metabolic pathways[2] and entire genomes.[3] Gene synthesis methods can be generally divided into polymerase cycling assembly (PCA)-dependent and independent ones, both depending on repeated cycles of conventional phosphoramidite chemistry consisting of base deprotection, coupling, capping, and oxidation.[4] On the one hand, PCA methods rely on the assembly of large DNA molecules from short fragments taking the advantages of DNA hybridization, annealing, and extension via the polymerase chain reaction (PCR). However, due to the use of naturally error-prone polymerases, the introduction of mistakes during the assembly processes cannot be avoided, among other disadvantages.[5] On the other hand, PCA-independent methods are based on the synthesis of oligonucleotides on solid supports followed by their assembly using enzymes involved in DNA repair and/or ligation instead of polymerases.[4] More recently, DNA microarrays have attracted increasing interest because unique sequences can be readily synthesized on a miniaturized chip with the advantages of reduced reagent consumption and high-throughput automatization.[5] In fact, several combinatorial libraries based on microarrays have been reported for engineering gene enhancers[6] and promoters,[7−10] plasmid stability,[11] or protein expression,[7,12] all of these examples being based on the Agilent technology.[13] However, the size of oligonucleotides using microarray technologies is still limited to about 200 bases so that their assembly into larger fragments is difficult and various errors can occur along the cloning process. Alternatively, two other approaches emerged in recent years: The "Blue Heron" solid support[14] and the "Sloning" building block[15] technologies. In the former, an oligonucleotide is covalently attached onto a solid support and assembled with bridging oligonucleotides to form an extended portion of the target polynucleotide. The cycle is repeated until the desired polynucleotide is completed, followed by its release upon denaturation. The Sloning technology uses a solid phase with a bound biotin-modified anchor oligo. To generate a target gene

Received: April 22, 2014
Published: June 12, 2014

sequence, a defined number of chemically synthesized building blocks or "splinkers" containing self-complementary regions are ligated to the anchor, immobilized, washed and cleaved, generating a subfragment of 18 bp in each cycle. The power of these two technologies is that full-length or large gene fragments can be produced in a fully automated and cost-effective manner. Furthermore, in contrast to PCA-based techniques, the solid-phase Blue Heron and Sloning platforms are believed to enable a faster and more accurate synthesis of genes of any complexity including GC-rich regions and repeat-intensive sequences. Although Blue Heron has the potential to generate libraries for directed evolution, only examples of the optimization of synthetic genes for heterologous gene expression are known (Blue Heron; personal communication). In contrast, the Sloning technology has been used successfully in protein engineering for the creation of libraries for expanding the genetic code[16] as well as for engineering binding affinity in anticalins[17] and antibodies.[18,19] In the production of these solid-phase combinatorial libraries,[16−19] primarily saturation mutagenesis was used, which allows the randomization of target gene sequences in a focused manner. We therefore concluded that Sloning has the potential to be applied in directed evolution of enzymes as biocatalysts in organic chemistry and biotechnology.

Several studies have shown that saturation mutagenesis is more efficient than error-prone PCR (epPCR) and DNA shuffling in the quest to evolve stereoselective and thermostable enzymes,[20] especially when executed iteratively (ISM).[21−23] Traditionally, when using PCR to create combinatorial libraries, the most common saturation mutagenesis technique employs the QuikChange protocol,[24] in which sense/antisense primers carrying degeneracies such as NNK (N = A/T/G/C; K = G/T) are used to introduce the mutations at a specific gene codon(s). Alternatively, primer mixtures according to the 22c-Trick[25] or 20c-Tang technique[26] encoding for all 20 amino acids can be used. Another saturation mutagenesis technique is based on MegaPrimer, in which one or multiple degenerate primers are used to introduce mutations at two distal sites by first creating a mutated large primer that is used in subsequent PCR cycles to extend the rest of the gene sequence, usually a plasmid.[27] More recently, the OmniChange protocol was introduced,[28] also based on degenerate but chemically modified primers that allow the introduction of up to 5 amino acid changes after many PCR cycles, a chemical treatment and hybridization. These may be too many steps for practical applications, especially when targeting a larger number of codons. To address these issues, we developed an extended form of the "Assembly of Designed Oligonucleotides" (ADO), which is a PCA-based two-step reaction method in which the degeneracy is not encoded in the amplifying primers, but encoded in primers that are used as backbones for the entire gene synthesis. Using ADO we targeted simultaneously 16 residues for saturation mutagenesis with various reduced amino acid alphabets.[29] Although successful in many cases, the problem with QuikChange, MegaPrimer, ADO, and any PCA-based gene synthesis tool is not only the likely introduction of errors in the PCR step but also that some templates are very "difficult-to-diversify".[30] The efficiency of these saturation mutagenesis techniques is dictated by several factors, including the position of the target codon and gene sequence, the length and GC content of the gene and the respective primer as well as the annealing temperature and quality of the primers in terms of randomized bases.[25] Indeed, in our hands the QuikChange protocol has failed many times. Even using optimized protocols for tricky templates based on MegaPrimer[30] and Gibson Assembly[31] methods, PCA-based tools often create a biased library. On the other hand, thus far the potential advantages of the Sloning technology has been realized only for challenging gene sequences compared to traditional approaches.[32] Unfortunately, the lack of a direct comparison between PCA-dependent and -independent libraries hinders the researcher to recognize the advantages and disadvantages of each technology, not only in practical terms regarding randomization efficiency but also taking into account direct and indirect costs. Moreover, it is not clear how the success rate of encountering improved mutants is linked to the amount of performed library coverage. Is screening of almost full coverage better or is a lower coverage equally successful, or alternatively how much lower?

In this study, taking as model enzyme the self-sufficient monooxygenase cytochrome $P450_{BM3}$ from *Bacillus megaterium* as biocatalyst, we compared, in terms of quality and screening effort, two sets of three combinatorial gene libraries, created using either an optimized MegaPrimer approach[30] or the solid-phase Sloning technology. Furthermore, we calculated the library sizes for 95% library coverage using the equation published by Patrick and Firth,[33] followed by creation of libraries, screening and comparison of the resulting data sets. Beside the direct comparison between PCA-dependent and -independent libraries, the question was addressed if and to what extent library coverage can be reduced while maintaining the probability to find the best or a second, third or nth best variant that might be equally sufficient for practical applications.[34]

RESULTS AND DISCUSSION

The model reaction chosen for library performance is the hydroxylation of testosterone, regio- and stereocontrol being a challenge in synthetic organic chemistry. Although wild type $P450_{BM3}$ does not oxidize steroids, mutant F87A shows an approximately 1:1 ratio of the monohydroxylated products 2β- and 15β-hydroxy testosterone (2β-OHT and 15β-OHT, respectively) (Figure 1A).[35]

Library Design. To evolve user-defined selectivity in enzymes, that is, chemo-, regio-, and enantioselectivity, we developed the Combinatorial Active-Site Saturation Test (CAST),[36] which is a strategy to target amino acids lining, and/or within, the active site of enzymes, usually based on the 3D structure but not restricted to it thanks to the power of structure homology models. The basic idea of CASTing is to generate small yet smart combinatorial libraries at sites composed of two or more amino acid residues that will be targeted simultaneously via saturation mutagenesis. In the case of $P450_{BM3}$, more than 25 amino acid positions are potential targets, from which we selected seven of the most relevant residues based on experimental evidence as explained elsewhere.[37] These seven residues were assigned to three different groups as shown in Figure 1B: Site A (R47/T49/Y51), B (V78/A82), and C (M185/L188).[35]

Library A and C were randomized using the same alphabet consisting of 12 amino acids (R, D, N, C, G, H, I, L, S, V, F, and Y) for reducing the screening effort.[39] These amino acids are encoded by the degenerate and nonredundant NDC codon that can be used in PCR-based saturation mutagenesis methods (Figure 1C). The Sloning codon choice for the same set of 12 amino acids differs only in the last nucleotide of 5 codons due to a codon optimization algorithm for *Escherichia coli* (Figure

dx.doi.org/10.1021/sb5002399 | ACS Synth. Biol. 2015, 4, 317−331

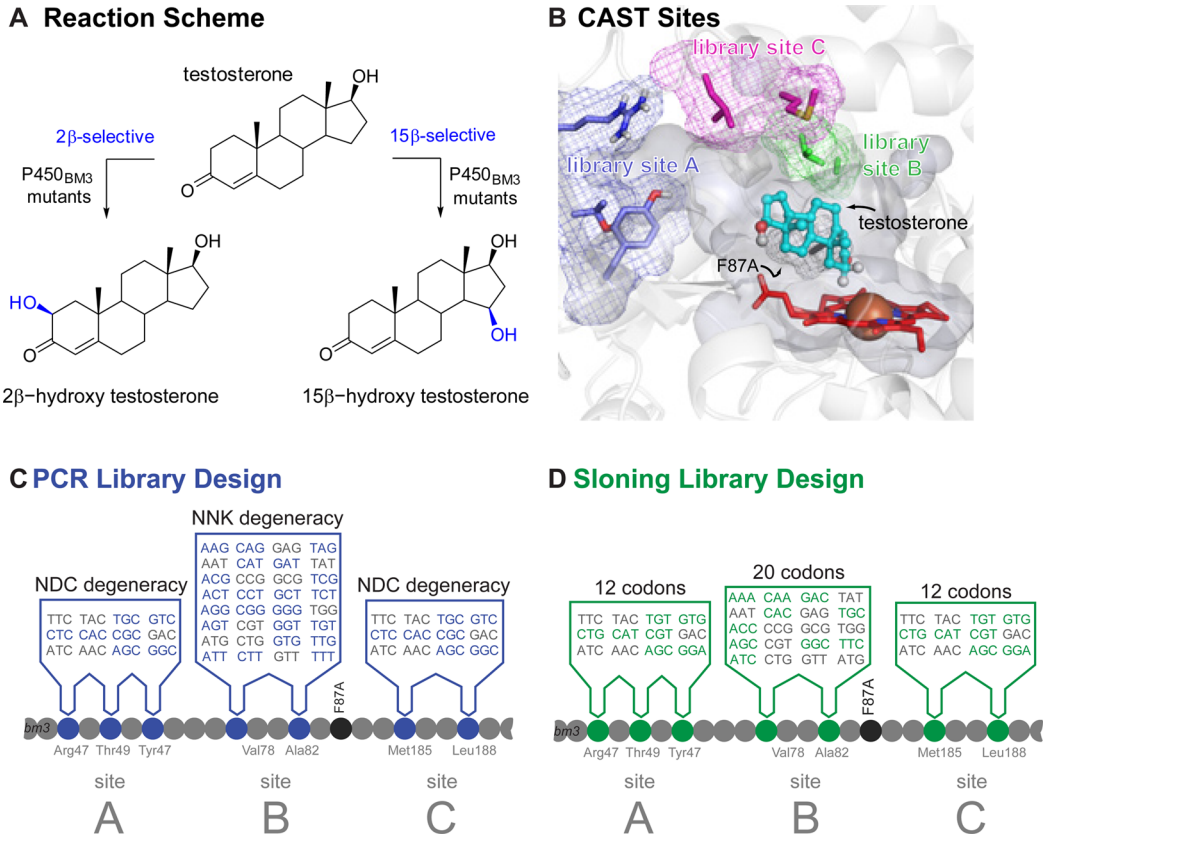


Figure 1. Model reaction and library design. (A) Testosterone hydroxylation by $P450_{BM3}$ mutants. (B) Active site of $P450_{BM3}$ mutant F87A. The three CAST sites and the F87A residue are highlighted. The structure was prepared with Schrödinger software (see Methods) and picture was created with PyMol.[38] (C) Diversity design of the combinatorial $P450_{BM3}$-F87A libraries used in this study. Library A consists of three simultaneously randomized positions, whereas library B and C consist of two. PCR-based libraries use either the nonredundant NDC codon (library A+C) or the redundant NNK codon (library B). (D) Sloning-based libraries encode the same set of amino acids using the displayed codon usages. Gray codons are present in both designs.

