Journal of Molecular Evolution (2024) 92:402–414
https://doi.org/10.1007/s00239-024-10179-8

ORIGINAL ARTICLE

Sampling Strategies for Experimentally Mapping Molecular Fitness Landscapes Using High-Throughput Methods

Steven K. Chen[1] · Jing Liu[1] · Alexander Van Nynatten[2] · Benjamin M. Tudor-Price[1] · Belinda S. W. Chang[1,3,4]

Received: 28 February 2024 / Accepted: 20 May 2024 / Published online: 17 June 2024

Abstract

Empirical studies of genotype–phenotype–fitness maps of proteins are fundamental to understanding the evolutionary process, in elucidating the space of possible genotypes accessible through mutations in a landscape of phenotypes and fitness effects. Yet, comprehensively mapping molecular fitness landscapes remains challenging since all possible combinations of amino acid substitutions for even a few protein sites are encoded by an enormous genotype space. High-throughput mapping of genotype space can be achieved using large-scale screening experiments known as multiplexed assays of variant effect (MAVEs). However, to accommodate such multi-mutational studies, the size of MAVEs has grown to the point where a priori determination of sampling requirements is needed. To address this problem, we propose calculations and simulation methods to approximate minimum sampling requirements for multi-mutational MAVEs, which we combine with a new library construction protocol to experimentally validate our approximation approaches. Analysis of our simulated data reveals how sampling trajectories differ between simulations of nucleotide versus amino acid variants and among mutagenesis schemes. For this, we show quantitatively that marginal gains in sampling efficiency demand increasingly greater sampling effort when sampling for nucleotide sequences over their encoded amino acid equivalents. We present a new library construction protocol that efficiently maximizes sequence variation, and demonstrate using ultradeep sequencing that the library encodes virtually all possible combinations of mutations within the experimental design. Insights learned from our analyses together with the methodological advances reported herein are immediately applicable toward pooled experimental screens of arbitrary design, enabling further assay upscaling and expanded testing of genotype space.

Keywords Molecular fitness landscape · Pooled genetic screen · Multiplexed assays of variant effect · Deep mutational scan · Multi-step mutagenesis · Sampling requirements · Epistasis · Higher-order interactions

Introduction

A quantitative understanding of the context-dependent effects of amino acid substitutions could reveal important insights into the evolutionary potential of proteins but is difficult to achieve due to the enormous number of possible variants and interactions between even a small number of sites. An early effort to study these context-dependent effects reconstructed several intermediate sequences and measured their thermostability as a factor of selection (Malcolm et al. 1990). Since then, the approach has expanded to reconstructing sets of all possible sequence intermediates in adaptive evolutionary trajectories, and has been applied to examine various proteins (Lunzer et al. 2005; Weinreich et al. 2006; Poelwijk et al. 2007; O'Maille et al. 2008; Lozovsky et al. 2009; Meini et al. 2015; Tufts et al. 2015; Yang et al. 2019) and other biomolecules (Domingo et al.

Handling editor: **David Liberles**.

Steven K. Chen and Jing Liu have contributed equally to this work.

✉ Belinda S. W. Chang
belinda.chang@utoronto.ca

1 Department of Cell & Systems Biology, University of Toronto, Toronto, ON, Canada

2 Department of Biological Science, University of Toronto Scarborough, Toronto, ON, Canada

3 Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON, Canada

4 Centre for the Analysis of Genome Evolution & Function, University of Toronto, Toronto, ON, Canada

2018; Baeza-Centurion et al. 2019). Although these studies have mapped parts of genotype–phenotype–-fitness landscapes, and provided valuable insights into the effects of higher-order epistatic interactions (Weinreich et al. 2013, 2018), this is clearly an area with huge potential for future growth, in explaining why specific evolutionary pathways were selected among the vast number of alternatives.

In principle, high-throughput experimental mapping of genotype space should substantially increase the size, resolution, and completeness of the resulting maps. Recent advances in DNA synthesis and deep sequencing technologies have given rise to high-throughput experimental designs, known as multiplexed assays of variant effect (MAVEs), that combine laboratory selection with deep sequencing to characterize the phenotypes and fitness effects for thousands, or even millions, of genetic variants. This high-throughput scalability is the primary advantage of using MAVEs, and is made possible by its unbiased approach to pool-screening genetic variant libraries (Fowler et al. 2010) rather than more traditional low-throughput testing of individual mutants. Given the screening capacity of MAVEs, the approach has been used widely to profile single nucleotide variants for clinical interpretation (Starita et al. 2017; Tabet et al. 2022; Weile and Roth 2018), and double mutants for probing pairwise molecular and biophysical interactions (Olson et al. 2014; Rollins et al. 2019; Schmiedel and Lehner 2019; Faure et al. 2022). More recently, researchers are starting to design even larger MAVE experiments to investigate complex evolutionary pathways and mechanisms (Li et al. 2016; Puchta et al. 2016; Sarkisyan et al. 2016; Wu et al. 2016; Starr et al. 2017; Domingo et al. 2018; Baeza-Centurion et al. 2019; Blanco et al. 2019) by mapping the variant effects of higher-order interactions (Weinreich et al. 2013, 2018) involving combinations of mutations. Such experimental designs may include studies of molecular fitness landscapes where single point mutations can lead to significant changes in activity or even the emergence of new protein functions (Romero and Arnold 2009; Soskine and Tawfik 2010), thus highlighting the importance of exhaustive or near-exhaustive screening. For combinatorial mutagenesis studies, however, the number of possible combinations of mutations can be immense and increases geometrically for each additional residue/site considered.

Various approaches have been taken to expand the combinatorial nature of MAVE experiments, with some focusing all of the mutagenesis effort on a small number of preselected residues/sites (Wu et al. 2016; Starr et al. 2017), while others have taken a random mutagenesis approach by introducing mutations across a large number of positions in the target gene (Sarkisyan et al. 2016). Between these two mutagenesis strategies exists a trade-off between being able to generate and assay all possible combinations of mutations and the number of positions considered for mutagenesis. This trade-off ultimately affects the scope of the MAVE experiment by constraining the number of positions selected for mutagenesis or the fraction of possible variants characterized at selected positions. For example, random mutagenesis across the full-length avGFP sequence profiled 92% of all single amino acid replacements but less than 2% of all pairwise and higher-order combinations (Sarkisyan et al. 2016). In contrast, highly focused mutagenesis limited variation to four coding positions in the GB1 gene but enabled 93.4% of all single, double, triple, and quadruple variants at the selected positions to be characterized (Wu et al. 2016). This trade-off is a consequence of experimental bottlenecks on variant pool size and/or complexity (Fowler et al. 2014; Faure et al. 2020), and is also impacted by the choice of codon mutagenesis scheme (NNK, NNB, NNN in reference to the IUPAC nucleotide codes) which may result in amino acid biases in the variant library. It is thus important to account for these factors when designing experiments aimed at screening combinatorial libraries of extraordinary diversity.

A general feature of MAVE workflows is the use of pooled approaches at every major experimental step, from initial library construction to deep sequencing. Since these methods involve synthesizing and sampling from extremely large variants pooled, it is imperative when designing such experiments to consider the number of possible variants in the library (that is, library size) and the minimum amount of random sampling (that is, sampling requirement) that should be conducted to overcome experimental bottlenecks. Yet, to our knowledge, methods for approximating sampling requirements for combinatorial MAVE experiments have yet to be developed. Laboratory testing of any such approximation approaches is also missing, thus limiting the widespread adoption, design, and deployment of combinatorial MAVE experiments.

