uSort-M: Scalable isolation of user-defined sequences from diverse pooled libraries

Micah B. Olivas[a,†], Patrick J. Almhjell[b,†], Jack D. Shanahan[a], and Polly M. Fordyce[a,c,d,e,*]

[a]Department of Genetics, Stanford University, Stanford, CA 94305
[b]Department of Biochemistry, Stanford University, Stanford, CA 94305
[c]Department of Bioengineering, Stanford University, Stanford, CA 94305
[d]Chan Zuckerberg Biohub, San Francisco, CA 94110
[e]Sarafan ChEM-H, Stanford University, Stanford, CA 94305
*Corresponding author
[†]These authors contributed equally

Abstract

High-throughput sequencing and computational protein design have created a growing gap between the discovery of new proteins and their functional characterization. In many instances, functional characterization requires one-to-one measurements—such as when detailed biochemical insights are desired or pooled selections are not possible—necessitating that individual variants be isolated and assayed. A major barrier to closing this gap is the cost to directly synthesize individual genes, which remains prohibitively expensive ($10–100 per sequence) and restricts these studies to small subsets of relevant variants, leaving many sequences without functional annotation. To address this, we developed user-defined Sorted Mutants (uSort-M), which combines pooled DNA synthesis, automated cell sorting of transformed *Escherichia coli*, and long-read sequencing to rapidly isolate and identify variants from diverse libraries. uSort-M can isolate, sequence, and validate individual variants from pooled libraries produced via diverse existing methods including multiplex assembly, error-prone PCR, or pooled site-directed mutagenesis. Sorting single bacterial clones into 384-well plates is efficient: eight plates (3,072 wells) can be filled in 1–2 hours, with up to 90% of wells yielding monoclonal cultures. Commercial long-read sequencing enables accessible, fast, and cost-effective identification of individual sequences from isolated clones while tolerating wide variation in fragment length and diversity across the library. Applying this workflow to a 328-member scanning mutagenesis library of a 300-bp gene recovered 96% of desired variants at fivefold lower cost than traditional synthesis. Numerical simulations identify key parameters governing library recovery and enable accurate prediction of the sampling effort required to achieve target coverage. As library size increases, this workflow offers substantial savings over traditional gene synthesis or cloning. Due to its generalizability, efficiency, and reliance on standard instrumentation, uSort-M removes a key barrier to large-scale protein functional characterization.

Introduction

Advances in DNA sequencing and computational protein design are identifying new sequences of interest at unprecedented scale—from clinical variants and environmental orthologs to de novo designs and high-throughput screening hits. Experimentally characterizing these sequences requires the ability to rapidly and reliably generate hundreds to thousands of sequence-verified gene isolates. Despite advances in DNA synthesis, preparation of "parsed" libraries (where each gene is physically isolated and its sequence known) remains prohibitively expensive, with synthesis typically costing $10–100 per sequence.

An alternative method is to prepare pooled DNA libraries and isolate single variants from the pool by stochastic sampling. This is a common approach in protein engineering, in which degenerate codons are used to create a library of specified diversity from which clones are then randomly tested for activity, often without concomitant sequence-verification.[1] The number of clones tested depends on the library size, with threefold oversampling returning 95% of unique members in an unbiased library.[1–3] This same sampling-based workflow can be applied to isolate and identify DNA variants by incorporating a

sequencing step after parsing.[4] However, existing methods typically require specific library design, specialized equipment, and complicated sequencing pipelines. A highly general, accessible workflow for parsing pooled libraries into a physically isolated, sequence-verified format would remove a major experimental bottleneck and accelerate protein functional characterization at scale.

To address this need, we developed user-defined Sorted Mutants (uSort-M), a workflow that combines pooled DNA synthesis, widely accessible cell-sorting instruments, and recent advances in accurate, multiplexed long-read sequencing to generate isolated, verified DNA constructs assembled within desired plasmids in as few as seven days and for as low as $2.58 per variant (see **Methods** and **Supplemental Methods**). uSort-M eliminates time-consuming colony picking, whether by hand or via single-purpose robotics, and is fully agnostic to input library composition. Because uSort-M uses pooled inputs, the method benefits from continuously decreasing oligonucleotide synthesis costs and advances in pooled assembly methods.[5] We describe the workflow, quantify how each step impacts library recovery efficiency, and directly compare results from Nanopore sequencing to Illumina sequencing-by-synthesis for pooled amplicons. We further provide tools for estimating recovery, simulating effects of critical variables like library skew, and demultiplexing sequencing reads to identify expected library members. We demonstrate the utility of uSort-M by efficiently generating a parsed library from a pool of 328 single mutants with 96.3% coverage. We anticipate that this workflow and accompanying tools will be broadly useful across library types and substantially reduce the time and cost of creating defined collections of isolated constructs.