1D). In the case of Library B, the randomization scheme includes all 20 canonical amino acids, which can be covered with the degenerate NNK codon for PCR-based saturation mutagenesis methodologies. NNK encodes 32 defined codons, including one stop codon, but bearing redundancy for the amino acids A, G, P, T, V and R, L, S with two and three codons, respectively. Sloning gene synthesis does not depend on degenerate codons, thus full randomization can be achieved with a nonredundant set of 20 codons (Figure 1D).

Library Construction and Quick Quality Control (QQC). The three PCR-based libraries were created using an optimized MegaPrimer PCR protocol,[30] as described earlier.[35] Since the target codons lie adjacent to each other in all cases, the use of a single primer, encoding the desired degeneracy, in combination with a nondegenerate primer is sufficient for megaprimer formation. Initial library creation trials did not result in any colonies; after optimizing primer length and reaction conditions, few (<100) colonies were obtained, however, without acceptable degeneracy. Finally, adjusting the annealing temperature and increasing the number of amplification cycles led to libraries with sufficient amounts of colonies per transformation (>10 000). Pooling these colonies followed by plasmid extraction and sequencing displayed acceptable diversities as judged by the "Quick Quality Control" (QQC)[25,40] (see Supporting Information (SI)). As discussed earlier,[25] the application of the QQC prior to screening is a very useful trick to avoid waste of time and resources for screening a library without the aimed diversity. The final PCR-based library A (A-PCR) showed acceptable degeneracy at all positions, but for the PCR-based library B (B-PCR) it was still not possible to randomize position Val78 as desired (SI Figure S1), where type AXX and XCX codons remained underrepresented. PCR-based library C (C-PCR) showed the lowest quality, especially at position Met185, where codons GGC (Gly), GAC (Asp) and GTC (Val) were also under-represented, due to the almost entire absence of G in the first nucleotide position (SI Figure S1). All libraries seemed to be free of primer misinsertions, but different amounts of parental sequences were found according to the QQC: Library A-PCR showed generally low (10−15%) while B-PCR medium (20%) and C-PCR high (25%) parental codon contents (SI Figure S1). Newly amplified template sequence originating from the PCR reaction is defined here as parental sequence and cannot be removed by DpnI digestion. In addition, control experiments yielded no *E. coli* colonies upon transformation with *DpnI*-digested, methylated template DNA. Therefore, we do not believe that the presence of parental sequences in our PCR-libraries is due to an inefficient *DpnI* treatment. The most likely reason for observing parental sequences is due to the use of a nondegenerate primer that forms a certain amount of nonmethylated, nonmutated megaprimer, resulting in parental sequences that cannot be *Dpn*I-digested. In addition, the chosen degeneracy for positions T49 and M185 lacks parental codons and consequently library A and C should not contain any parental sequences. In

dx.doi.org/10.1021/sb5002399 | *ACS Synth. Biol.* 2015, 4, 317−331

Table 1. Results of 96-Well Plate Sequencing, which Represent a Random Sample Data Set for Each Library

	A		B		C	
	PCR	Sloning	PCR	Sloning	PCR	Sloning
samples sequenced	96	96	96	96	96	96
controls on plate	2	4	2	2	2	0
failed sequencing	2	1	3	3	27	3
deletions	6	5	2	0	0	8
misc. problems	5[f]	3[g]	3[f]	0	3[h]	6[f]
parental sequences[a]	11	0[k]	18	0[k]	10	0[k]
stop codons	0[k]	0[k]	2	0[k]	0[k]	0[k]
non-designed codons	7[b]	1[c]	0	3[d]	2[e]	0
no. of randomized sequences	70	83	68	91	54	79
yield (single seq)[i]	76%	91%	75%	100%	81%	85%
yield (QQC)[j]	~85%	~95%	~80%	~90%	~75%	~95%

[a]$P450_{BM3}$ mutant F87A was used as template. [b]In R47: TGT(Cys), TCT(Ser); in T49: Ser (TCC), 2× Thr (ACC), Leu (CTG); in Y51: Leu (CTG). [c]In R47: Arg (CGG). [d]In V78: Gln (CAG), in A82: Arg (CGG), Phe (TTT). [e]In Met188: Arg (CGG), His (CAT). [f]Deletions or insertions of one or two bases at one of the saturated positions. [g]One 31 bp insertion and two in frame codon deletion at target site. [h]A single base insertion at codon 188, a deletion of 20 bp starting at residue 188 and one double sequence at the addressed positions. [i]Considering parental content, frame shifts, deletions, etc. as side products of the PCR reaction. For example, 96 − 2 − 2 = 92 subtotal samples, 11 + 6 + 5 = 22 nondiversified transformants and junk constructs; 92 − 22 = 70 randomized samples or 76% yield. [j]Yield of main plasmid construct as estimated from the QQC data. [k]Not part of the probed sequence space, due to the used randomization scheme.

contrast, the parental codon is included in the target randomization for both positions of library B. Thus, although the MegaPrimer method is useful, it cannot provide fully bias-free libraries as judged by the QQC. Overall, the time for producing these medium-high quality libraries including primer design, PCR optimization, digestion, and transformation trials as well as QQC analysis amounted to 3 months.

In the case of the Sloning libraries, gene fragments of 683 bp (954 bp with flanking sequences) were synthesized and cloned into the target plasmid (see Methods), followed by host transformation and delivery as glycerol stocks. The delivered library stock contained 5000 clones for library A (A-SLO), 11000 for library B (B-SLO) and 5000 for library C (C-SLO). Due to the randomization scheme of the Sloning libraries and the limitations of the QQC, the results were too ambiguous for conclusions concerning the amount of parental contamination (see SI). Nevertheless, the QQC can reveal other problems such as multiple DNA-fragment insertions, wrong degeneracies, etc. The results of the Sloning libraries provide an indication that the designed degeneracy was not fully achieved in all positions but due to the nature of the QQC technique the results were not conclusive enough (SI Figure S1). On average, it took 2 months per library to be delivered. Library B took 4 months, since the first library failed the internal quality control and needed to be created a second time. This event supports our own observation that library B was difficult to be diversified and suggests that the AT rich region near the codons 78 and 82 of the *bm3* gene is generally more challenging to diversify than other positions, as experienced with ADO[29] and Gibson assembly[31] PCR-based methods.

Although it took twice as long to receive the three outsourced combinatorial libraries (6 versus 3 months of the homemade libraries), the former showed higher quality regarding the absence of parental sequences as judged by the QQC (SI Figure S1). Nevertheless, an advantage of the PCR-based libraries is that these could be optimized in parallel, which would be beneficial if screening requires only a few hours up to a few days. Conversely, the Sloning libraries were synthesized sequentially, but this was of no great disadvantage because the HPLC-based steroid hydroxylation screening process took days to a few weeks. It is important to bear in mind that these estimations are based on an experienced researcher, but longer times may be required for inexperienced ones or when dealing with "difficult-to-diversify" templates. On the other hand, if the researcher is experienced and is running several projects in parallel, it might be better to outsource the combinatorial libraries in order to save time. Importantly, all these factors have to be taken into account when assessing overall costs for obtaining one or various combinatorial libraries, which will depend on the project and whether success can be achieved at an early stage. Finally, another important factor is library quality in terms of expected diversity. In order to gain more insights on the codon and amino acid frequency distribution at each randomized position, we thus sequenced single clones of each library.

Library Quality and Diversity. All libraries were inserted into the host *E. coli* BL21-Gold(DE3), followed by plating and single colony picking in LB media using 96 deep-well plates. Due to the large costs of sequencing at the time when the project was running, only one (replica) plate per library was sent for plasmid extraction and sequencing analysis to GATC (Germany). The results are presented in Table 1. On average 1−3 sequencings failed per plate, but in the case of library C-PCR, 27 samples failed, most likely due to suboptimal growth conditions or poor plasmid extraction. Despite this drawback, we continued with the analysis using the available sequencing data. Importantly, none of the Sloning library samples contained parental sequences, whereas 11, 18, and 10 parental sequences were found for the PCR-based data sets A, B, and C, respectively (Table 1). In the PCR sequencing data set, two sequences encoded a stop codon at residue 82. Additionally, besides a few frame-shifts, caused by single base deletions or insertions at the targeted codons, we also observed partial gene deletions in C-SLO library, where in-frame residues 2−251 of BM3 were missing. Surprisingly, we observed the appearance of nonexpected codons mostly in the PCR-based libraries, but to a lesser extent in the Sloning ones. For example, the C-PCR data set contained the nondesigned $P450_{BM3}$ mutants M185M-(ATG)/L188R(CGG)/F87A and M185N(AAC)/L185L-(CTG)/F87A, but three of these codons (i.e., ATG, CGG,

dx.doi.org/10.1021/sb5002399 | ACS Synth. Biol. 2015, 4, 317−331

and CTG) are not part of the NDC degeneracy. These findings could be interpreted as to arise from pairing of two different single-strand megaprimers, one originating from the non-degenerated primer (carrying no mutation, since ATG and CTG are the parental codons), and one from a mutated strand and misread upon amplification *in vivo*. Similar observations were made in the B-PCR data set, where the parental GCA (Ala) codon remained at position 82 in four sequences, but position Val78(GTA) was randomized and yielded TTT(Phe), CAT(His), GTT (Val), and GTG (Val) variants. Here, the observed codons in position 78 were included in the used NNK degeneracy. Finally, some Sloning libraries also contained unexpected codons. For instance, codons TTT, CGG, and CAG were found in the B-SLO data set, while CGG was found in that of A-SLO. The reason for the presence of these unexpected codons is unknown.