To address the need for exhaustive characterization of combinatorial libraries in MAVE experiments, we propose statistics calculations and simulation methods for approximating minimum sampling requirements. Although the calculations and simulations are, in principle, generalizable to any number of sites and codon mutagenesis schemes, we show that these methods robustly approximate minimum sampling requirements for combinatorial libraries. By analyzing the simulation process, we not only confirmed our calculation method but determined quantitatively how demand for sampling effort increases as sampling progresses toward recovery of all variants in the library. We further resolved sampling trajectory differences between simulations of nucleotide versus amino acid variants and among mutagenesis schemes. To enable laboratory testing of our approximation approaches, we developed a new high-efficiency laboratory protocol that brings together intramolecular plasmid assembly and ultradeep multiplexed sequencing

to create high-quality combinatorial libraries. Using this protocol, we demonstrated the practical utility of our approximation approaches by creating two four-codon combinatorial libraries that encode for virtually all possible combinations of mutations within the experimental design.

Results

Minimum Sampling Requirements for Testing Combinatorial Libraries

To investigate sampling requirements for designing combinatorial MAVE experiments, we consider a hypothetical set of sequences where four codon positions are mutated combinatorially to encode all possible mutational pathways involving four amino acid sites. Mutagenesis was done using NNK randomization to reduce codon usage and access to premature stop codons while maintaining the ability to encode all 20 amino acids. When NNK randomization is applied to mutate four codon positions combinatorially, the resulting library of sequences comprises 32^4 or 1,048,576 possible nucleotide variants, which encode for 160,000 amino acid sequences.

To determine minimum sampling requirements for this combinatorial library, we applied a solution to a classical problem in probability theory (Newman 1960) to approximate the minimum number of sequence combinations that need to be sampled randomly to have sampled each unique nucleotide variant at least once. For the purpose of making a general approximation, we assume that each degenerate oligonucleotide used to construct such a library binds the target sequence to incorporate changes with equal probabilities. We also assume that the pool of oligonucleotides is in excess as the number of oligonucleotides used to construct such libraries typically outnumber experimentally recoverable sequences by many orders of magnitude (Supplemental Calculation S1–S3). Assuming equal probabilities of synthesizing each unique nucleotide variant, the expected sampling requirement $E(n)$ for n number of unique variants is given by

$$E(n) = nH_n \tag{1}$$

where n is the number of unique nucleotide variants in the library and H_n is the nth harmonic number. Since n is large, we can approximate H_n using

$$H_n \approx \ln(n) + \gamma \tag{2}$$

where γ is the Euler–Mascheroni constant. Using (1) and (2), we calculate that 1.52×10^7 sequences need to be synthesized to exhaustively sample a four-codon combinatorial library generated using NNK mutagenesis (Supplemental Calculation S4), matching results from a previously reported probabilistic model (Nov 2012). By applying this approach toward an increasing number of codons considered for mutagenesis, we highlight the crux of the problem: that sampling requirement for recovery of all variants in the library increases geometrically with library size, and quickly exceeds the sampling and sequence recovery capabilities of most MAVE screens (Fig. 1).

By taking basic experimental consideration into account, this simple calculation method enables initial approximations of minimum sampling requirements for recovering all possible nucleotide variants in combinatorial mutation libraries to be made but is only a representation of ideal sampling behavior and does not consider uneven representation of oligonucleotides and stochasticity in the sampling process.

Tracking the Sampling Process in Simulations

To better account for problems related to stochasticity in the sampling process, we developed simulations where sequence variants may be generated at unequal frequencies due to random chance. These simulations were designed with all of the assumptions from the above theoretical approximation approach, which allows for direct comparisons to be made. We have also made these simulations flexible to user-specified mutagenesis schemes, number of codons to mutate, and percent recovery of all possible nucleotide sequences in the library as inputs in order to return a list of randomly

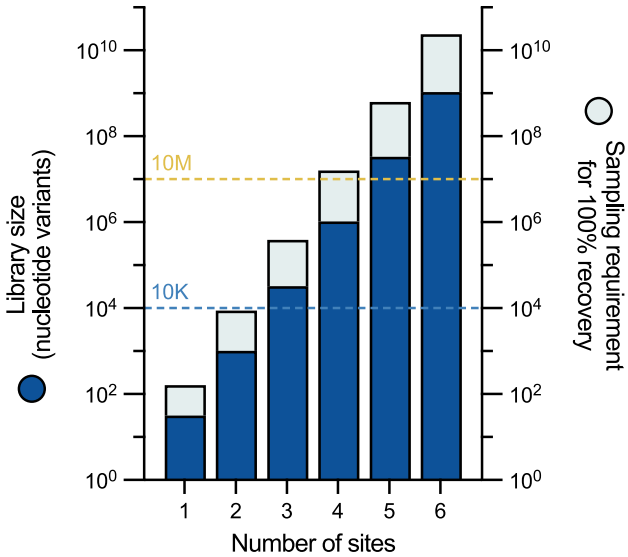


Fig. 1 Theoretical diversity and sampling requirements undergo combinatorial explosion for each additional site/codon considered for mutagenesis. Overlay bar chart showing the theoretical library diversity (dark blue bars) and sampling requirements (light blue bars) as the number of sites combinatorially mutated using NNK codon randomization increase. Dashed lines show the range of typical sampling capabilities of MAVE screens (Color figure online)

generated sequences that satisfy these specifications (Fig. 2A and "Methods").

Using this setup, we simulated the sampling requirements for the above NNK combinatorial library at increasing recovery from 5 to 100%. As expected, sampling requirement increases as recovery increases (Fig. 2B). As recovery approaches 100%, the curve becomes nearly vertical (Fig. 2B), revealing that sampling effort is concentrated within the last 1% to full recovery (Fig. 2C). At full recovery, the average sampling requirement for having sampled all possible nucleotide combinations is $1.67 \times 10^7 \pm 2.32 \times 10^6$ sequences across four replicate simulations (Fig. 2C), a result that is robust to error when compared to the 1.52×10^7 sequences calculated from theory.

To quantify the effects of codon degeneracy on the sampling process, we performed a similar set of simulations but instead constrained percent recovery to test for amino acid sequences encoded by the same codon randomization scheme (see "Methods"). Upon making this change, the sampling requirement for full recovery dropped by 25.7% to $1.13 \times 10^7 \pm 1.85 \times 10^6$ sequences (Fig. 2C). This difference, however, does not scale linearly over the course of the sampling process. Rather, we find that marginal gains in percent recovery demand increasingly greater sampling effort when testing for nucleotide sequences compared to their encoded amino acid equivalents (Fig. 2B).