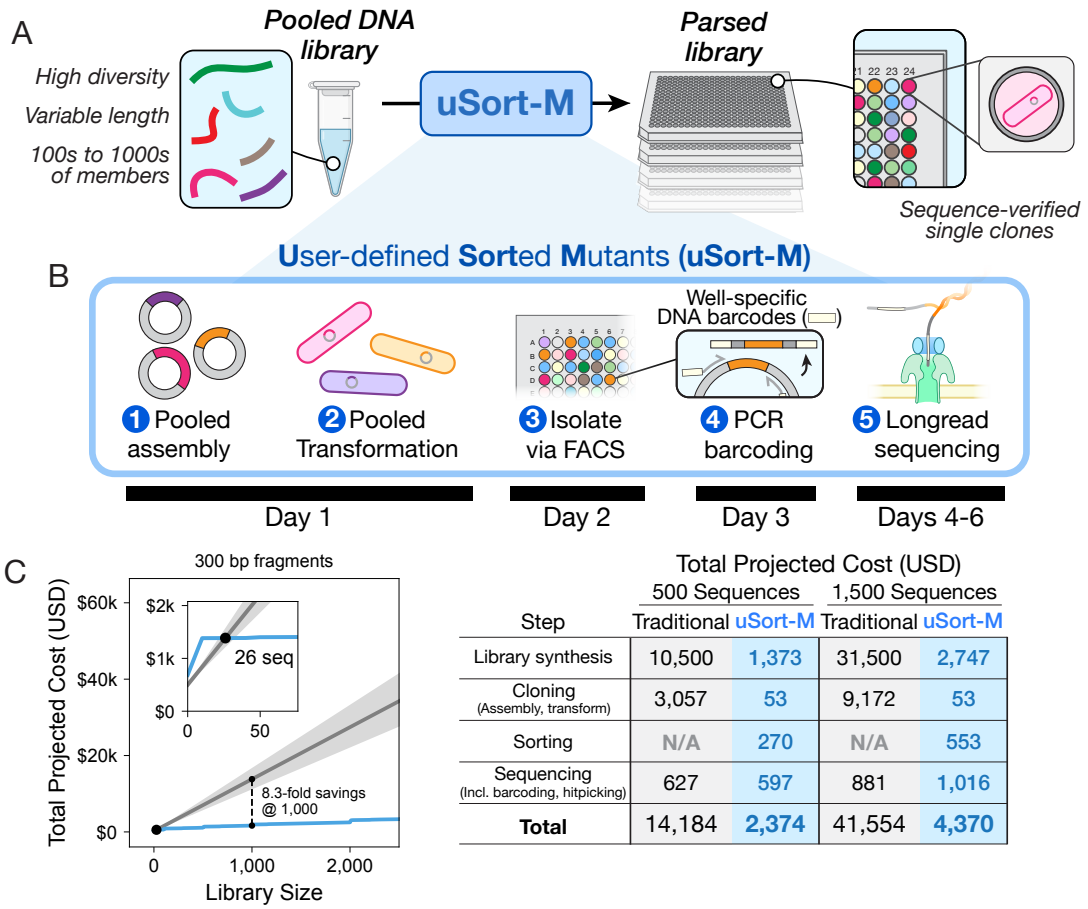
Results

Overview of the uSort-M pipeline.

The uSort-M pipeline rapidly isolates and sequences individual DNA variants from a pre-synthesized oligo pool via stochastic sampling, yielding a final parsed and sequence-validated library (**Figure 1A**). uSort-M involves: 1) assembling a pooled library into a plasmid vector of choice; 2) transforming *E. coli* cells with the pooled library and culturing overnight; 3) sorting individual transformed *E. coli* into wells of 384-well plates via FACS and culturing overnight under antibiotic selection; 4) amplifying DNA templates and appending plate/well-based barcodes via colony PCR;[6,7] and 5) pooling PCR amplicons prior to multiplexed sequencing (**Figure 1B**). Each sequencing read contains the sequence of the variant and a barcode specifying the plate/well-based position, thereby identifying the DNA variant within each cultured well. Desired library members can be selected from these plates and organized for downstream applications.