Although the same number of clones per library was sent for sequencing, the yield of successfully randomized constructs was higher for the Sloning data set than for the PCR-based one: 83 (A-SLO), 91 (B-SLO) and 79 (C-SLO) versus 70 (A-PCR), 68 (B-PCR) and 54 (C-PCR). Considering parental sequences and other "junk" constructs as side products, the yields of randomization amount to 76%, 75% and 81% for the PCR libraries A, B, and C and to 91%, 100% and 85% for the Sloning ones, respectively (Table 1). Both quality analyses, the random single sequencing and the QQC, suggest that the Sloning libraries are better randomized than the PCR-based ones. This conjecture is reasonable because undesired parental sequences contaminations are absent and yields of correctly mutated constructs are higher. To assess in more detail the diversity on codon and amino acid level, we subjected the results from single sequencing to descriptive statistics and probability distribution analysis.

Absolute codon occurrence per residue was first analyzed by box plotting, including the nondesigned codons/amino acids. Codons and resulting amino acids from parental constructs were excluded, to assess the created diversity devoid of the background of parental contamination. Figure 2A shows that Sloning data sets have higher randomization efficiency per targeted residue than the PCR data sets, because they exhibit a slightly higher mean amino acid appearance. Important for library quality is the nonzero minimum amino acid appearance. The analyzed random samples of the Sloning libraries met this criterion, with V78 being the only exception. However, the picture is different for the PCR data set, where six out of seven positions miss one or more amino acids completely. While this may not hold true for the complete library considering the limited number of sequences per library, it certainly means that those amino acids have a very low probability to appear in the respective library. Furthermore, several outliers were found within the sequencing data sets: three in the PCR case, but only one within the Sloning data set. Application of the Grubbs test confirmed that the occurrence of 20 phenylalanines at position 47, 10 asparagines at 188 and 15 glycines at 185 are indeed statistical outliers in the PCR-based libraries, whereas the occurrence of 10 asparagines at position 82 was statistically possible, when assuming that codon incorporation during PCR was nonparametric. Although our data sets are relatively small and based on limited sequencing results, the boxplot analysis indicates that the Sloning libraries may have a more diversified population of amino acids compared to the PCR libraries.

These findings were confirmed by statistical analysis of the sequencing data sets via the χ-square test (Figure 2B). A common assumption in PCR-based library creation is that the annealing behavior of all diversified primer constructs leads to a uniform distribution of the designed sequences. For NDC degeneracy (sites A and C), this results theoretically in a uniform distribution for the amino acids at protein level, too, whereas for NNK degeneracy (site B) those amino acids that are encoded manifold should appear with higher probability. We used these resulting distributions as null hypothesis for the χ-square analysis of the data sets of the PCR libraries. For the Sloning libraries the null hypothesis is a uniform distribution of the targeted amino acids. If the resulting χ-square probabilities are $\leq$5%, the null hypothesis is rejected; that is, we have evidence for nonuniform data sets and consequently a biased library sample. Figure 2B shows that six out of seven PCR data sets are extremely biased, with a probability of <1% that such an observed distribution occurs by chance. Only position Y51 from the A-PCR sample was found to exhibit the aimed diversity. In contrast, four positions of the Sloning libraries passed the test, except positions Y51 (0.6%), V78 (1.4%), and L188 (3.6%). We have likewise analyzed the sequencing results from random samples taken from the Sloning libraries prior to library delivery, and found that positions R47 (3.3%) and L188 (3.2%) did not pass our χ-square test (SI Figure S2A), but it should be noted that less samples (~50) were considered, which is borderline for a statistically solid conclusion. Overall, the χ-square test points to the fact that the Sloning libraries are closer to the desired diversity than the PCR ones.

Since a rigorous χ-square test requires that each amino acid is represented at least 3−5 times in the random sample, our sequencing sample sizes (see Table 1) are at the lower limit for a robust statistical analysis to reject the null hypothesis. However, on the one hand the sequencing costs were limiting at the onset of this study and, on the other hand, an appropriate sample size can be determined only after having a first insight into the randomization efficiency. In any case, some amino acids are found far too seldom, while others too often for allowing us to classify the libraries as uniform. This contrasts with what often is assumed,[33,34,41,42] namely that "... *in saturation mutagenesis, the randomization at the DNA level at each position is typically uniform...*".[34] Yet the data of the present sequencing analysis suggests that this assumption is false, especially for the PCR library samples but also in a few Sloning cases. We attribute this lack of uniformity to intrinsic variables such as the gene sequence secondary structure, target sequence, and coupling efficiency of bases during primer synthesis. In particular, the thermodynamic and kinetic effects during hybridization of the template and different target sequences might cause the nonuniform randomization.[43,44] Importantly, Figure 2B suggests that the parent $P450_{BM3}$ mutant F87A is more difficult to diversify via PCR than with Sloning. Thereafter, we analyzed in detail the amino acid frequency of each library data set.

Figure 2C−H breaks down the observed distributions per randomized position to the identity and occurrence frequencies of the amino acids, revealing different degrees of bias in the sequencing data sets. For example, Gly and Arg containing variants in the C-PCR sample are rare, whereas these are very well represented in the C-SLO sample. The statistical outliers R47F (A-PCR), A82N (B-PCR), and L188N (C-PCR) are particularly visible (Figure 2C−E). Generally, the three Sloning data sets show a smoother amino acid distribution compared to the PCR-based data set. However, in samples B-PCR and B-SLO, there is an interesting pattern: P (CCG, CCT), Q

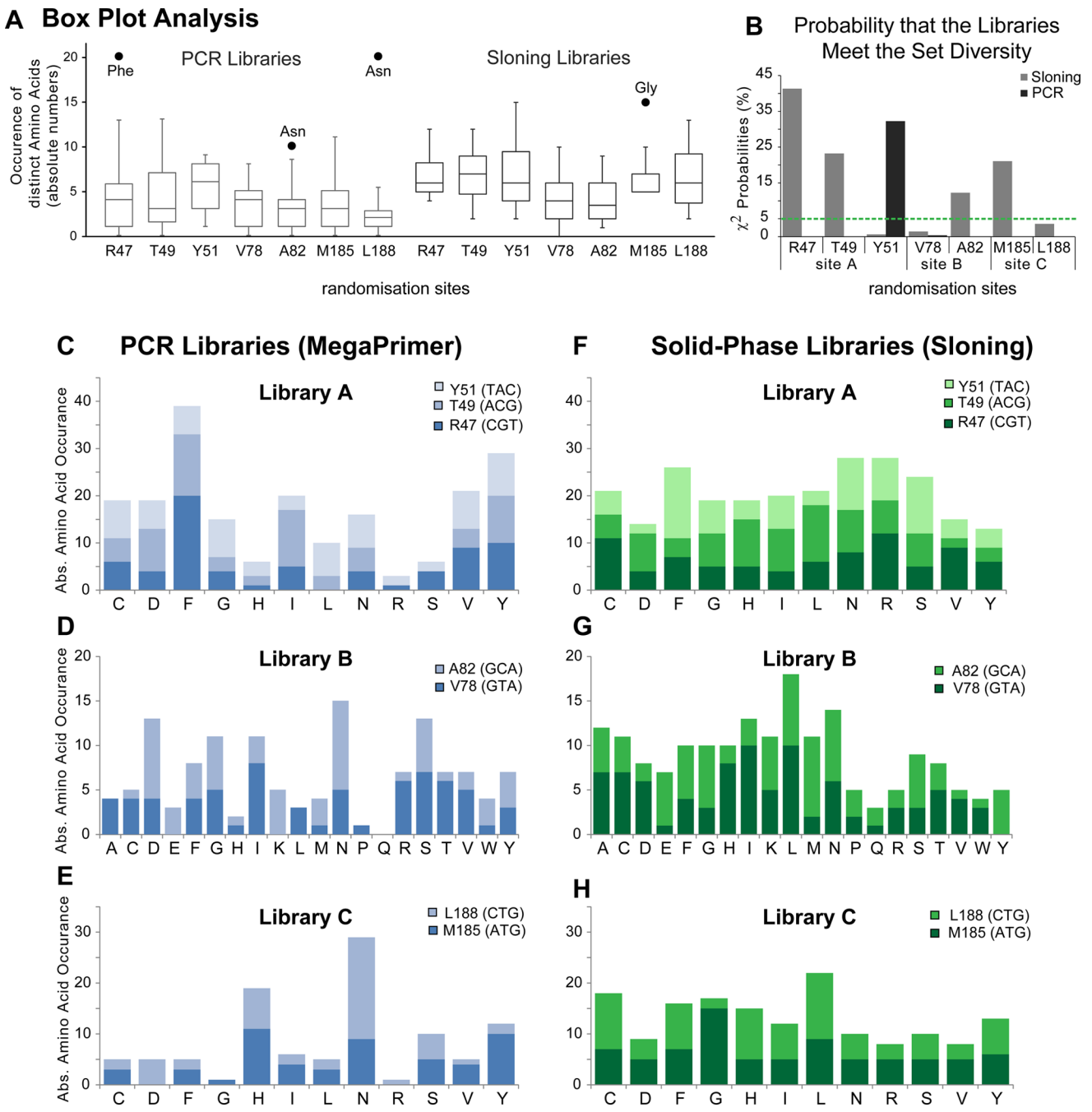


Figure 2. Statistical analyses of the sequencing data sets. (A) Boxplot analysis of absolute amino acid occurrence observed per randomized position for PCR and Sloning libraries data sets. (B) χ-Square test analyzing whether the desired amino acid (aa) frequencies were obtained. For positions with a probability below 5%, the hypothesis is rejected and the set diversity was not created. (C–E) PCR data sets (blue) of absolute amino acid occurrence per mutated position in each of the random library samples, respectively containing 68/66/69 sequencing results for library sample A (R47/T49/Y51), 68/62 for B (V78/A82), and 53/50 for C (L188/M185). (F–H) Sloning data sets (green) containing 83/83/83, 90/89, and 79/79 samples from A, B, and C, respectively. Observed nondesigned amino acids and the two stop codons (B-PCR) are excluded for simplicity. The corresponding codons are summarized in the Supporting Information (Table S1 + S2).