Having resolved differences in sampling effort in these two scenarios, we sought to understand these differences by examining the composition of sequences generated during each simulation. Given that sampling is done randomly and from a limited pool of unique sequences, as recovery increases it is logical to expect both the frequency that any

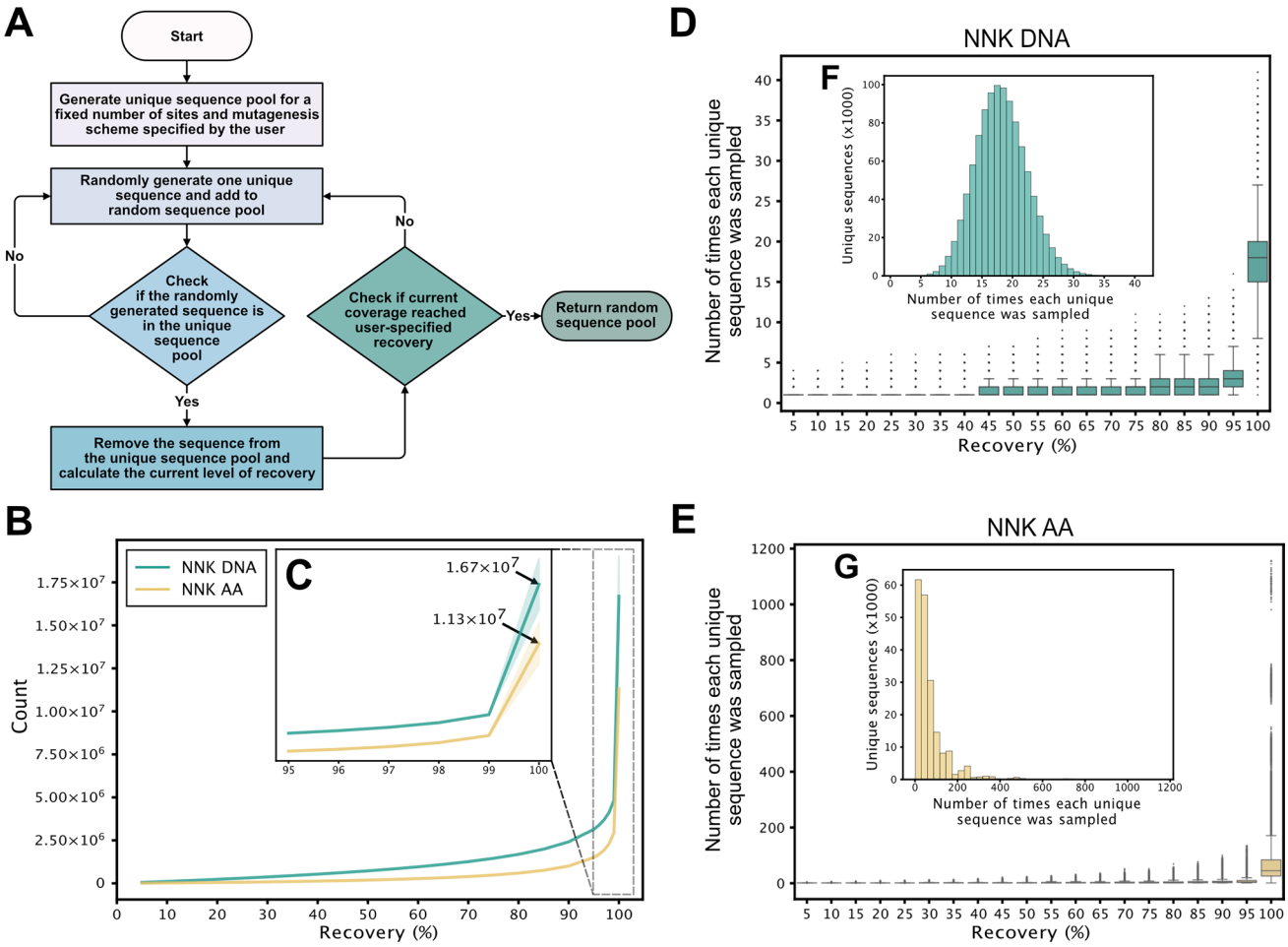


Fig. 2 Simulating the sampling requirements for combinatorial libraries. **A** Algorithm flowchart describing the simulation process. **B** Line plot of sampling requirements simulated for combinatorial libraries that contain four codons mutated using NNK randomization. Lines connect mean sampling requirements ± standard deviation of $n = 4$ replicate simulations at increasing percent recovery of all possible nucleotide (DNA) and amino acid (AA) variants in the library. For the range of 5–90% recovery, sampling requirements were simulated at 5% increments. From 90 to 100%, simulations were conducted at 1% increments. **C** Inset plot of the last 5 percent of sampling to full recovery. **D** and **E** Box plots showing the number of times that each unique sequence was sampled at increasing percent recovery. **F** and **G** Distribution showing the number of times each unique sequence was sampled when sampling for all possible sequences in the library

given sequence is sampled more than once and the number of times that any given sequence is sampled repetitively to increase. This is indeed what we observed, irrespective of whether the simulation was set up to test for nucleotide or amino acid sequences (Fig. 2D, E). However, when sampling for 100% recovery of nucleotide sequences, the number of times each unique sequence was sampled follows a unimodal distribution (Fig. 2F), which is in contrast to the heavily right-skewed distribution generated when sampling for amino acid sequences (Fig. 2G).

To test whether these observations hold up against combinatorial libraries created using other mutagenesis schemes, we performed an equivalent set of calculations and simulations for a library generated using NNB randomization, where we found consistent results (Supplemental Calculation S5 and Supplemental Fig. S1).

Collectively, these findings confirm that sampling effort becomes increasingly expensive as demand for sampling sequence variants approaches 100%. Although this problem can be partially remedied by testing for amino acid sequences rather than nucleotide, sampling bias effects introduced by codon degeneracy remain a key feature of these types of randomization schemes.

Experimental Validation of Minimum Sampling Requirements for Combinatorial Libraries

To test the practical utility of the sampling requirement predicted by theory and simulations in experiments, we devised a method for constructing combinatorial gene variant libraries in the laboratory. To enable direct comparisons between theoretical, simulated, and experimental results, we constructed two combinatorial libraries, Library 1 and 2, each following the above design specifications of mutating four codon positions using NNK mutagenesis.

We generated our libraries by building on a previously reported plasmid library construction method (Galka et al. 2017). Our protocol is different, however, as it makes use of different reagents and includes additional attention to optimizing primer design specifications and PCR cycling conditions (Fig. 3A and "Methods"). These changes enabled us to selectively reduce the proportion of sequences synthesized with primer tandem repeats from as high as 44% to nearly none (Supplementary Fig. S2). This improved design maximizes the number of sequences generated with mutations at the desired positions, which is critical when attempting to construct and screen combinatorial libraries with sequence complexity ranging in the millions. Using the protocol reported here, we were able to generate and recover sequence variants at scale, exceeding our calculated minimum sampling requirement of 1.52×10^7 by fourfold in single reactions (see "Methods"). To further account for the problem of stochastic drift, we performed two replicate constructions of each library, which enabled the number of sequences recovered to exceed the minimum sampling requirement by eightfold (see "Methods" and Supplemental Fig. S3).