uSort-M provides two main advantages relative to previously published techniques. First, most prior approaches isolate single transformed clones from a pool by either manual colony picking (which is labor-intensive) or using a specialized and expensive colony-picking robot.[1,4] uSort-M employs a fluorescence-activated cell sorting (FACS) instrument to isolate single *E. coli* cells directly from liquid culture. Commercial and widely available FACS instruments can sort single bacteria at high rates (~70 wells per minute in single-cell sorting mode) even in the absence of a fluorescent signal, making this a highly general approach. Second, most prior approaches have sequenced amplicons using Illumina-based sequencing that limits read lengths to ~300–700 nucleotides (paired-end 150- or 600-cycle MiSeq v3 kits).[4] The need to associate DNA barcodes with sequence information throughout an amplicon either limits amplicon sizes to ~100 amino acids or requires additional transposase-based fragmentation. uSort-M instead leverages Nanopore sequencing, which returns reads that can span >25 kilobases[8,9] and is available through several commercial providers (e.g., Plasmidsaurus) with turnaround times ranging from overnight to a few days for tens of thousands or multi-gigabase (Gb) read depths, respectively.

uSort-M becomes more cost-effective than traditional gene fragment synthesis for 300-bp libraries with as few as 26 constructs (**Figure 1C** and **Supporting Information, Figure S1** and **Table S1**). As library size increases, savings grow substantially, reaching ~80% cost reduction at 1,000 variants.



Step	Total Projected Cost (USD) 500 Sequences Traditional	500 Sequences uSort-M	1,500 Sequences Traditional	1,500 Sequences uSort-M
Library synthesis	10,500	1,373	31,500	2,747
Cloning (Assembly, transform)	3,057	53	9,172	53
Sorting	N/A	270	N/A	553
Sequencing (Incl. barcoding, hitpicking)	627	597	881	1,016
Total	14,184	**2,374**	41,554	**4,370**

Figure 1. Overview of the uSort-M pipeline. A. uSort-M converts an arbitrary pooled library of DNA variants into a "parsed" library where each variant is isolated in its own well with validated sequence information. **B.** uSort-M isolates single plasmid variants within a bacterial host via pooled assembly (1) and transformation (2), isolates individual bacteria via high-throughput cell sorting (FACS) (3), amplifies and barcodes DNA amplicons from cultured clones via PCR (4), and then pools amplicons for multiplexed long-read sequencing (5) that associates plate/well-specific barcodes with DNA amplicons. **C.** Projected costs of traditional library preparation *vs.* uSort-M. *Left*: cost as a function of library size for traditional synthesis methods (light grey shading indicates estimated minimum and maximum costs; grey line indicates mean) *vs.* u-Sort-M (blue line) for libraries of 300 bp fragments. Dashed black line indicates 8.3-fold savings for uSort-M vs traditional synthesis. *Right*: table indicating the estimated cost of individual steps required for traditional gene synthesis or uSort-M for 500- and 1,500-variant libraries, respectively, of 300 bp fragments.

uSort-M enables efficient generation of a scanning mutagenesis library.

To demonstrate the capabilities and potential of uSort-M and directly compare results from Illumina and Nanopore sequencing, we ordered a 328-variant pooled oligonucleotide library and isolated clones via the uSort-M pipeline. The library encoded four different codon substitutions at each of 85 residue positions (excluding wild-type amino acids) within the small human enzyme acylphosphatase-2 (hAcyP2)[10] (**Figure 2A**). hAcyP2's small size (~100 amino acids) allowed us to encode the entire structured region within a ~250-bp oligonucleotide, simplifying synthesis (**Supporting Information, Table S2**). After appending ~350 bp of DNA sequence required for plate/well position barcoding and Illumina sequencing (see below), the entire construct remained ≤650 bp and could be sequenced using a MiSeq (**Supporting Information, Figure S2**).

We amplified the oligonucleotide pool using a universal flanking primer set and then assembled it into a destination vector containing an ampicillin resistance marker via pooled Golden Gate cloning.[11] After transformation, plating, and overnight growth, we obtained 5,250 transformants, representing a 16-fold oversampling of the 328-member library. By Poisson sampling statistics from a 328-member population

with replacement, this level of oversampling represents an expected >99.9% completeness of the library represented in the sampled transformants[1–3] assuming uniform representation of each DNA variant (see **Methods** for calculations).

After culturing the pool of transformed cells overnight, we used a BD FACSAria II to sort individual bacteria into 384-well plates (**Figure 2A**). Gating based on forward and side-scattering signals discriminated bacteria from other debris in the absence of a fluorescent signal; and an additional gate based on the height and area of the forward scattering signals identified single bacterial cells. Anticipating some loss at each subsequent stage of the pipeline, we sorted eight 384-well plates over ~2 hours of instrument time (3,072 wells total; 9.4-fold oversampling and >99.9% expected completeness, **Figure 2A**). After sorting and culturing overnight, 2,057 of 3,072 wells showed growth from a single sorting event (as assessed by quantifying optical density), equivalent to 67% efficiency, 6.3-fold oversampling, and 99.8% expected completeness (**Figure 2B** and **Supporting Information, Figure S3** and **Table S3)**. Because plates are cultured without shaking, colonies are visible at the bottom of many wells with growth (others may appear suspended), making it possible to manually examine or image plates to detect doublets ahead of sequencing (**Supporting Information, Figure S4**).