(CAG), and E (GAG, only in position 78) are underrepresented amino acids. Although one may attribute these findings to mere probabilities, it is likely that the "difficult-to-diversify" $P450_{BM3}$ gene limits the introduction of the target codons at the respective positions V78(GTA) and A82(GCA) due to the amplification process either *in vitro* during the PCR (first case) or *in vivo* (when amplified, transcribed, and translated by *E. coli's* machinery). These findings indicate that the Sloning libraries are of greater quality than the PCR ones, thus fulfilling the most important requirement in library design and construction: a more balanced and complete coverage of the targeted sequence space.

Library Screening Effort. The statistical analysis and probabilistic distributions of the library samples show that the Sloning libraries are better than the PCR-based ones, but an extrapolation to the complete library size is not appropriate, given the limited amount of variants sequenced. For this reason, and instead of sequencing more samples, we opted to screen

dx.doi.org/10.1021/sb5002399 | *ACS Synth. Biol.* 2015, 4, 317–331

Table 2. Number of Theoretical Gene and Protein Variants, Library Sizes and Effort Savings, as Well as Fractional Completeness under Ideal Conditions of the Six Combinatorial Libraries

	A		B		C	
	PCR	Sloning	PCR	Sloning	PCR	Sloning
degeneracy	NDC	12 codons	NNK	20 codons	NDC	12 codons
no. of gene/protein variants	1728	1728	1024/400	400	144	144
library size[a]	5076	3864	3008	1222	752	672
savings (Sloning over PCR)		24%		59%		11%
fractional completeness (library coverage on protein level)[b]	95%	89%	98%	95%	99%	99%

[a]Excluding screened controls. [b]Nonredundant libraries have equal completeness on DNA and protein level, while the redundant NNK library has 95% completeness on DNA level and 98% on protein level. Value calculated using available online tools.[34,47]

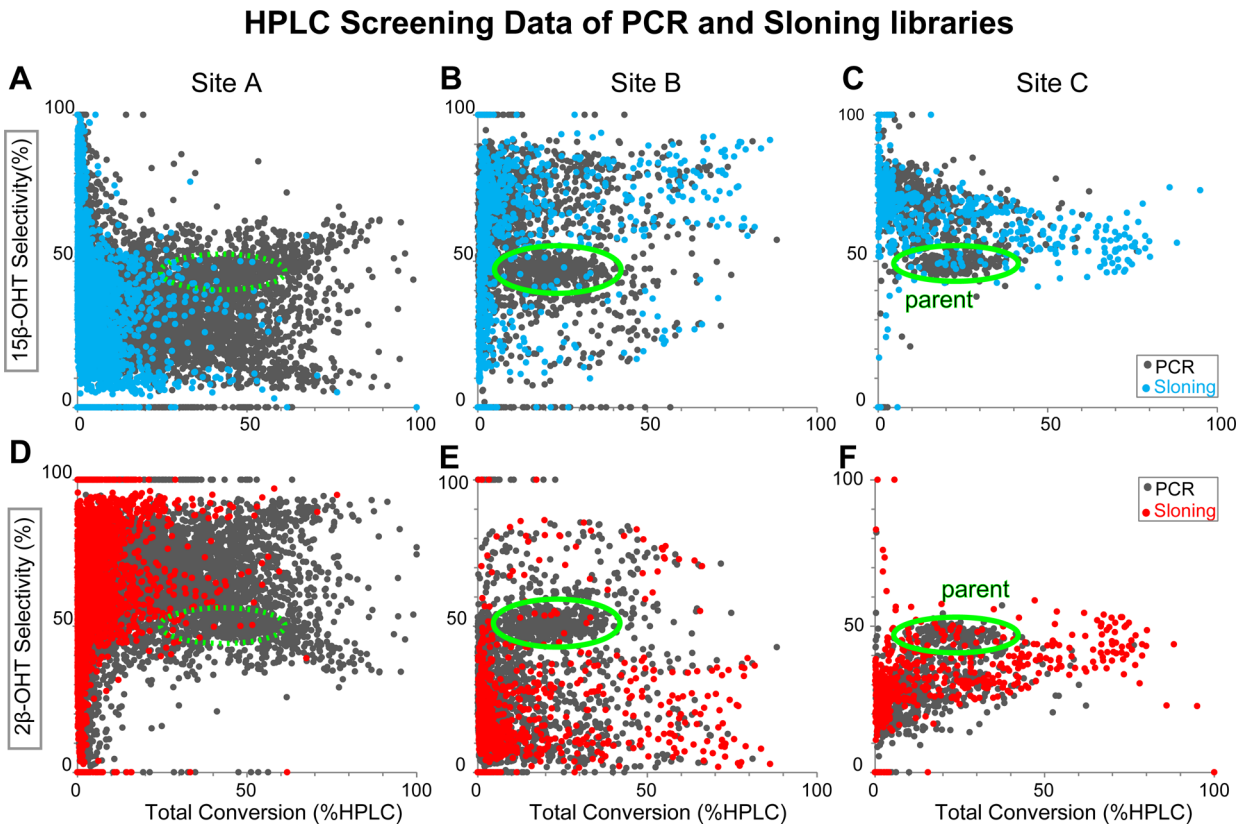


Figure 3. Library screening results. Total testosterone conversion (%HPLC) of the six combinatorial libraries is shown as a function of either 15β-OHT or 2β-OHT regioselectivity. Colored entries show the Sloning libraries data, while gray entries represent the PCR library results. The green circle highlights a cluster corresponding to parental transformants in PCR libraries.

each of the six libraries in order to determine and compare their biocatalytic landscapes. As noted above, $P450_{BM3}$ mutant F87A catalyzes the hydroxylation of testosterone with notable activity (20% total conversion as determined by HPLC analysis), but it is not regioselective (1:1 mixture of 2β- and 15β- hydroxylation products). Thus, the set goal is to find highly stereo- and regioselective mutants for 2β- or 15β-hydroxy testosterone formation (2β-OHT or 15β-OH, respectively) but also showing acceptable activities. Six libraries were created, whose sizes were calculated[33] aiming to cover a large portion (87−99%) of the targeted sequence space. The original Patrick and Firth equation was modified to take the parental contamination into consideration (see SI). This, in addition to limitations imposed by practical constraints,[45] resulted in 5076 (A-PCR), 3008 (B-PCR), and 752 (C-PCR) transformants per PCR-based library and 3864 (A-SLO), 1222 (B-SLO), and 672 (C-SLO) transformants per Sloning library (Table 2). These sizes correspond to library coverage on protein level of 95% and 89% (A libraries), 98% and 95% (B libraries), and 99% both (C libraries) for the PCR-based and Sloning cases, respectively. Importantly, despite comparable library coverage, the absolute library sizes differ very much: 1222 in B-SLO compared to 3008 transformants in B-PCR (Table 2). The latter screening effort is three times as high and results from the necessity of screening a redundant PCR-based library with only 400 variants at the protein, but 1024 at the gene level.

We screened in total 8836 and 5758 samples for the PCR-based and Sloning libraries, respectively (Table 2). Clearly, the effort invested in Sloning corresponds to 65% of the effort needed to screen the PCR-based libraries; thus, 35% of screening resources and time could have been generally saved. In particular, 24% of resources could be saved for library A-SLO compared to A-PCR, which corresponds to a total of 12 multititer plates (MTP), a total HPLC measuring time of 57.6

dx.doi.org/10.1021/sb5002399 | ACS Synth. Biol. 2015, 4, 317−331

h (3 min per sample; 288 min per MTP), 2 weeks of work and resources for preparing and processing the libraries. When comparing B-SLO with B-PCR, especially the removal of redundancy at the gene level contributes to the overall 59% screening effort savings, which amounts to 20 MTP plates and the respective preparation time: 96 h or 4 days. It should be noted, however, that the removal of NNK/S redundancy at the gene level can be achieved by mixing various primers, representing the latest trick for creation of smaller yet smarter libraries in directed evolution and protein engineering,[25,26] besides other emerging techniques.[46] Finally, due to the large oversampling of transformants necessary to reach 99% coverage in the C libraries, only 11% savings were achieved with the Sloning library in this case.

The library screening results are presented in Figure 3. Interestingly, the smaller Sloning libraries essentially show the same scattering pattern as the remaining entries. Regioisomers 2β-OHT and 15β-OHT were found as the dominant products in all six libraries, but 1β-OHT, 16β-OHT, and 19-OHT species were also observed in various amounts as side products (SI Figure S3B). Notably, the A-PCR library pattern is stretched horizontally, due to a longer reaction time (72 h) resulting in an overall higher conversion level (at that time this set of measurements was in an optimization process), compared to the standard reaction time of 24 h for the remaining five libraries. Despite this, selectivity values were generally not affected by the overall activity and therefore the scattering patterns are comparable.

$P450_{BM3}$ mutants from the A libraries show mostly 2β-OHT selectivities (Figure 3D), and the observed 15β-OHT selectivities did not exceed 60% (Figure 3A). In contrast, B libraries contain both 15β- and 2β-OHT producing mutants, with selectivities scattering more toward 15β-OHT (Figure 3B). Still, a few selective 2β-OHT producers were observed in the B libraries (Figure 3E), although these mutants show a slightly lower selectivity compared to those in the A libraries. In addition, more library B variants than those in library A exhibit an activity greater than 50%. Most notable are the two different scattering patterns of the C libraries. The Sloning library C contains a significant amount of mutants with high activity (>50% HPLC conv., Figure 3F) that are missing in C-PCR (Figure 3C). The selectivity values of library C variants are generally not satisfactory (around 60%), showing a small preference toward 15β-OHT.

Strikingly, in libraries B-PCR and C-PCR an obvious cluster of entries becomes apparent at around 50% selectivity and ~20% activity, which is absent in the Sloning libraries, and possibly hidden in library A-PCR owing to the longer reaction time. Comparison with the plot of controls (parent F87A, SI Figure S3A) leads to the conclusion that these clusters correspond to parental transformants (F87A). This analysis therefore demonstrates how contaminated with parental template (F87A) the "difficult-to-diversify" PCR-based libraries are. It also shows the advantage of synthesizing solid-phase combinatorial libraries, because the parental contamination in the PCR libraries needs to be compensated by a significantly increased screening effort to achieve the targeted sequence space coverage.

In summary, variations at residues R47, T49, and Y51 (library A) show a preference toward 2β-OHT-selectivity, whereas those at V78 and A82 have a greater influence on regioselectivity control, since selective variants producing either 2β-OHT or 15β-OHT were observed (Figure 3). The targeted residues M185 and L188 mainly influence activity as measured by total testosterone conversion. Finally, the two-parameter biocatalytic landscapes clearly show that mutants with high selectivity for hydroxylated testosterone regioisomers as well as testosterone conversion are present in both PCR-based and solid-phase libraries. Nevertheless, a closer look at each library is needed to assess in more detail the number and identity of those mutants to determine the advantage of each method.