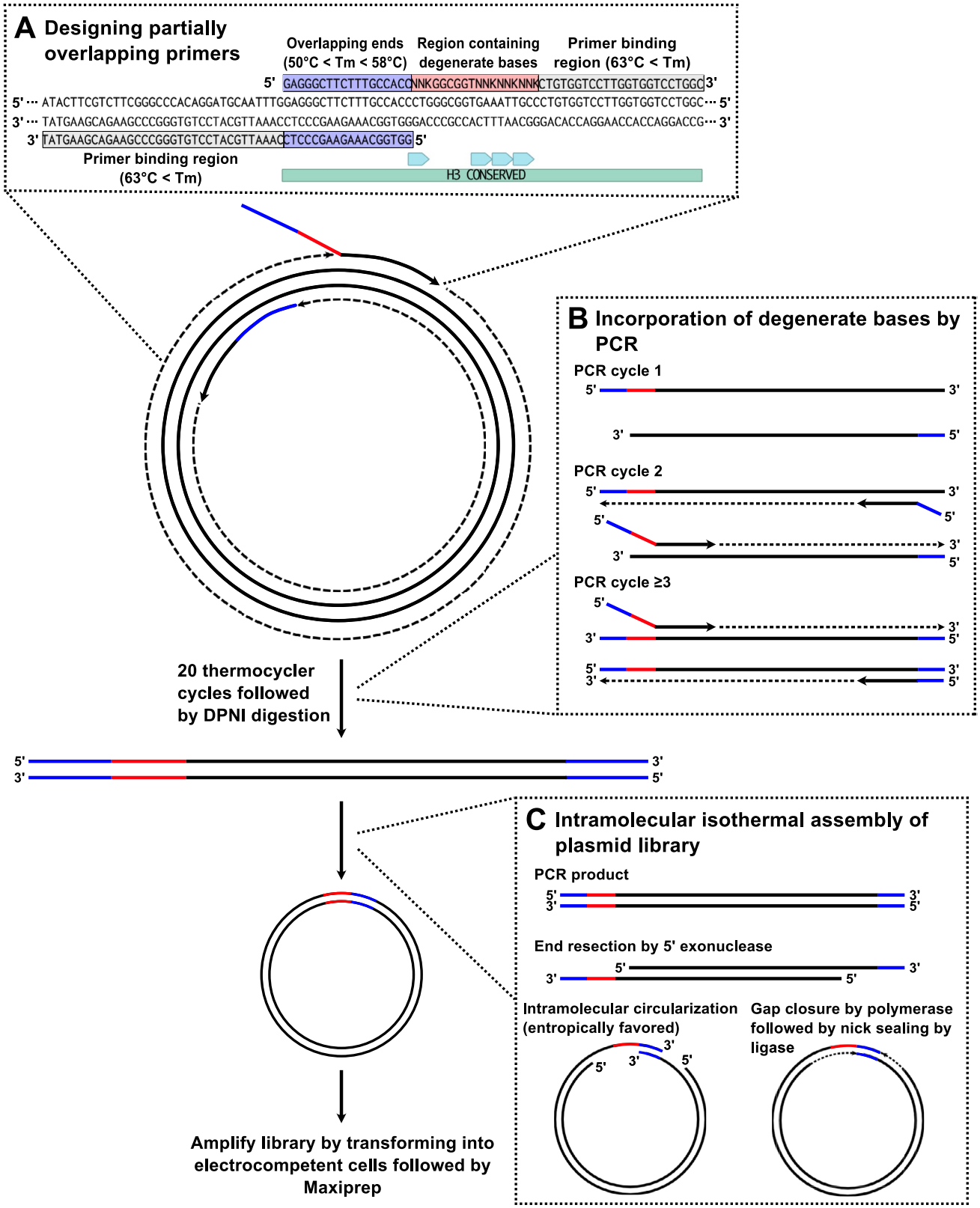
To assess the quality of our variant libraries, we Sanger sequenced 40 colonies drawn from each library. By examining the trace data, we found that every single plasmid that we sequenced contained codon replacements at the designated positions for mutagenesis (Supplemental Fig. S2B). At these positions, reads from three out of the 80 colonies sampled were found to contain multiple peaks at the mutated positions, indicating possible concatemer formation during library assembly. However, re-transformation of the corresponding minipreps resolved these colonies to be multiple vector transformants. Indeed, multiple vector transformants have been found to occur at detectable frequencies when transforming competent bacteria, especially when large amounts of DNA are used (Weston et al. 1979; Goldsmith et al. 2007). However, this should not be an issue here as all of the plasmids recovered during library construction were ultimately purified into the same pool.

Outside the designated positions for mutagenesis, manual inspection of the reads found no mutations, except for a single nucleotide insertion, across all 80 colonies sampled.

Examining the Complexity of Laboratory-Generated Combinatorial Libraries

To determine whether our approach for creating laboratory-generated combinatorial libraries enables exhaustive sampling of all possible sequence variants within the experimental design, we used ultradeep targeted sequencing to sequence the mutated region within each library at an on-target coverage of at least 45 × the library size. To determine the percent recovery of all variants in each library, we adapted analysis methods and algorithms designed for detecting variants from metagenomic data (Masella et al. 2012; Rosen et al. 2012; Edgar 2013; Callahan et al. 2016), based on an error model that considers per-base quality information, concordance between read pairs, and abundance thresholds to suppress errors ("Methods"). After error filtering, we identified 99.1% of all possible nucleotide variants, which encode 99.7% of all amino acid variants in Library 1 (Supplemental Table S4). For Library 2, we identified 99.0% of all possible nucleotide variants and 100% of all amino acid variants (Supplemental Table S5). For each of the libraries, mutagenesis was specific to the designated positions and base frequencies are as expected for NNK codon randomization (Fig. 4A, B).

Although minor variations exist between the two libraries, the amino acid frequencies encoded by the mutagenized codons are also as expected despite the reduced but nevertheless present levels of codon degeneracy inherent



to NNK randomization (Fig. 4C, D). Amino acids encoded by the greatest number of codons—Leucine, Serine, and Arginine—are among the top represented amino acids at the randomized sites but are not overly inflated as would be expected if PCR amplification biases severely reduced uniformity. Moreover, we found that abundances of base

◂**Fig. 3** Laboratory construction of a combinatorial library using degenerate oligonucleotides. **A** Primer design principles for the incorporation of degenerate bases. The forward primer contains a primer binding region (beige), a region containing degenerate bases (red), and an overlapping end sequence (blue), while the reverse primer only contains a primer binding region and an overlapping end sequence. The primer binding regions are designed to have higher melting temperatures compared to the overlapping regions to reduce primer dimerization. **B** The process of incorporating degenerate bases (red) and overlapping ends (blue) into the sequence by PCR. The resulting PCR products contain degenerate bases at the targeted sequence and overlapping regions at each end. **C** Circularization of the PCR products into double-stranded plasmids using isothermal assembly, which involves resection at the 5′ ends, intramolecular circularization, polymerase-mediated gap closure, and nick sealing by a ligase. The assembled plasmid library is then transformed into competent cells and maxiprepped to form the variant library (Color figure online)

combinations identified in both of the independently constructed libraries are correlated (Fig. 4E), indicating reproducible mutagenesis efficiencies despite differences in target sequences. One key factor that helped maintained uniformity is minimizing the number of PCR cycles used to add adapters and indexes for sequencing ("Methods"), a strategy employed in single-cell sequencing studies to preserve copy number information (Zahn et al. 2017) that should be adopted as standard practice in MAVE studies.