To enable pooled sequencing of amplicons while preserving information about their well location, we used colony PCR to amplify DNA templates directly from the cultures and append primer barcodes encoding plate and well positions. Illumina's Nextera indices served as plate-specific outer barcodes (**Supporting Information, Table S4**) and evSeq indices served as well-specific inner barcodes according to a standardized mapping scheme[6] (**Figures 2C and S2**, see **Methods** and **Supporting Information**). The Nextera barcoding primers also appended the i5 and i7 adapter sequences required to perform Illumina sequencing on a MiSeq. As Oxford Nanopore sequencing directly ligates its required adapters to arbitrary DNA molecules prior to sequencing, this allowed us to sequence the same exact pool of DNA via both methods and directly compare outcomes. The full cost of barcoding variants was ~$0.25 per well, or ~$100 per 384-well plate (**Supporting Information, Table S1**).

After obtaining sequencing data from both methods, we first examined how individual reads were mapped to plate-well positions using their respective demultiplexing pipelines. As expected, Illumina MiSeq reads were efficiently mapped to individual wells using existing methods designed for this purpose (**Methods**, see **Supporting Information, Table S5** for compiled demultiplexing results). We obtained 5.54 M paired-end reads containing the correct Nextera barcodes using Illumina's bcl2fastq demultiplexing software, which assigned reads to their respective plates; we then used the evSeq demultiplexing software to further assign the inner barcodes to their respective wells within each plate. Demultiplexing via evSeq assigned 3.45 M reads to wells, with the remaining reads failing due to low read quality or the absence of an expected barcode pair. There were two clear distributions for wells with high (>100 reads) vs. low sequencing depth, which corresponded to wells with high vs. low bacterial growth in the input culture (**Figure 2D**). Overall, 98% (2,017/2,057) of wells with high bacterial growth after sorting were assigned >100 reads, confirming robust amplification of DNA templates directly from bacterial culture. A lack of reads from wells without growth additionally confirmed minimal contamination or index hopping during multiplexed sequencing library preparation.

Equivalent demultiplexing results were obtained using the Nanopore pipeline. Multiplexed Nanopore sequencing typically leverages 24-bp barcodes to compensate for a higher expected level of base-calling noise[7,9,12,13], which is longer than either the 12-bp outer Nextera barcode or 7-bp inner evSeq barcode used here. Using these barcodes, Nanopore sequencing reads were readily demultiplexed by Oxford Nanopore's Dorado software[14] (**Methods**), assigning 1.47 M of 4.31 M reads to specific wells (42.6% of the final assigned read depth, compared to MiSeq). Per-well sequencing depths were ~2-fold lower than, but highly correlated with, MiSeq-based results across all 8 plates (**Figure 2E**, r^2 = 0.98, slope = 0.42). Despite higher reported sequencing error rates for Nanopore sequencing, the MiSeq+evSeq and Nanopore+Dorado pipelines had comparable rates of assigning reads to wells (62.2% vs. 64.4%, **Supporting Information, Table S5**) and Nanopore sequencing returned >100 reads for 2,005 of the 2,057 wells with growth (97%).[15] Nanopore reads were no more likely than MiSeq reads to be assigned

to wells lacking obvious growth (**Supporting Information, Figure S5**), suggesting that random miscalling and misassignment of the shorter barcodes was not a problem even with noisier data.

We next aligned well-specific reads to the reference AcyP2 sequence in order to determine the variant contained within each well. For the MiSeq data, we continued to use the established evSeq software to determine the identity of each well; for the Nanopore data, we used minimap2[16] to align reads to the reference sequence and custom software to determine mismatched bases and confirmed final variant calls using the samtools consensus software[17] (**Methods**). Across the 2,005 wells that had >100 Nanopore and MiSeq reads, 1,676 (83.6%) had identical variant calls across both platforms. Of the 329 that did not match, 284 (86%) of these were detected to be multiclonal. Of the remaining 45 wells, all but