Hit Analysis. Although higher randomization quality and absence of codon redundancy correlates with significant economic savings, the overall library sizes (~95% coverage) as calculated by the Patrick and Firth algorithm[33,41,47] are still rather labor intensive for most chromatography-based screening assays. Of course, lower library coverage than 95% could be chosen and screened, hoping that a good mutant is found, but the correlation between the probability of encountering that mutant and an arbitrarily chosen low library coverage is unknown. Recent work by Nov suggests that it is not necessary to strive for 95% library coverage, which aims to find *the best* mutant, because one out of the best two, three or *n*th mutants may be equally sufficient for practical applications.[34] Nov also argues that "*...existing criteria and practices for determining the library size in saturation mutagenesis experiments are often too conservative...*".[42] Therefore, to determine the effect of library coverage reduction on the probability of finding 1 out of the *n*th-best mutants, we assessed the relationship between the traditional Patrick–Firth and Nov metrics (Figure 4).

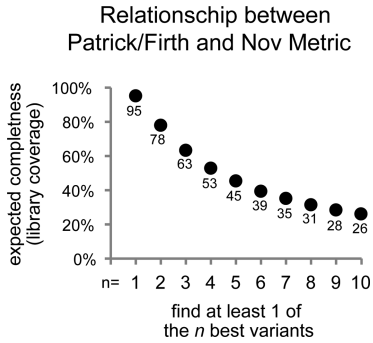


Figure 4. Patrick–Firth versus Nov statistical metrics. Relationship between the expected completeness (i.e., library coverage) algorithm by Patrick and Firth as computed by GLUE-IT,[47] and the concept of finding at least one of the *n*th-best variants (with a 95% probability) by Nov as computed by TopLib.[34] This mathematical relationship is independent of the number of positions randomized and of the randomization scheme.

Interestingly, aiming for finding *the best* mutant corresponds to 95% library coverage as shown elsewhere,[34] whereas finding at least 1 out of the second, third, fourth, and fifth best mutants corresponds to 78%, 63%, 53%, and 45% library coverage, respectively, regardless of library size and design (Figure 4). It is remarkable to observe that there are manifold publications in protein directed evolution where the effort invested in screening combinatorial libraries targeting 2 or more residues with medium-to-large amino acid alphabets is very diverse: from extremely large to extremely low library coverage (Table 3). From this data, it appears that improved mutants can be found without screening a large portion of any combinatorial library.

In order to determine whether >95% screening effort can be generally avoided while still having a high probability (95%) of

Table 3. Different Screening Efforts of Combinatorial Libraries Targeting at Least 2 Amino Acids with Saturation Mutagenesis as Calculated by TopLib[34] Assuming a 95% Probability and 100% Yield

no. targeted positions (no. libraries)	randomization alphabet(s)[a]	theoretical no. combinations of amino acids (codons)	no. required for 95% library coverage	screened numbers	library coverage performed	at least one of the *n* best Hits	ref.
3 (1)	WHK (9 aa, 1 stop codon) WHW (6 aa, 1 stop codon) RHK (10 aa/codons)	540 (700)	2096	14 330	99.99%	1	48
2 (1)	NNS (20 aa/32 codons) NDT (12 aa/codons)	240 (384)	967	1216	~97%	1	49
2 (5)	NDT (12 aa/codons)	144 (144)	430	504	97%	1	50
9 (1)	2 to 4 aa/codons	1024 (1024)	3067	2400	90%	2	51
2 (1)	NNK (20 aa/32 codons)	400 (1024)	2130	1500	90%	2	52
2 (1)	NNS (20 aa/32 codons)	400 (1024)	2130	747	74%	3	53
2 (1)	NNK (20 aa/32 codons)	400 (1024)	2130	470	~59%	4	54
3 (1)	NDT (12 aa/codons) NDT (12 aa/codons) YHC (6 aa/codons)	864 (864)	2587	600	~53	5	55
2 (1)	NNK (20 aa/32 codons)	400 (1024)	2130	300	~44%	6	54
5 (4)	5 to 6 aa/codons	5400 (5400) or 6480 (6480)	16176 or 19 411	1472	~20% or ~18%	>10	56
4 (1)	NNK (20 aa/32 codons)	160 000	1×10^6	5000	0.025%	>100	57

[a]W = A/T; H = A/C/T; K = G/; R = A/G; N = A/C/T/G; S = C/G; D = A/G/T.

Table 4. Amino Acid Identity, Occurrence at Different Library Coverage Stages, and Performance of Hits after Re-culturing of Selected Transformants of Libraries A[a]

mutant		PCR						Sloning					biocatalytic parameters	
library size		1036	1294	1726	2584	3864	5076	1036	1294	1726	2584	3864		
library coverage		45%	53%	63%	78%	89%	95%	45%	53%	63%	78%	95%		
at least 1 out of *n* best		5	4	3	2	2−1	1	5	4	3	2	1	2β-selectivity (%)	total conversion (%)
	47/49/51													
KSA-22	VLN	−	−	−	−	−	−	++	++	++	++	++	95.5 ± 1.3	44 ± 5
KSA-23	ILN	− −	− −	− −	− −	− −	++	−	−	−	−	−	94.4 ± 1.3	50 ± 10
KSA-24	VII	−	−	−	−	−	−	−	−	+	+	+	93.9 ± 1.1	54 ± 9
KSA-3	IIV	− −	− −	− −	+ −	+ −	++	−	−	−	−	−	93.9 ± 1.1	54 ± 11
KSA-25	IIN	−	−	−	−	−	+	−	−	−	−	−	93.7 ± 1.3	64 ± 9
KSA-2	III	+ −	+ −	+ −	++	++	++	+	+	+	+	+	93.7 ± 1.3	52 ± 12
KSA-26	VIN	− −	− −	+ −	+ −	+ −	++	−	−	−	−	−	93.7 ± 1.1	51 ± 12
KSA-27	VFN	+ −	+ −	+ −	+ −	+ −	++	−	−	−	−	−	92.4 ± 1.6	44 ± 14
KSA-28	NII	−	−	−	−	−	−	+ −	+ −	+ −	+ −	++	92.0 ± 1.8	63 ± 10
KSA-29	NLI	−	−	−	−	−	−	−	−	−	+	+	91.9 ± 1.6	55 ± 7
KSA-30	NIV	−	−	−	−	−	−	−	−	−	+	+	91.4 ± 1.9	48 ± 9
KSA-31	−II	−	−	−	−	−	−	+	+	+	+	+	90.7 ± 1.5	42 ± 3
KSA-32	−LI	−	−	−	−	−	−	− − −	− − −	− − −	++ −	+++	89.2 ± 4.0	47 ± 15
KSA-33	CLI	−	−	−	−	−	−	+	+	+	+	+	89.0 ± 2.1	38 ± 10
KSA-34	NLL	−	−	−	−	−	−	+	+	+	+	+	88.4 ± 1.0	30 ± 3
KSA-35	SII	−	−	−	−	−	−	+	+	+	+	+	86.2 ± 2.7	45 ± 9
KSA-36	VCI	−	−	−	−	−	−	+	+	+	+	+	85.5 ± 2.9	20 ± 3
KSA-37	−VI	−	−	−	−	−	−	+	+	+	+	+	84.5 ± 2.3	32 ± 8
hit numbers		2	+0	+1	+2	+0	+6	10	+0	+1	+4	+2	cumulative	
		11						17					total	

[a]Observed mutations are listed in one letter code in column two sequentially for the positions R47/T49/Y51. Occurrence of the WT amino acid is indicated by a dash (−).

finding at least 1 out of the *n* best mutants, we first screened the six combinatorial libraries with almost complete library coverage. Of special interest is the hit quality and quantity in our libraries when coverage is stepwise reduced from 95% to 45% as many studies have shown (Table 3). Thus, using the raw screening data from six large data sets, we then analyzed hit identity and frequency of 4 scenarios with reduced library coverage, and its equivalent to finding at least 1 out of the *n* best variants or hits.

We define "hits" as those mutants showing at least 85 or 80% regioselectivity and ≥35% activity (total testosterone conversion, HPLC) as in the case of library A or B, respectively, but in library C, no selective mutants were found, so we defined as hits those mutants exhibiting at least 50% activity. Only the top 20 transformants from each of the six libraries were selected

dx.doi.org/10.1021/sb5002399 | ACS Synth. Biol. 2015, 4, 317−331

Table 5. Amino Acid Identity, Occurrence at Different Library Coverage Stages and Performance of Hits after Re-culturing of Selected Transformants of Libraries B[a]

mutant		PCR						Sloning					biocatalytic parameters	
library size		240	300	399	598	1222	3008	240	300	399	598	1222		
library coverage		45%	53%	63%	78%	86%	98%	45%	53%	63%	78%	95%		
at least 1 out of *n* best		5	4	3	2	2−1	1	5	4	3	2	1	selectivity (%)	total conversion (%)
	78/82												15*β*	
KSA-5	LF	+ −	+ −	+ −	+ −	++	++	+ − − − −	+ − − − −	++ − − −	++ − − −	+++++	91.8 ± 1.3	64 ± 14
KSA-38	MF	−	−	−	−	−	−	−	−	−	+	+	91.3 ± 0.3	50 ± 23
KSA-6	TF	++− −	+++−	+++−	+++−	++++	++++	−	−	−	−	−	91.1 ± 0.7	58 ± 13
KSA-39	LH	−	−	−	−	−	−	− −	− −	− −	+ −	++	90.8 ± 1.1	21 ± 8
KSA-7	IF	− −	+ −	+ −	+ −	+ −	++	−	−	−	−	−	89.5 ± 1.4	77 ± 2
KSA-40	TM	−	−	−	−	+	+	−	−	−	−	−	86.0 ± 1.8	48 ± 6
KSA-41	MM	−	+	+	+	+	+	−	−	−	−	−	85.9 ± 0.7	58 ± 12
KSA-42	LM	−	−	−	−	−	−	+ − −	+ − −	++ −	++ −	+++	84.0 ± 6.4	50 ± 15
KSA-43	−M	+ −	+ −	+ −	+ −	+ −	++	−	−	−	−	−	82.4 ± 1.0	77 ± 5
KSA-44	ML	−	−	−	−	−	−	−	−	−	−	+	81.1 ± 0.7	62 ± 11
													2*β*	
KSA-46	LD	−	−	−	−	−	+	+ −	+ −	+ −	+ −	++	83.7 ± 1.8	47 ± 15
KSA-8	ID	+ − − −	+ − − −	+ − − −	+ − − −	+++ −	++++	− − −	− − −	− − −	+ − −	+++	82.9 ± 8.8	48 ± 14
KSA-47	MD	− − −	− − −	− − −	− − −	++ −	+++	+	+	+	+	+	80.7 ± 2.1	60 ± 15
KSA-48	−D	−	−	−	−	−	−	+ −	+ −	+ −	+ −	++	80.1 ± 2.2	11 ± 7
hit numbers		5	+3	+0	+0	+7	+5	5	+0	+2	+3	+10	cumulative	
		20						20					total	

[a]Observed mutations are listed in one letter code in column two sequentially for the positions V78/A82. Occurrence of the WT amino acid is indicated by a dash (−).

dx.doi.org/10.1021/sb5002399 | ACS Synth. Biol. 2015, 4, 317−331

as hits and sequenced (see Methods). When less than 20 transformants qualified as hits, the selectivity threshold was loosened stepwise by 1% until 20 transformants were reached. In general, all A and B libraries contained sufficient transformants for at least one of the main regioisomers 2β-OHT or 15β-OHT; for example, A-PCR contained 148 transformants fulfilling the hit criteria of which 44 showed selectivities above 90% for 2β-OHT under screening conditions (SI Table S4). Importantly, those transformants qualifying as hits were recultured and assayed as triplicates on the same MTP to ensure comparable reaction conditions for direct comparison. Thereafter, hits from both types of libraries were analyzed as if these would have been obtained in a reduced library size to mimic a screening process with lower coverage. We chose the absolute numbers predicted by ~95%, 78%, 63%, 53%, and 45% library coverage, which equates when striving for finding at least one out of n = 1, 2, 3, 4, 5 best variants (Table 4−6).