To compare our experimental results against ideal sampling behavior, we computationally downsampled the number of reads to match the minimum sampling requirements determined using theory and simulations (Table 1). In the downsampled dataset, Library 1 preserved 85.2% of all nucleotide variants and 94.6% of all amino acid variants, while 93.6% of all nucleotide variants and 99.8% of all amino acid variants were identified for Library 2. Doubling the number of reads to two times the minimum sampling requirement enables marginally more variants to be recovered, which is reflective of diminishing returns observed in our simulation trajectories (Fig. 2B).

Taken together, these results constitute a complete workflow that starts with defining the sampling criteria for testing combinatorial libraries, followed by a demonstration of how our proposed calculation and simulation methods can be applied in a new experimental protocol for constructing combinatorial libraries. Our approach ensures capture of the diversity of sequences envisioned in the experimental design.

Discussion

In this study, we brought together combinatorial statistics calculations, simulations, and molecular biology experiments to produce high-quality combinatorial variant libraries that contain all possible combinations of amino acid replacements for fixed number of amino acid sites. Having confirmed the usefulness of our approximation approaches for generating combinatorial libraries with exhaustive or near-exhaustive levels of variant recovery, here we discuss the implications of our findings for pooled approaches in general and their possible applications toward developing combinatorial MAVE studies for large-scale testing of molecular fitness landscapes.

Benefits and Costs of Pooled Approaches

Our assessment of the sampling criteria for assaying multi-site combinatorial libraries highlights key benefits and drawbacks associated with pooled experimental approaches. Although pooled approaches allow for an extremely large number of variants to be created and tested in multiplex, the sampling process comes at the cost of being inherently random. Due to this randomness, the probability of sampling the same variant multiple times must be considered when determining the sampling requirement for full recovery. The resulting cost in sampling efficiency is accounted for in our calculations here and in other probabilistic models developed for discovering top-performing variants in directed evolution experiments (Patrick et al. 2003; Bosley and Ostermeier 2005; Firth and Patrick 2008; Kong 2009; Nov 2012). For the four-codon combinatorial library investigated here, our calculated minimum sampling requirement for full recovery exceeds library size by ~ 1.5 orders of magnitude, indicating that most variants will likely be sampled multiple times—an expectation that was also confirmed in our simulations and laboratory experiments. This difference between minimum sampling requirement and library size will only grow as the number of mutated codons increases. Although this problem of diminishing returns can be partially addressed by using more efficient mutagenesis schemes, such as MAX (Hughes et al. 2003, 2005), SILM (Tang et al. 2012), and the 22c-trick (Kille et al. 2013), these library synthesis methods are more complicated to implement and are also inherently random. However, it is also important to recognize that experimenters may choose not to use mutagenesis schemes that minimize codon usage to near one codon per amino acid as this would preclude the possibility of testing the effects of synonymous codons in MAVE experiments. For these reasons, we designed our simulation software to be amenable to various mutagenesis schemes, leaving these design choices to users themselves. In summary, we posit that the foremost barrier to exhaustively testing combinatorial libraries using MAVEs is contingent on experimental design and the inherently random sampling process rather than a lack of experimental throughput.

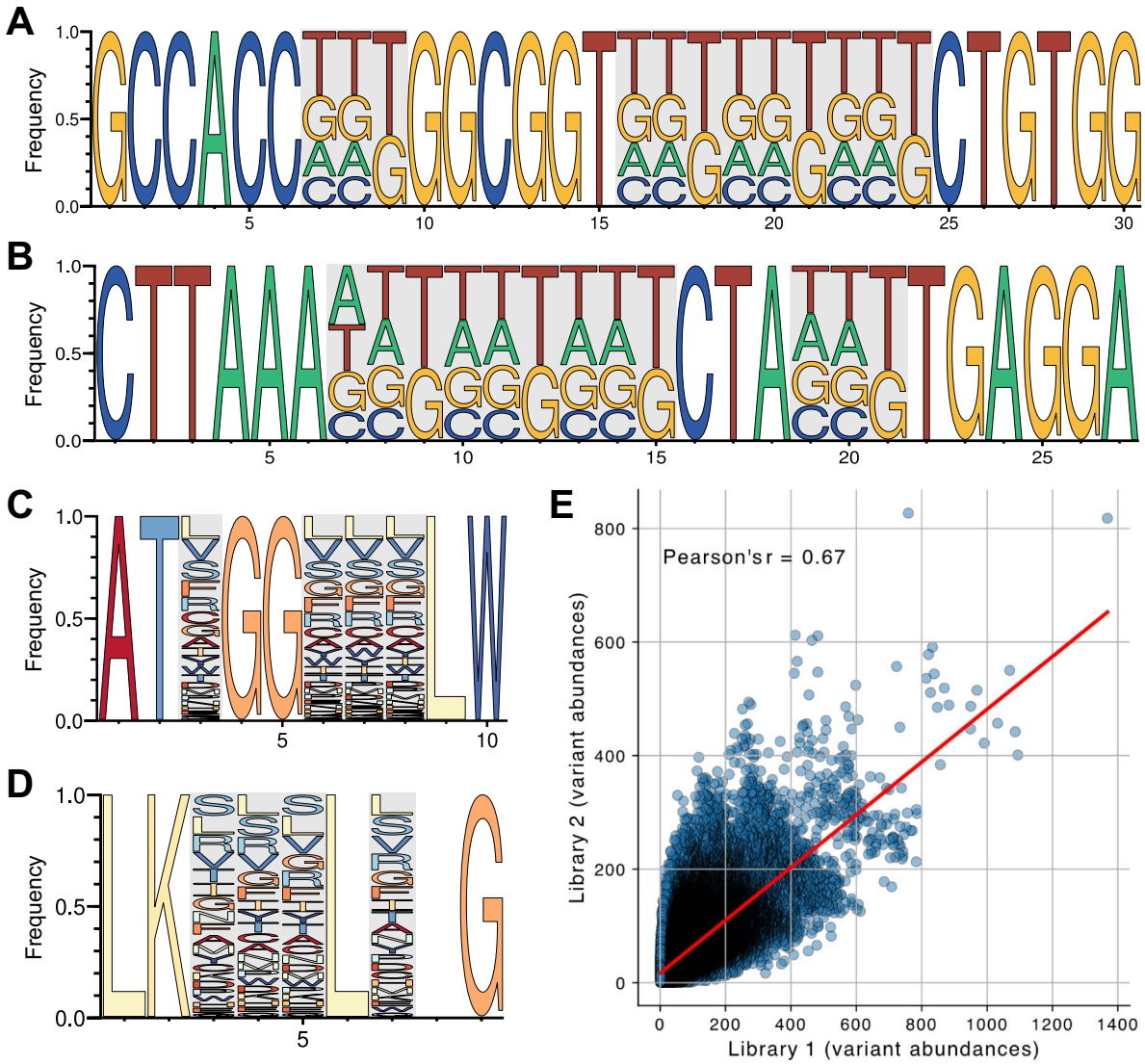


Fig. 4 Base compositions of mutated codons in combinatorial libraries and their encoded amino acids distributions. **A** Logo plot showing the base frequencies at combinatorially mutated codons and surrounding base positions in Library 1. Designated positions for codon replacement are shaded gray. **B** The same as in **A** except for Library 2. **C** and **D** Amino acid frequencies encoded by the libraries shown in **A** and **B**, respectively. Invariant positions encoded by stop codons are blank as stop codons do not encode any amino acid. **E** Dot plot showing abundances of base combinations identified in both of the constructed libraries