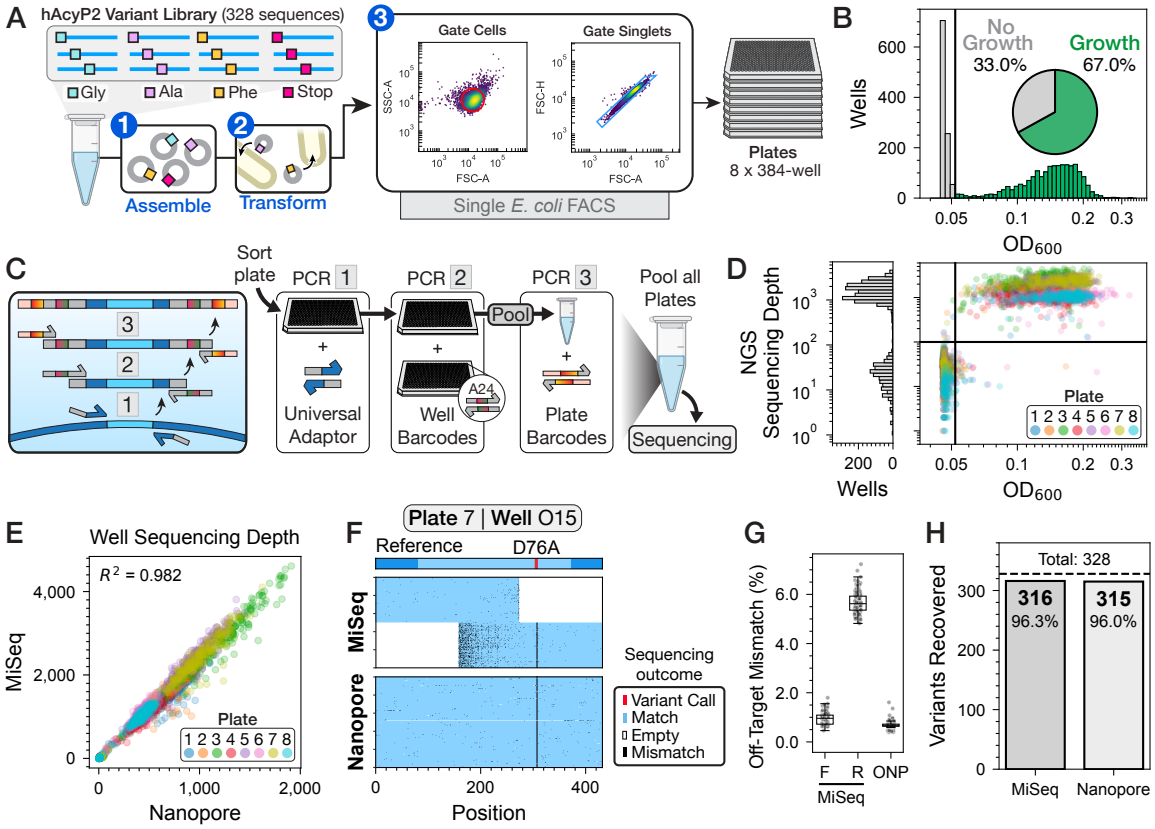


Figure 2. Efficient recovery of a 328-variant scanning library with uSort-M. A. Schematic illustrating 328-variant human acylphosphatase 2 (hAcyP) library processing, including (1) assembly, (2) transformation, and (3) FACS sorting. Left scatter plot: selection of cell events by gating the distribution of side scatter area (SSC-A) *vs.* forward scatter area (FSC-A) values. Right scatter plot: selection of singlets from gating the distribution of forward scatter height (FSC-H) *vs.* FSC-A. **B.** Histogram of the number of wells as a function of measured OD_{600} values. Wells with bacterial growth (green bars representing 2,057 wells) were defined as having OD_{600} > 0.052 (grey vertical line). Inset pie chart indicates fraction of wells with (green, 67%) and without (grey, 33%) growth. **C.** Plate-well barcoding scheme for multiplexed sequencing of variant DNA in each well. Three iterations of PCR encode source plate and well location on each read. **D.** Number of demultiplexed MiSeq reads (sequencing depth) *vs.* well OD_{600}; marker color indicates plate of origin; vertical and horizontal lines are shown for OD_{600} = 0.052 and sequencing depth = 100, respectively. Histogram (left) indicates number of wells with a given read depth. **E.** Scatter plot comparing number of reads per well from demultiplexed MiSeq data *vs.* Nanopore data; annotation specifies Pearson correlation coefficient. **F.** Read alignment plots from a representative well comparing paired-end MiSeq data and long-read Nanopore data. Sequence positions are aligned along the x-axes of each plot, with bases matching the reference are colored in light blue, mismatches in black, and empty positions in white. *Top*: read schematic showing aligned positions corresponding to the hAcyP2 WT reference sequence (light blue), called variant for the given well (red), and DNA appended during indexing (dark blue). *Middle*: Subsampled paired-end MiSeq reads ordered by orientation, with forward reads on top and reverse reads on bottom. *Bottom*: Subsampled Nanopore reads. **G.** Quantified fidelity across all reads from all wells, defined as the frequency of finding a mismatch outside of the called variant bases. Each point shows the fidelity of a single read, boxes denote the median, 25th, and 75th percentiles of the distribution, and error bars represent the standard deviation. **H.** Number of variants recovered for the 328-member library from MiSeq or Nanopore data.