When analyzing the complete screening data, libraries A yielded together 18 mutants qualifying as hits with selectivities ranging from 95.5 to 84.5% for 2β−OHT with activities of 30−64% total testosterone conversion (Table 4). Interestingly, the best mutants are very close to each other regarding selectivity because 8 of them have 2β−OHT selectivities between 93 and 95% with conversions between 47 and 64%. Important for high stereoselective 2β-OHT production seems to be the occurrence of aliphatic amino acids (I, L, or V) at positions R47/T49 or T49/Y51, whereas R47 or Y51 can either be replaced by N, which is slightly smaller than V and bears hydrogen donor/acceptor functionality. Among the cluster of eight best hits, N is found only in position Y51, whereas the hydrogen donor functionally in position R47 leads to a reduction in selectivity. Furthermore, hits are not equally spread between both A libraries. Whereas A-PCR contains 11 hits among the apparent eight best, A-SLO contains only three of these (KSA-22, KSA-24, and KSA-2) but 10 other unique hits not encountered in A-PCR. A possible reason why these 10 hits (KSA-28 to KSA-37) were not encountered could be the low abundance of amino acids R and S, which is supported by the single sequencing results (Figure 2C). Importantly, even though some hits occur only in A-PCR and others only in A-SLO, both sets contain hits of similar biocatalytic fitness. When reducing the screening effort of A-PCR to equal that of A-SLO, four of the apparent best hits (KSA-3, KSA-2, KSA-26, and KSA-27) are still encountered, while two (KSA-23 and KSA-25) would be missed. Of course, stepwise reduction of the absolute library size reduces the amount of overall encountered hits in both libraries, but it still enables the researcher to find at least some out of the best hits. Even the heavily reduced library sets of only 1036 clones (45% coverage) still contain the best hits KSA-2, KSA-22, and KSA-27. This would correspond to finding at least 1 out of the 5 best mutants, but we found in both cases two of the five best hits. Overall, the reduction of library coverage has not hampered the discovery of improved biocatalysts in either library type. However, more hits were found in A-SLO compared to A-PCR (17 vs 11), which demonstrates the advantage of screening better quality libraries. In any case, these findings suggest that the reduction of library coverage could be an alternative to the conservative 95% library coverage screening if the goal is to find a practical biocatalyst.

In the case of library B, 10 hits with selectivities varying from 92 to 81% for 15β-OHT were identified, which displayed a broad range of activities from 21 to 77%. Also, four sequence related mutants producing mainly 2β-OHT were found with selectivities ranging from 84 to 80% and activities between 11 and 66% (Table 5). Comparing the observed mutations at position 82, it becomes clear that for 15β-hydroxylation the occurrence of bulky, hydrophobic residues is important. Indeed, a shift toward 16β-hydroxylation is observed when the biggest amino acid W was encountered (Figure S3B) and 2β-hydroxylation is observed by incorporating D. The role of position 78 could not be revealed from the collected data, but improvement was achieved when L, M or I replace V regardless of regioselectivity (Table 5). Different to libraries A, where most hits were encountered only once or twice, libraries B yielded five hits from three to five times. Similarly, when >95% library coverage is done, B-PCR and B-SLO show the same number of hits (20). We can attribute this finding to the fact that both libraries display a uniform distribution of those amino acids required for the factors determining regioselectivity (see Figure 2D+G).

Overall, the same trend as seen in A libraries emerges: A high coverage promotes the discovery of various closely related hits, providing interesting insights into the biocatalytic landscape which describes the $P450_{BM3}$ hydroxylation of testosterone. Just as in the A libraries, when library coverage is reduced, at least four improved mutants are still found for a library size of 240 transformants or 45% library coverage (Table 5).

The best hits of the C libraries show conversions up to 80% with selectivities between 54 and 74% in favor of 15β-OHT (Table 6). Interestingly 5 out of the 6 best mutants were found in C-SLO more than once (e.g., KSA-52 was found 6 times). Regarding C-PCR, it is interesting to note that this library failed to contain hits having more than 54% conversion. In contrast, C-SLO bears transformants with as twice as much (80%) conversion. Sequencing revealed the presence of small residues at positions 185 and 188 like G, C and S, but also R contributed to an increase in activity as seen with mutants KSA-52, KSA-12, KSA-13, KSA-53, and KSA-54. Importantly, a look at Figure 2E shows that C-PCR sequenced data set failed to contain variants with G or R in either addressed position. It is therefore not surprising that these mutants were missed in the C-PCR library.

Hence, libraries C-PCR and C-SLO are important tutorial examples because they reveal the pitfalls a researcher in protein evolution may face: Figure 3C+F shows two different data scattering patterns, which we attribute to the pronounced differences in library quality at the DNA level. The lower diversity quality means that areas of the targeted sequence space were not created in C-PCR. Unfortunately, other than in libraries A-PCR and B-PCR, where the quality difference was not so pronounced, this missed area(s) of the sequence space encoded variants with improved catalytic parameters, as can be seen in C-SLO, which contains more active hits than C-PCR (Figure 3C+F). These results confirm that achieving the right diversity is crucial for the subsequent screening effort and the discovery of hits. The second apparent pitfall is that it is important to consider the target residues for randomization wisely. Residues M185 and L188 of library site C influenced only the activity toward steroid hydroxylation, whereas residues R47, T49, Y51, V78, and A82 are important for controlling regioselectivity and having a likewise positive influence on total activity. When not observing any good hits in a library, and low diversity quality can be excluded, then and only then can we conclude that the targeted residues do not contribute to the examined catalytic parameter(s).

Importantly, upon analyzing carefully the hits in all libraries, we found that aiming for reduced library coverage as predicted

Table 6. Amino Acid Identity, Occurrence at Different Library Coverage Stages and Performance of Hits after Re-culturing of Selected Transformants of Libraries C[a]

mutant		PCR						Sloning				biocatalytic parameters	
library size		86	108	143	215	672	752	86	108	143	215	672			
library coverage		45%	53%	63%	78%	86%	98%	45%	53%	63%	78%	95%			
at least 1 out of the n best		5	4	3	2	2−1	1	5	4	3	2	1	total conversion (%)	15β-selectivity (%)	
KSA-52	185/188 GC	−	−	−	−	−	−	+++ − −	+++++ −	+++++ −	+++++ −	++++++	80 ± 11	54.2 ± 2.8	
KSA-12	SC	−	−	−	−	−	−	− −	− −	− −	− −	++	78 ± 7	53.8 ± 11.3	
KSA-13	GG	−	−	−	−	−	−	+	+	+	+	+	70 ± 5	53.9 ± 3.4	
KSA-53	RC	−	−	−	−	−	−	− −	− −	− −	− −	++	64 ± 5	53.4 ± 1.0	
KSA-54	RN	−	−	−	−	−	−	−	−	−	+	+	56 ± 5	66.8 ± 3.2	
KSA-11	NN	−	−	−	−	+	+	−	−	−	−	−	54 ± 10	69.4 ± 2.7	
KSA-55	SS	−	−	−	−	+	+	−	−	−	−	+	53 ± 10	58.5 ± 9.2	
KSA-56	SG	−	−	−	−	−	−	−	−	−	−	+	53 ± 7	58.6 ± 6.1	
hit numbers		+0	+0	+0	+0	+2	+0	+5	+1	+0	+1	+7	cumulative total		
		2						14							

[a]Observed mutations are listed in one letter code in column two sequentially for the positions M185/L188.

by Nov is a beneficial guide for the goal of finding an improved mutant for biocatalysis. In fact, around 50% reduction of the screening effort still offers the potential to find at least one of the best hits. For the interested reader, we also provide a closer look on the 63% coverage or finding "at least 1 out of the 3 best variants" data set in direct comparison with the 95% coverage data set (see SI). We believe that using the Nov metric gives more confidence to the researcher because its user-friendly online tool can help in designing libraries of any size and randomization up to 12 target codons,[34] whereas computing library coverage percentages is limited to defined codon degeneracies of maximally 6 codons and does not allow to consider the yields of construct synthesis using GLUE-IT.[33]

Since sometimes the encountered hits do not fulfill the set project goal in the first round of mutagenesis and screening, instead of further screening that library to increase its coverage, a different strategy is to perform Iterative Saturation Mutagenesis (ISM). For example, we chose the template of a very active (84%) yet unselective variant (62%) from library A-PCR not qualifying here as hit (KSA-4: R47Y/T49F/F87A)[35] and addressed the residues of library B as before, finding a quintuple mutant (KSA-14: R47Y/T49F/V78L/A82M/F87A) with excellent selectivity (96%) and activity (85% HPLC conversion) as best hit. Interestingly, KSA-14 contains the mutations of KSA-42 (V78L/A82M), a previously identified apparent eighth best hit with 84% selectivity and 50% activity (Table 5). In a second example, the mutations observed in KSA-6 (V78T/A82F/F87A), the apparent third best hit from B-PCR (Table 5), were encountered in the third best hit KSA-16 (R47Y/T49F/V78T/A82F/F87A) after ISM showing 90% selectivity for 15β−OHT and 60% activity as reported elsewhere.[35] This indicates that the rational combination of mutations is not always straightforward, and that ISM in combination with reduced screening effort is a very fast strategy to develop biocatalysts. This reduction of screening effort is of prime importance when aiming for implementing an industrial process because biocatalyst development time can contribute significantly to the overall costs. Of course, it also remains to be demonstrated whether such a library coverage reduction also applies in combinatorial libraries of larger complexity (e.g., 5 or more residues randomized with similar or smaller amino acid alphabets as reported here), but the message is that at least one of the n best hits can be found in a practical manner. Last but not least, reducing the screening effort gives a glimpse on the biocatalytic and/or fitness landscape, yet the exploration of almost complete library coverage has given insight into our understanding of ISM-based directed evolution and the emergence of epistasis, thereby rationally improving the strategies for evolving better enzymes.[58]