Table 1 Effect of downsampling read counts to match minimum sampling requirements

	Sampling depth (multiple of minimum sampling requirement)	Unique DNA variant calls	Recovery of DNA variants (%)	Unique amino acid variant calls	Recovery of amino acid variants (%)
Library 1	1 ×	893,874	85.2	184,055	94.6
	2 ×	1,004,056	95.8	191,845	98.6
Library 2	1 ×	981,959	93.6	194,091	99.8
	2 ×	1,032,594	98.5	194,478	100.0

Exhaustive and Near-Exhaustive Sampling of Genotype Space

By incorporating the statistical assumptions underlying our calculations into simulation algorithms, we expand beyond theoretical calculations to quantify sampling requirements directly. Simulations allow for stochasticity in the sampling process and for variability among replicate sampling trajectories to be assessed. Notably, when we changed the parameters of our simulation to sample the library at increasing levels of recovery, simulation trajectories show that the vast majority of sampling effort is concentrated within the last 1% of full recovery, which indicates that sampling effort can be significantly reduced with minimal losses in recovery. Moreover, we show quantitatively that sampling effort is substantially reduced by sampling for amino acid sequences instead of nucleotide. For studies of evolution aimed at mapping molecular fitness landscapes, testing 99% of all possible variant combinations may be sufficient, given the immense sampling cost associated with sampling every last variant (Nov 2012).

Our simulations approach provided valuable insights into the combinatorial mutation library sampling process but can be improved further by incorporating user-specified abundance thresholds for overcoming issues related to stochastic drift and additional parameters that account for biases in oligonucleotide synthesis and binding efficiencies. These improvements will be of interest to experimenters and will be made available in future versions of our simulation program.

Calculated and Simulated Sampling Requirements are Useful Guides for Designing Combinatorial MAVE Experiments

Testing our calculations and simulations in laboratory experiments informs us about the practical usefulness of our approximations. To conduct such a test, we built on a previously published plasmid library construction method (Galka et al. 2017) by changing experimental conditions and reagents to significantly improve the proportion of sequences synthesized with mutations at the desired positions. Using the protocol presented here, we created two four-codon combinatorial libraries, with one library achieving 99.7% of all possible amino acid combinations and the other achieving 100%. Given the high efficiency of our experimental protocol, the number of sequences synthesized, recovered, and deep sequenced here surpassed our predicted minimum sampling requirements for full recovery. This is generally recommended for MAVE experiments in order to account for stochastic drift and experimental errors (Fowler et al. 2014; Faure et al. 2020). Thus, in practice, the approximated minimum sampling requirements are useful guides. However, these approximations are indeed 'minimum' and should be exceeded whenever possible to account for randomness and to maximize variant recovery.

Even for high-efficiency experimental methods, exhaustive or near-exhaustive sampling of combinatorial libraries beyond the fourth codon will be challenging due to further geometric increases in variant combinations. Despite this, exhaustively mutating and characterizing only a few codons already enable a much larger proportion of genotype space to be explored over what was conventionally possible and is likely to yield significant discoveries. We anticipate that improvements in DNA synthesis technologies coupled with further decreases in sequencing costs will enable ever-larger molecular fitness landscapes to be mapped using MAVE experiments.

A Priori Determination of Sampling Requirements May also Help Inform How Bottlenecks Can Be Overcome at all Major Experimental Steps in MAVE Workflows

Our methods for determining sampling requirements are useful for constructing high-complexity combinatorial mutation libraries, but may also help inform how other bottlenecks in MAVE experiments can be overcome. Possible points in MAVE workflows where bottlenecks may arise include synthesis of the input library either by amplification or subcloning, transfection of the library into cells, selection, post-selection extraction of the library, sample preparation for sequencing, and deep sequencing (Faure et al. 2020). Minimum sampling requirements must be exceeded for all of these steps in order to ensure a high probability of capturing all or most sequence variants. For fluorescence-activated cell sorting (FACS)-type MAVE experiments, meeting minimum sampling requirements as defined here may be sufficient for assessing how most variants have separated. However, growth-based MAVE experiments may have even higher sampling demands for accurately capturing negative changes in variant frequencies (Fowler et al. 2014). Further modeling of sampling requirements for all major experimental steps and selection methods is needed to accommodate different MAVE designs and the ever-increasing size of combinatorial MAVE experiments.

Materials and Methods

Simulations

We developed a Java script to simulate the sampling requirements for combinatorial gene variant libraries using user-specified mutagenesis schemes (NNK, NNB, NNN, etc.), number of codons to mutate, and percent recovery of all

possible sequences in the library (Fig. 2A). Using these parameters, the program will first generate a 'unique sequence pool,' comprising of all possible nucleotide sequences in the library. Next, a random nucleotide sequence will be generated, added to a 'random sequence pool,' and compared to the sequences in the unique sequence pool. If the randomly generated sequence is found in the unique sequence pool, the sequence will be removed from the unique sequence pool and the current level of variant recovery will be calculated. Until the user-specified percent recovery is reached, the program will continue to generate random sequences by iterating through this sampling process.

The script can also be used in 'AA mode' to account for codon degeneracy by determining the sampling requirements for amino acid variants. This was done by introducing an extra step where the randomly generated nucleotide sequence is converted to its corresponding amino acid sequence and compared to a sequence pool comprising all possible amino acid variants in the library. In this case, user-specified recovery is calculated from a pool of all possible amino acid variants instead of a pool of all possible nucleotide variants.

Combinatorial Variant Library Construction

We created a heavily modified version of the QuikLib protocol (Galka et al. 2017) for building synthetic plasmid libraries. Bacterial glycerol stocks containing plasmids with our target sequences for mutagenesis were grown overnight at 37 °C in LB liquid cultures with the appropriate antibiotic. Plasmids were purified using GenElute™ Plasmid DNA Miniprep Kit (Millipore Sigma, Burlington, MA, USA) and used as the input DNA template for PCR during library construction.

To design primers for library construction, we ordered PAGE Ultramer™ DNA oligonucleotides (Integrated DNA Technologies, Coralville, IA, USA) that contain degenerate bases at specific codon positions. In our primer design, we incorporated design principles originally developed for site-directed mutagenesis (Liu and Naismith 2008) that limit the number of overlapping bases between the forward and reverse primers to decrease primer dimerization (Fig. 3A). PCR was set up using reagents from KOD Xtreme™ Hot Start DNA Polymerase kit (MilliporeSigma, Burlington, MA, USA) according to the manufacturer's protocol, except we replaced the KOD Xtreme™ DNA Polymerase with Q5 Hot Start High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA, USA) to increase DNA replication fidelity. 20 cycles of PCR were run with 10 ng of DNA input using the following conditions: Initial denaturation at 94 °C for 2 min, denaturation at 98 °C for 10 s, annealing at 62.6 °C for 30 s, elongation at 68 °C for 1 min per kilobase of plasmid, final elongation at 6 °C for 5 min, and indefinite hold at 4 °C. To suppress primer dimerization and the incorporation of tandem primer repeats (Supplemental Fig. S4A, B), the annealing temperature used during PCR exceeded the melting temperature of the overlapping ends but not that of the primer binding regions (Fig. 3A). The resulting PCR products contain overlapping ends and degenerate bases at the target site (Fig. 3B).