one were determined to have indels. These indels were directly detected from Nanopore sequencing by samtools consensus but were filtered to a low-but-nonzero number (<100) of spurious indel-filtered reads by evSeq processing of Illumina-sequenced samples. The single remaining well that passed all quality-control thresholds but did not match was caused by an instance where the Nanopore pipeline identified a mismatch in the same codon(s) as the MiSeq pipeline but was unable to resolve its specific identity. Overall, Nanopore errors were randomly distributed across reads such that true variation was easy to identify (**Figure 2F**). By contrast, Illumina read errors were similarly distributed in the forward reads, but highly concentrated near the end of the reverse reads (**Figure 2F,G**). Both sequencing methods recovered variants at similar frequencies across positions in the hAcyP2 amplicon, with no evidence of positional bias affecting variant recovery (**Supporting Information, Figure S6**).

Finally, we assessed the number of desired variants recovered by MiSeq and Nanopore sequencing, respectively. For the 2,017 wells with >100 MiSeq reads, 1,375 (68.2%) contained desired single-mutant sequences as their most abundant variant. This suggests that ~32% of wells contained off-target variation (e.g., multiple mutations or indels). For our 300-mer pooled oligos with a synthesis error rate of 1/3000 bases, we expect 10% of oligos to contain errors, suggesting that PCR amplification of the ssDNA pool, Golden Gate assembly, and *E. coli* processes during transformation contribute the remaining ~22% of off-target variation. Of these 1,375 samples, 79 of these had low clonal purity (<90% of reads indicated the same variant; **Supporting Information, Figure S7**), leaving a total of 1,296 in-library monoclonal wells (64.2%, 4-fold oversampling of the 328-member library). Nanopore variant calling results were similar, with 1,293 (64.5%) of 2,005 wells with >100 reads containing high-confidence monoclonal single-mutant sequences. From these sampling numbers, we expected ~98% coverage of the 328-member library. Consistent with this prediction, we identified 316 (96.3% completeness) and 315 (96.0% completeness) unique variants using the MiSeq and Nanopore pipelines, respectively (**Figure 2H**). Overall, Nanopore offered lower cost and higher flexibility while returning similar information, even when using libraries with sequence features optimized for Illumina sequencing.

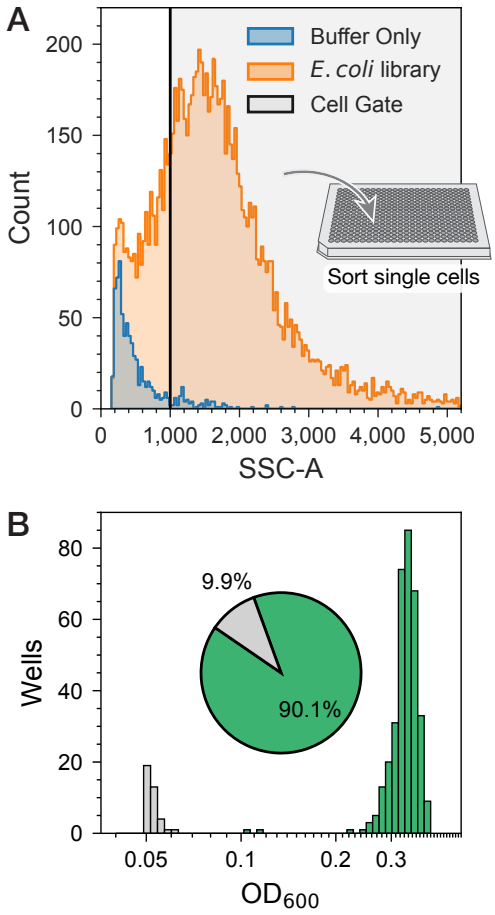


Figure 3. Optimizing recovery of *E. coli* clones. A. Event counts *vs.* side scatter area signal (SSC-A) from FACS analysis of either buffer-only (blue) or transformed *E. coli* (orange) samples. **B.** Measured OD_{600} for 384 wells sorted using a more stringent SSC-A gating threshold. Positive growth is colored in green, while no growth is colored in gray. *Inset*: proportion of wells with growth (green, 90.1%) to no growth (gray, 9.9%).

Highly selective gating optimizes *E. coli* clone recovery during cell sorting.