Conclusion and Outlook. We have shown that three diverse combinatorial libraries based on saturation mutagenesis created via solid-phase gene synthesis (Sloning technology) have greater quality at both the DNA and protein level than the same libraries created by an optimized PCR-based megaprimer approach. This conclusion is based on the following experimental results: First, no restriction to existing degenerate codons; second, as a consequence, removal of redundancy, resulting in lowered screening effort; third, the contamination with parental constructs and incorrect constructs is significantly lower, resulting in higher quality libraries on the DNA level, which ensures a higher coverage of the targeted sequence space. All these factors together contribute to a significantly reduced screening effort by up to 59% (library B-SLO vs B-PCR),

dx.doi.org/10.1021/sb5002399 | ACS Synth. Biol. 2015, 4, 317−331

providing an economic advantage over the PCR created libraries. Currently, it is difficult to assess financial differences, because it is problematic to compare the materials plus labor that are necessary when creating PCR-based libraries with the costs of buying commercially produced libraries. However, if the costs of gene syntheses continue to go down, designed solid-phase combinatorial gene libraries for saturation mutagenesis accessible on a commercial basis may prove to be the favored option in future directed evolution.

Our results also provide evidence that the Nov statistical approach of finding at least 1 out of the *n*th best mutants, or its equivalent library coverage reduction based on the Patrick and Firth algorithm, are useful strategies with the goal of finding an improved biocatalyst because the screening effort can be decreased significantly irrespective of library design and creation methodology. Of course, more hits can be discovered when more screening is performed, something that is advisable when exploring sequence-to-function relationships as a goal. In addition, it is important to remember that in order to find a practical catalyst or extensively exploring sequence space, the aimed diversity has to be properly achieved. If not, more exhaustive screening of a library will never provide satisfactory results, as seen for library C-PCR. Application of the QQC is mandatory. Likewise, if the chosen residues for saturation mutagenesis have no effect on the investigated biocatalytic parameters, it is likely that there will be no success (as with selectivity in library C). However, if the right residues are chosen, the likelihood of finding more hits increases, as shown for library A and B in terms of selectivity.

Finally, the development of useful guidelines for ISM,[37] together with high-quality solid-phase gene synthesis as described here for combinatorial libraries in combination with less conservative screening (partial library coverage) seems to be the most powerful strategy for beating the numbers problem in directed evolution. Retrospectively, it is impressive to realize that combining these strategies could have saved 92% screening effort (240 versus 3008 transformants) while still finding at least 1 out of the best hits (Table 5) as in the case of library B.

METHODS

Figure Preparation. $P450_{BM3}$ structure was obtained by manually mutating residue Phe87 into Alanine with Maestro software[59] using 1JPZ.pdb[60] (chain B) as starting point, followed by refinement through the Protein Preparation Wizard[59] (standard settings). Testosterone was prepared with LigPrep[59] and docked using Glide[61] employing a modified heme force field.[62] Docking resulted in several poses, from which one, orienting testosterone for hydroxylation in 2β position, is shown.

Library Construction, Quality Control, Screening, and Analysis. Sloning prepared *bm3* monooxygenase (*cyp*102A1) diversified gene pools with the codons given in Figure 1C and D. The gene fragments had a size of 683 bp (954 with flanking sequences) and were cloned into pETM11 with *Noc*I and *Nhe*I prior to *E. coli* BL21 Gold(DE3) transformation. PCR library construction and screening of PCR and Sloning libraries was performed as previously described.[35] Briefly, *E. coli* cells were expressed in 96 deep-well plates containing TB media supplemented with $Fe(III)Cl_3$ and glutamate to enhance gene expression levels, harvested, washed, and supplemented with potassium phosphate based reaction buffer containing 1 μM testosterone and a glucose/glucose dehydrogenase based recycling system. The 600 μL-reaction was carried out for 24 h at 25 °C, followed by extraction with 350 μL ethyl acetate, evaporation, and uptake of the dried sample in 150 μL acetonitrile prior to HPLC analysis. HPLC chromatograms were automatically integrated using standard settings for peak integration by the HPLC analysis software (LSsolution Version 1.22 SP1 by Shimadzu Corporation). Assigned peaks, as defined by the programmed compound table, were automatically exported, resulting in a single Excel file, which contains all HPLC results of one library. Entries were sorted descending by selectivity. False positives were eliminated by using Excel's filter function to exclude those transformants, where the regioisomer peak had an area count below 10 000. A cutoff activity threshold of ≥35% for total testosterone conversion was applied and the first 10−20 entries were selected as hits and taken for validation and sequencing. Hits were recultured and assayed under identical conditions as above for validation. Hits were sequence by preparing plasmid DNA using a Qiagen Miniprep Kit and sending the DNA to GATC Biotech AG (Germany). Sequencing results used for the statistical analysis were obtained by sending variants in a 96 multititer plate filled with LB agar containing 5 μg/mL kanamycin, which was inoculated from a library plate using a replicator, to GATC Biotech AG (Germany) for plasmid preparation and sequencing with the T7 primer. Grubbs test was performed as previously reported,[25] and Microsoft Excel was used for boxplot analysis of the absolute number of amino acids observed in the six random samples and chi-square analysis of the obtained amino acid frequencies (raw data available in SI Table S3).

Information about the results of the QQC, the observed codons of the random samples and the number of transformants qualified as hits can be found in Supporting Information.

ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Email: reetz@kofo.mpg.de.

Author Contributions

‖S.H. and F.E.Z. contributed equally. F.E.Z., S.H., and M.T.R. developed the idea. S.H. and partially F.E.Z. performed the experimental work. S.H. analyzed and together with C.G.A.R. interpreted the data. M.Z., S.H., and F.E.Z. performed the statistical analyses. S.H. and C.G.A.R. wrote the manuscript with help of M.T.R., F.E.Z., and M.Z.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Heike Hinrichs for technical performance of the HPLC measurements. S.H. thanks Yuval Nov for fruitful discussions concerning the different metrics. This work was supported by the Max Planck Society and the Arthur C. Cope Foundation.

REFERENCES

(1) Bornscheuer, U. T., Huisman, G. W., Kazlauskas, R. J., Lutz, S., Moore, J. C., and Robins, K. (2012) Engineering the third wave of biocatalysis. *Nature 485*, 185−194.