After PCR, the reaction products were treated with DPNI (New England Biolabs, Ipswich, MA, USA), run on a 1% agarose gel to confirm single DNA bands, and column purified using QIAquick PCR purification kit (Qiagen, Hilden, NRW, Germany). To circularize the PCR products, we added 2 µg of purified PCR product to 200 µL of 2 × NEBuilder HIFI DNA Assembly Master Mix (New England Biolabs, Ipswich, MA, USA), topped-up the final volume to 400 µL with water, and incubated the reaction at 50 °C for 1 h. In the NEBuilder reaction, the overlapping ends, incorporated during PCR, enable intramolecular ligation and circularization of the PCR products into double-stranded plasmids (Fig. 3C). Circularized PCR products were purified using DNA Clean & Concentrator Kit-5 (Zymo Research, Irvine, CA, USA) followed by transformation into 50 µL of NEB 10-beta Electrocompetent *E. coli* (New England Biolabs, Ipswich, MA, USA) with a 2.0 kV electroporation pulse in a 0.1 mm cuvette. For each variant library, cells from two replicate transformations were pooled and inoculated into 500 mL of LB liquid media containing the appropriate antibiotic. From these 500 mL cultures, aliquots were serially diluted and test plated on selective media to estimate transformation yield (Supplemental Fig. S3), which averaged 1.28×10^8 transformants across both variant libraries. The remaining cultures were grown overnight for 16 h and subsequently maxiprepped using PureLink™ HiPure Plasmid Maxiprep Kit (Invitrogen, Waltham, MA, USA) to form the final variant libraries.

Testing the Quality of Sequences in the Constructed Libraries

Transformants plated during library construction were grown into colonies, picked, grown in selective media, miniprepped using QIAprep® Spin Miniprep Kit (Qiagen, Hilden, NRW, Germany), and Sanger sequenced (Eurofins Genomics, Huntsville, AL, USA). Chromatograms were aligned and analyzed using CLC Genomics Workbench v 22.0.1.

Generating Amplicons for Deep Sequencing

To survey the sequence diversity for each variant library, we generated amplicons of the mutated sites using custom PAGE Ultramer™ DNA primers (Integrated DNA Technologies, Coralville, IA, USA) that contain all the necessary

elements (Supplemental Tables S2, S3) for direct sequencing on Illumina flow cells. For the first variant library, primers (Supplemental Table S2) were designed to bind 58 bp upstream and 64 bp downstream from the 18 bp region containing four combinatorially mutated codons (Supplemental Fig. S4A). For the second library, primers (Supplemental Table S3) were designed in the same way to survey a 15 bp region also containing four combinatorially mutated codons. For each variant library, four primer pairs, each containing unique index barcodes and heterogeneity spacers of variable length, were used in a combinatorial indexing matrix (Supplemental Fig. S4B) to generate 16 uniquely indexed amplicon pools. Heterogeneity spacer sequences were designed using a custom script to counteract the issue of low nucleotide diversity during Illumina sequencing (see "Data and resource availability" section). We first tested these primers for non-specific binding in 30 cycle PCRs using NEBNext Ultra™ II Q5 Master Mix (New England Biolabs, Ipswich, MA, USA) with the following conditions: Initial denaturation at 98 °C for 30 s, denaturation at 98 °C for 15 s, annealing at 69 °C for 15 s, elongation at 72 °C for 10 s, final elongation at 72 °C for 2 min, and indefinite hold at 4 °C. After confirming specific amplification of the target regions (Supplemental Fig. S4C), we conducted eight cycles of minimal PCR using the same kit and thermocycling conditions to generate amplicons (Supplemental Fig. S4D) for deep sequencing.

Amplicon Sequencing

Amplicon processing and sequencing were performed by the Centre for the Analysis of Genome Evolution & Function at the University of Toronto. The amplicons were first purified using 0.9× ratio of AMPure® XP magnetic beads (Beckman Coulter, Brea, CA, USA) followed by quantification using Qubit™ dsDNA HS kit (Thermo Fisher, Waltham, MA, USA). The samples were then pooled in equal amounts, followed by dual size selection with 0.7 ×/0.2 ratios of AMPure beads. The pooled samples were then denatured and diluted to a sequencing concentration of 650 pM and sequenced on an Illumina NextSeq™1000 using the P1 chemistry with 150 × 2 Paired-End reads (Illumina, San Diego, CA, USA). 5% PhiX spike-in was included during sequencing. Raw data generated from the sequencing run were converted to FASTQ format using BCL convert v 4.0.3 (Illumina, San Diego, CA, USA).

Sequence Quality Filtering, Trimming, and Merging

The deep sequencing data were filtered to remove reads that contain ambiguous base calls and/or expected errors (Edgar and Flyvbjerg 2015) using parts of the DADA2 package v 1.26.0 (Callahan et al. 2016). Filtered reads were then globally trimmed using Cutadapt v 4.2 (Martin 2011) to remove primer sequences, heterogeneity spacers, and low-quality base positions (Phred + 33 score lower limit < 28) at the beginning and end of reads, leaving only our target region, comprising of the mutated codons and surrounding invariant bases. Since the Paired-End read lengths used during sequencing exceed the length of our target regions for mutagenesis, the trimmed forward and reverse reads within each read pair overlap completely. Any read pairs that did not overlap completely were excluded from the analyses. Read pairs that passed quality filtering and trimming were merged using PANDAseq (Masella et al. 2012). Merging using either the default simple_bayesian (Masella et al. 2012) or the UPARSE/USEARCH (Edgar 2013) algorithm implementations in PANDAseq yielded highly similar (± 0.005%) number of unique sequences. Merged reads were then converted to FASTA format and used to generate logo plots using WebLogo (Crooks et al. 2004) to show nucleotide and amino acid frequencies (Fig. 4).

Determining Recovery of Variants in the Library

To determine the percent recovery of all possible nucleotide and amino acid sequences in each combinatorial library, merged reads were first put through a 'NNK filter' to remove sequences that did not conform to NNK randomization at the designated positions for mutagenesis. Next, we inferred the proportion of false positive variant calls given all reads at various abundance thresholds from the number of mismatch base calls at invariant base positions (Supplemental Tables S4, S5 and Supplemental Fig. S5). After looking at the abundance distributions, as previously suggested (Rosen et al. 2012), we set an arbitrary but computationally necessary 'abundance filter' that removed low abundance reads (detected < 4 ×) to decrease the proportion of false positives (Supplemental Fig. S5). After filtering, the remaining reads were used to call variants and calculate the proportion of all nucleotide and amino acid sequences detected in each library. The effects of these filtering steps are summarized (Supplemental Tables S4, S5).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-024-10179-8.

Acknowledgements We thank Frederick P. Roth for comments and suggestions, and the Centre for the Analysis of Genome Evolution and Function at the University of Toronto for providing deep sequencing services and consultations on sequencing strategies. This work was supported by a NSERC Discovery grant to B.S.W.C., and an Ontario Graduate Scholarship and a Vision Science Research Program Fellowship to S.K.C.

Author Contributions S.K.C. and B.S.W.C. conceived and designed the study; S.K.C. and J.L. performed calculations, in silico simulations, in vitro experiments, and bioinformatics analyses; B.M.T.P. provided deep sequencing primer design; A.V.N. contributed to the

bioinformatics analyses; S.K.C. wrote the manuscript with input from J.L., A.V.N., and B.S.W.C.; and B.S.W.C. supervised the study. All co-authors read and approved the manuscript.