During hAcyP2 library generation, only 67% of 3,072 sorted wells yielded successful outgrowth, suggesting that cell sorting was a major source of inefficiency. This inefficiency could arise from failed cell deposition or failure of deposited wells to grow.[18] To distinguish between these possibilities, we compared side-scatter area (SSC-A) distributions from a stationary-phase *E. coli* culture with a buffer-only negative control. Both samples contained a debris population with a peak near 500 RFU (**Figure 3A**); this peak persisted even when buffer was filtered through a 0.22-μm filter, suggesting that debris contamination—rather than cell viability—was the primary source of failed deposition events. To test whether more stringent SSC-A gating could reject this debris and improve outgrowth, we sorted cells into a single 384-well plate using a refined gate and assessed growth after overnight incubation. The stringent gating increased the fraction of wells

with growth from 67% to 90.1% (**Figure 3B**). Because the number of sorted cells required (1,000–10,000) will always be far smaller than the total number of cells in the culture (~10^7), stringent gating does not risk excluding rare library members.

Simulations recapitulate experimental results and reveal factors that influence library recovery.

Standard sampling statistics predict library recovery without considering the potential impact of inefficiencies throughout the uSort-M pipeline, potentially explaining why our measured recovery rate (96%) was slightly below the predicted rate (98%). To estimate the number of wells that must be sorted and sequenced for a given library coverage under real-world conditions, we simulated how inefficiencies in DNA synthesis and assembly, bacterial transformation, sorting and outgrowth, and sequencing library preparation impact variant recovery. Each simulation considers five primary user-defined quantities: (1) **library skew**, or the fold-difference in relative abundance between the most- (90th percentile) and least- (10th percentile) represented sequences; (2) **off-target variation**, or the fraction of amplicons that do not match the target library due to errors during DNA synthesis, plasmid assembly, transformation, and