(2) Carothers, J. M., Goler, J. A., and Keasling, J. D. (2009) Chemical synthesis using synthetic biology. *Curr. Opin. Biotechnol. 20*, 498−503.
(3) Leprince, A., van Passel, M. W., and dos Santos, V. A. (2012) Streamlining genomes: Toward the generation of simplified and stabilized microbial systems. *Curr. Opin. Biotechnol. 23*, 651−658.
(4) Xiong, A. S., Peng, R. H., Zhuang, J., Liu, J. G., Gao, F., Chen, J. M., Cheng, Z. M., and Yao, Q. H. (2008) Non-polymerase-cycling-assembly-based chemical gene synthesis: Strategies, methods, and progress. *Biotechnol. Adv. 26*, 121−134.
(5) Xiong, A. S., Peng, R. H., Zhuang, J., Gao, F., Li, Y., Cheng, Z. M., and Yao, Q. H. (2008) Chemical gene synthesis: Strategies, softwares, error corrections, and applications. *FEMS Microbiol. Rev. 32*, 522−540.
(6) Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Jr., Kinney, J. B., Kellis, M., Lander, E. S., and Mikkelsen, T. S. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol. 30*, 271−277.
(7) Kosuri, S., Goodman, D. B., Cambray, G., Mutalik, V. K., Gao, Y., Arkin, A. P., Endy, D., and Church, G. M. (2013) Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A. 110*, 14024−14029.
(8) Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2012) Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U.S.A. 109*, 19498−19503.
(9) Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol. 27*, 1173−1175.
(10) Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol. 30*, 521−530.
(11) Liachko, I., Youngblood, R. A., Keich, U., and Dunham, M. J. (2013) High-resolution mapping, characterization, and optimization of autonomously replicating sequences in yeast. *Genome Res. 23*, 698−704.
(12) Quan, J., Saaem, I., Tang, N., Ma, S., Negre, N., Gong, H., White, K. P., and Tian, J. (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol. 29*, 449−452.
(13) LeProust, E. M., Peck, B. J., Spirin, K., McCuen, H. B., Moore, B., Namsaraev, E., and Caruthers, M. H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res. 38*, 2522−2540.
(14) Mulligan, J. T., Parker, H.-Y. Solid phase methods for polynucleotide production. U.S. Patent No. 7,482,119, Dec. 16, 2008.
(15) Van den Brulle, J., Fischer, M., Langmann, T., Horn, G., Waldmann, T., Arnold, S., Fuhrmann, M., Schatz, O., O'Connell, T., O'Connell, D., Auckenthaler, A., and Schwer, H. (2008) A novel solid phase technology for high-throughput gene synthesis. *BioTechniques 45*, 340−343.
(16) Kuhn, S. M., Rubini, M., Fuhrmann, M., Theobald, I., and Skerra, A. (2010) Engineering of an orthogonal aminoacyl-tRNA synthetase for efficient incorporation of the non-natural amino acid O-methyl-L-tyrosine using fluorescence-based bacterial cell sorting. *J. Mol. Biol. 404*, 70−87.
(17) Gebauer, M., Schiefner, A., Matschiner, G., and Skerra, A. (2013) Combinatorial design of an Anticalin directed against the extra-domain b for the specific targeting of oncofetal fibronectin. *J. Mol. Biol. 425*, 780−802.
(18) Bowers, P. M., Neben, T. Y., Tomlinson, G. L., Dalton, J. L., Altobell, L., Zhang, X., Macomber, J. L., Wu, B. F., Toobian, R. M., McConnell, A. D., Verdino, P., Chau, B., Horlick, R. A., and King, D. J. (2013) Humanization of antibodies using heavy chain complementarity-determining region 3 grafting coupled with *in vitro* somatic hypermutation. *J. Biol. Chem. 288*, 7688−7696.
(19) Zhai, W., Glanville, J., Fuhrmann, M., Mei, L., Ni, I., Sundar, P. D., Van Blarcom, T., Abdiche, Y., Lindquist, K., Strohner, R., Telman, D., Cappuccilli, G., Finlay, W. J., Van den Brulle, J., Cox, D. R., Pons, J., and Rajpal, A. (2011) Synthetic antibodies designed on natural sequence landscapes. *J. Mol. Biol. 412*, 55−71.
(20) Parikh, M. R., and Matsumura, I. (2005) Site-saturation mutagenesis is more efficient than DNA shuffling for the directed evolution of β-fucosidase from β-galactosidase. *J. Mol. Biol. 352*, 621−628.
(21) Reetz, M. T. (2011) Laboratory evolution of stereoselective enzymes: A prolific source of catalysts for asymmetric reactions. *Angew. Chem., Int. Ed. Engl. 50*, 138−174.
(22) Reetz, M. T., Prasad, S., Carballeira, J. D., Gumulya, Y., and Bocola, M. (2010) Iterative saturation mutagenesis accelerates laboratory evolution of enzyme stereoselectivity: Rigorous comparison with traditional methods. *J. Am. Chem. Soc. 132*, 9144−9152.
(23) Reetz, M. T. (2013) Biocatalysis in organic chemistry and biotechnology: Past, present, and future. *J. Am. Chem. Soc. 135*, 12480−12496.
(24) Hogrefe, H. H., Cline, J., Youngblood, G. L., and Allen, R. M. (2002) Creating randomized amino acid libraries with the QuikChange Multi Site-Directed Mutagenesis Kit. *BioTechniques 33*, 1158−1160 1162, 1164−1155.
(25) Kille, S., Acevedo-Rocha, C. G., Parra, L. P., Zhang, Z. G., Opperman, D. J., Reetz, M. T., and Acevedo, J. P. (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth. Biol. 2*, 83−92.
(26) Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M., and Jiang, R. (2012) Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques 52*, 149−158.
(27) Sarkar, G., and Sommer, S. S. (1990) The "megaprimer" method of site-directed mutagenesis. *BioTechniques 8*, 404−407.
(28) Dennig, A., Shivange, A. V., Marienhagen, J., and Schwaneberg, U. (2011) OmniChange: The sequence independent method for simultaneous site-saturation of five codons. *PLoS One 6*, e26222.
(29) Acevedo-Rocha, C. G., and Reetz, M. T. (2014) Assembly of designed oligonucleotides: A useful tool in synthetic biology for creating high quality combinatorial DNA libraries. In *Directed Evolution Library Creation: Methods and Protocols* (Ackerley, D., Copp, J., and Gillam, E., Eds.) Springer, New York, In Press. DOI: 10.1007/978-1-4939-1053-3_13.
(30) Sanchis, J., Fernandez, L., Carballeira, J. D., Drone, J., Gumulya, Y., Hobenreich, H., Kahakeaw, D., Kille, S., Lohmer, R., Peyralans, J. J., Podtetenieff, J., Prasad, S., Soni, P., Taglieber, A., Wu, S., Zilly, F. E., and Reetz, M. T. (2008) Improved PCR method for the creation of saturation mutagenesis libraries in directed evolution: Application to difficult-to-amplify templates. *Appl. Microbiol. Biotechnol. 81*, 387−397.
(31) Agudo, R., Roiban, G. D., and Reetz, M. T. (2012) Achieving regio- and enantioselectivity of P450-catalyzed oxidative CH activation of small functionalized molecules by structure-guided directed evolution. *ChemBioChem 13*, 1465−1473.
(32) Mahon, C. M., Lambert, M. A., Glanville, J., Wade, J. M., Fennell, B. J., Krebs, M. R., Armellino, D., Yang, S., Liu, X., O'Sullivan, C. M., Autin, B., Oficjalska, K., Bloom, L., Paulsen, J., Gill, D., Damelin, M., Cunningham, O., and Finlay, W. J. (2013) Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. *J. Mol. Biol. 425*, 1712−1730.
(33) Patrick, W. M., Firth, A. E., and Blackburn, J. M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng. 16*, 451−457.
(34) Nov, Y. (2012) When second best is good enough: Another probabilistic look at saturation mutagenesis. *Appl. Environ. Microbiol. 78*, 258−262.
(35) Kille, S., Zilly, F. E., Acevedo, J. P., and Reetz, M. T. (2011) Regio- and stereoselectivity of P450-catalysed hydroxylation of steroids controlled by laboratory evolution. *Nat. Chem. 3*, 738−743.

(36) Reetz, M. T., Carballeira, J. D., Peyralans, J., Hobenreich, H., Maichele, A., and Vogel, A. (2006) Expanding the substrate scope of enzymes: Combining mutations obtained by CASTing. *Chemistry 12*, 6031−6038.
(37) Acevedo-Rocha, C. G., Kille, S., and Reetz, M. T. (2014) Iterative saturation mutagenesis: A powerful approach to engineer proteins by systematically simulating Darwinian evolution. In *Directed Evolution Library Creation: Methods and Protocols* (Ackerley, D., Copp, J., and Gillam, E., Eds.) Springer, New York, In Press. DOI: 10.1007/978-1-4939-1053-3_7.
(38) *The PyMOL Molecular Graphics System*, Version 1.5.0.4; Schrödinger, LLC: New York.
(39) Reetz, M. T., Kahakeaw, D., and Lohmer, R. (2008) Addressing the numbers problem in directed evolution. *ChemBioChem. 9*, 1797−1804.
(40) Bougioukou, D. J., Kille, S., Taglieber, A., and Reetz, M. T. (2009) Directed evolution of an enantioselective enoate-reductase: Testing the utility of iterative saturation mutagenesis. *Adv. Synth. Catal. 351*, 3287−3305.
(41) Bosley, A. D., and Ostermeier, M. (2005) Mathematical expressions useful in the construction, description, and evaluation of protein libraries. *Biomol. Eng. 22*, 57−61.
(42) Nov, Y. (2013) Fitness loss and library size determination in saturation mutagenesis. *PLoS One 8*, e68069.
(43) Horne, M. T., Fish, D. J., and Benight, A. S. (2006) Statistical thermodynamics and kinetics of DNA multiplex hybridization reactions. *Biophys. J. 91*, 4133−4153.
(44) SantaLucia, J., Jr., Allawi, H. T., and Seneviratne, P. A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry 35*, 3555−3562.
(45) The need of controls per plate and an even number of 96-deep-well plates for simplier laboratory processing.
(46) Ashraf, M., Frigotto, L., Smith, M. E., Patel, S., Hughes, M. D., Poole, A. J., Hebaishi, H. R., Ullman, C. G., and Hine, A. V. (2013) ProxiMAX randomization: a new technology for non-degenerate saturation mutagenesis of contiguous codons. *Biochem. Soc. Trans. 41*, 1189−1194.
(47) Firth, A. E., and Patrick, W. M. (2008) GLUE-IT and PEDEL-AA: New programmes for analyzing protein diversity in randomized libraries. *Nucleic Acids Res. 36*, W281−285.
(48) Evans, B. S., Chen, Y., Metcalf, W. W., Zhao, H., and Kelleher, N. L. (2011) Directed evolution of the nonribosomal peptide synthetase AdmK generates new andrimid derivatives *in vivo*. *Chem. Biol. 18*, 601−607.
(49) Blikstad, C., Dahlström, K. M., Salminen, T. A., and Widersten, M. (2013) Stereoselective oxidation of aryl-substituted vicinal diols into chiral α-hydroxy aldehydes by re-engineered propanediol oxidoreductase. *ACS Catal. 3*, 3016−3025.
(50) Carlsson, A. J., Bauer, P., Ma, H., and Widersten, M. (2012) Obtaining optical purity for product diols in enzyme-catalyzed epoxide hydrolysis: Contributions from changes in both enantio- and regioselectivity. *Biochemistry 51*, 7627−7637.
(51) Sandstrom, A. G., Wikmark, Y., Engstrom, K., Nyhlen, J., and Backvall, J. E. (2012) Combinatorial reshaping of the *Candida antarctica* lipase A substrate pocket for enantioselectivity using an extremely condensed library. *Proc. Natl. Acad. Sci. U.S.A. 109*, 78−83.
(52) Hougland, J. L., Gangopadhyay, S. A., and Fierke, C. A. (2012) Expansion of protein farnesyltransferase specificity using "tunable" active site interactions: Development of bioengineered prenylation pathways. *J. Biol. Chem. 287*, 38090−38100.
(53) Jakoblinnert, A., van den Wittenboer, A., Shivange, A. V., Bocola, M., Heffele, L., Ansorge-Schumacher, M., and Schwaneberg, U. (2013) Design of an activity and stability improved carbonyl reductase from *Candida parapsilosis*. *J. Biotechnol. 165*, 52−62.
(54) Ji, D., Wang, L., Hou, S., Liu, W., Wang, J., Wang, Q., and Zhao, Z. K. (2011) Creation of bioorthogonal redox systems depending on nicotinamide flucytosine dinucleotide. *J. Am. Chem. Soc. 133*, 20857−20862.
(55) Zhang, W., Moden, O., Tars, K., and Mannervik, B. (2012) Structure-based redesign of GST A2-2 for enhanced catalytic efficiency with azathioprine. *Chem. Biol. 19*, 414−421.
(56) Parra, L. P., Agudo, R., and Reetz, M. T. (2013) Directed evolution by using iterative saturation mutagenesis based on multiresidue sites. *ChemBioChem. 14*, 2301−2309.
(57) Reetz, M. T., Wilensek, S., Zha, D., and Jaeger, K. E. (2001) Directed evolution of an enantioselective enzyme through combinatorial multiple-cassette mutagenesis. *Angew. Chem., Int. Ed. Engl. 40*, 3589−3591.
(58) Reetz, M. T. (2013) The importance of additive and non-additive mutational effects in protein engineering. *Angew. Chem., Int. Ed. Engl. 52*, 2658−2666.
(59) Suite 2012: Maestro, version 9.3, Schrödinger; LigPrep, version 2.5, Schrödinger; Protein Preparation Wizard: Epik version 2.3, Impact version 5.8; Prime version 3.1, Schrödinger, LLC: New York, 2012.
(60) Haines, D. C., Tomchick, D. R., Machius, M., and Peterson, J. A. (2001) Pivotal role of water in the mechanism of P450BM-3. *Biochemistry 40*, 13456−13465.
(61) Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem. 47*, 1739−1749.
(62) Seifert, A., Tatzel, S., Schmid, R. D., and Pleiss, J. (2006) Multiple molecular dynamics simulations of human p450 monooxygenase CYP2C9: The molecular basis of substrate binding and regioselectivity toward warfarin. *Proteins 64*, 147−155.