Data Availability Code for simulating sampling requirements is available at https://github.com/belindachang-lab/Comb_lib_Mutagenesis_Simulator. Code for generating heterogeneity spacers can be found at https://github.com/belindachang-lab/Comb_lib_hetero_gen_spacer. The computational pipeline for analyzing the deep sequencing data is available at https://github.com/belindachang-lab/Comb_lib_NGS_analysis. The colony segmentation and counting tool for determining transformation efficiencies can be accessed at https://github.com/belindachang-lab/Comb_lib_ColonyCounter. All raw deep sequencing reads generated in this study have been submitted to Zenodo (https://zenodo.org) under the accession 8356598.

References

Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B (2019) Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. Cell 176:549-563.e23

Blanco C, Janzen E, Pressman A, Saha R, Chen IA (2019) Molecular fitness landscapes from high-coverage sequence profiling. Annu Rev Biophys 48:1–18

Bosley AD, Ostermeier M (2005) Mathematical expressions useful in the construction, description and evaluation of protein libraries. Biomol Eng 22:57–61

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583

Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190

Domingo J, Diss G, Lehner B (2018) Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature 558:117–121

Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods 10:996–998

Edgar RC, Flyvbjerg H (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–3482

Faure AJ, Schmiedel JM, Baeza-Centurion P, Lehner B (2020) DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. Genome Biol 21:207

Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B (2022) Mapping the energetic and allosteric landscapes of protein binding domains. Nature 604:175–183

Firth AE, Patrick WM (2008) GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in randomized libraries. Nucleic Acids Res 36:W281–W285

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S (2010) High-resolution mapping of protein sequence-function relationships. Nat Methods 7:741–746

Fowler DM, Stephany JJ, Fields S (2014) Measuring the activity of protein variants on a large scale using deep mutational scanning. Nat Protoc 9:2267–2284

Galka P, Jamez E, Joachim G, Soumillion P (2017) QuickLib, a method for building fully synthetic plasmid libraries by seamless cloning of degenerate. PLoS ONE 12:e0175146

Goldsmith M, Kiss C, Bradbury ARM, Tawfik DS (2007) Avoiding and controlling double transformation artifacts. Protein Eng Des Sel 20:315–318

Hughes MD, Nagel DA, Santos AF, Sutherland AJ, Hine AV (2003) Removing the redundancy from randomised gene libraries. J Mol Biol 331:973–979

Hughes MD, Zhang Z-R, Sutherland AJ, Santos AF, Hine AV (2005) Discovery of active proteins directly from combinatorial randomized protein libraries without display, purification or sequencing: identification of novel zinc finger proteins. Nucleic Acids Res 33:e32–e32

Kille S, Acevedo-Rocha CG, Parra LP, Zhang Z-G, Opperman DJ, Reetz MT, Acevedo JP (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. ACS Synth Biol 2:83–92

Kong Y (2009) Calculating complexity of large randomized libraries. J Theor Biol 259:641–645

Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. Science 1979(352):837–840

Liu H, Naismith JH (2008) An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. BMC Biotechnol 8:91

Lozovsky ER, Chookajorn T, Brown KM, Imwong M, Shaw PJ, Kamchonwongpaisan S, Neafsey DE, Weinreich DM, Hartl DL (2009) Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proc Natl Acad Sci 106:12025–12030

Lunzer M, Miller SP, Felsheim R, Dean AM (2005) Evolution: the biochemical architecture of an ancient adaptive landscape. Science 1979(310):499–501

Malcolm BA, Wilson KP, Matthews BW, Kirsch JF, Wilson AC (1990) Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature 345:86–89

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBNET J 17:10

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. BMC Bioinform 13:31

Meini M-R, Tomatis PE, Weinreich DM, Vila AJ (2015) Quantitative description of a protein fitness landscape based on molecular features. Mol Biol Evol 32:1774–1787

Newman DJ (1960) The double dixie cup problem. Am Math Mon 67:58

Nov Y (2012) When second best is good enough: another probabilistic look at saturation mutagenesis. Appl Environ Microbiol 78:258–262

O'Maille PE, Malone A, Dellas N, Andes Hess B, Smentek L, Sheehan I, Greenhagen BT, Chappell J, Manning G, Noel JP (2008) Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. Nat Chem Biol 4:617–623

Olson CA, Wu NC, Sun R (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Curr Biol 24:2643–2651

Patrick WM, Firth AE, Blackburn JM (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. Protein Eng Des Sel 16:451–457

Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445:383–386

Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, Kudla G (2016) Molecular evolution: network of epistatic interactions within a yeast snoRNA. Science 1979(352):840–844

Rollins NJ, Brock KP, Poelwijk FJ, Stiffler MA, Gauthier NP, Sander C, Marks DS (2019) Inferring protein 3D structure from deep mutation scans. Nat Genet 51:1170–1176

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

Rosen MJ, Callahan BJ, Fisher DS, Holmes SP (2012) Denoising PCR-amplified metagenome data. BMC Bioinformatics 13:283

Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O et al (2016) Local fitness landscape of the green fluorescent protein. Nature 533:397–401

Schmiedel JM, Lehner B (2019) Determining protein structures using deep mutagenesis. Nat Genet 51:1177–1186
Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nat Rev Genet 11:572–582
Starita LM, Ahituv N, Dunham MJ, Kitzman JO, Roth FP, Seelig G, Shendure J, Fowler DM (2017) Variant interpretation: functional assays to the rescue. Am J Hum Genet 101:315–325
Starr TN, Picton LK, Thornton JW (2017) Alternative evolutionary histories in the sequence space of an ancient protein. Nature 549:409–413
Tabet D, Parikh V, Mali P, Roth FP, Claussnitzer M (2022) Scalable functional assays for the interpretation of human genetic variation. Nature 56:441–465
Tang L, Gao H, Zhu X, Wang X, Zhou M, Jiang R (2012) Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. Biotechniques 52:149–158
Tufts DM, Natarajan C, Revsbech IG, Projecto-Garcia J, Hoffmann FG, Weber RE, Fago A, Moriyama H, Storz JF (2015) Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude Pikas. Mol Biol Evol 32:287–298
Weile J, Roth FP (2018) Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. Hum Genet 137:665–678
Weinreich DM, Delaney NF, Depristo M, a,, Hartl DL. (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312:111–114
Weinreich DM, Lan Y, Wylie CS, Heckendorn RB (2013) Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev 23:700–707
Weinreich DM, Lan Y, Jaffe J, Heckendorn RB (2018) The influence of higher-order epistasis on biological fitness landscape topography. J Stat Phys 172:208–225
Weston A, Humphreys GO, Brown MGM, Saunders JR (1979) Simultaneous transformation of *Escherichia coli* by pairs of compatible and incompatible plasmid DNA molecules. Mol Gen Genet 172:113–118
Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. Elife. https://doi.org/10.7554/eLife.16965
Yang G, Anderson DW, Baier F, Dohmen E, Hong N, Carr PD, Kamerlin SCL, Jackson CJ, Bornberg-Bauer E, Tokuriki N (2019) Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. Nat Chem Biol 15:1120–1128
Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, Aparicio S, Hansen CL (2017) Scalable whole-genome single-cell library preparation without preamplification. Nat Methods 14:167–173