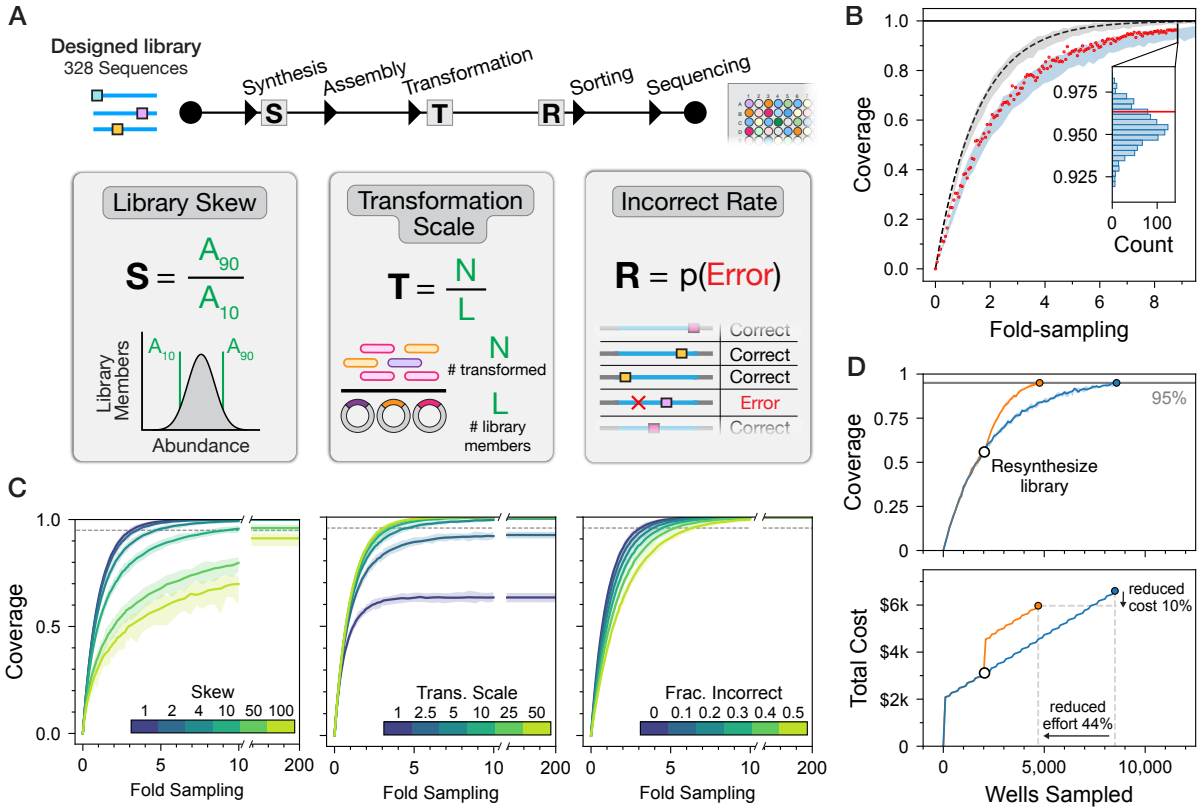


Figure 4. Simulations demonstrating the effects of uSort-M parameters on library coverage. **A.** Simulation pipeline illustrating key parameters from each step in the uSort-M workflow and how they are defined. **B.** Simulated coverage as a function of fold-sampling of a 328-member library, where the number of wells sorted is equal to the product of the library size and fold-sampling. Red points represent bootstrap downsampling from the collected hAcyP2 library sequencing data for each fold-sampling value; blue shaded area shows 99% confidence interval from 100 independent simulations using known parameters of the experiment (skew = 2, off-target variation = 0.35, transformation scale = 5,250/328, sorting efficiency = 0.67, PCR failure rate = 0.025). Grey shaded area shows 99% confidence interval from 100 independent numerical simulations assuming a 'perfect' input library (skew = 1, off-target variation = 0, transformation scale = 100, same sorting and PCR parameters); dashed black line indicates the analytical solution to sampling from this population. Inset shows simulated coverage distribution at 3,072 sorted wells in blue; red line indicates the observed coverage (0.963, 316 variants). **C.** Coverage *vs.* fold-sampling plots for a 1000-member library with varying skew, off-target variation, and transformation scale, holding non-varied parameters at 'realistic' values of 1, 0, and 50, respectively. The x-axis is discontinuous between 10x sampling and >380x sampling to illustrate where the curves plateau. **D.** Sampling *vs.* coverage (*upper*) or *vs.* cost (*lower*) for either exhaustive sampling (blue curve) or sampling with a 'targeted resynthesis' step (orange curve) of a 1,000-member library. In 'targeted resynthesis', a set number of wells are sorted and sequenced to a lower coverage value, and then the remaining unsampled members of the library are reordered as a new, smaller library.

outgrowth, (3) **transformation scale**, or the ratio of the number of transformed cells to the number of unique library variants, (4) **sorting and outgrowth efficiency**, or the fraction of wells that show growth, and (5) **PCR failure rate**, or the fraction of wells with growth that do not yield product after colony PCR (**Figure 4A**).

To test if these parameters are sufficient to accurately explain our observed uSort-M recovery rate, we simulated recovery based on the known parameter values from the initial hAcyP2 scanning library (a library size of 328, a library skew of 2, a total of 5,250 transformants, 35% off-target variation, 67% sorting outgrowth efficiency, and a 2.5% PCR failure rate; **Supporting Information, Figure S8**). These simulations predicted a median recovery of 96% at 3,072 wells, in good agreement with the observed median recovery of 96% (**Figure 4B, inset**). Predicted recoveries also consistently agreed with downsampling simulations that mimicked picking fewer wells from the hAcyP2 experiment (**Figure 4B, red points**), further establishing that the uSort-M process was well-approximated by these simulations.

Next, we systematically varied library skew, transformation scale, and off-target variation for a synthetic 1000-member library to assess how each parameter affects coverage (**Figure 4C** and **Supporting Information, Table S9**). The simulations revealed that differences in library skew have the largest impact on recovery: at 1-, 10-, and 50-fold skew, achieving 90% coverage requires oversampling by 2.4, 5.8, and 27.6-fold respectively for the simulated library. For high (50-fold) skew, even 200-fold oversampling (200,000 wells, in this case) recovers only 95% of variants, as rare variants must compete against substantially more abundant sequences during sampling. Transformation scale sets an upper bound on the coverage that can be achieved, as variants that don't appear in the initial transformed pool can never be recovered. At a transformation scale of 1 (the number of transformants is equal to the library size), only 63% of variants from a perfectly uniform library are expected to be present in the transformed cells (again calculated using Poisson sampling statistics). Increasing transformation scale to 2.5 enables 90% recovery at 7-fold oversampling, while transformation scales of 5 or higher ensure near-complete representation. However, this value can change for highly skewed libraries, as rare variants are less likely to be transformed. These results highlight the importance of transforming a sufficient number of cells during the cloning phase. Off-target variation acts as a constant tax on sampling efficiency by introducing "wasted" samples in which wells contain incorrect sequences. Off-target variation typically stems from errors that occur during DNA synthesis (Twist Biosciences reports a 1:3000 bp error rate) and during PCR amplification and sequencing library preparation (e.g., template switching) and can range from 10–40%.[4] At a 30% off-target rate (near the 35% observed in the hAcyP2 library), achieving 90% coverage requires 4-fold oversampling compared to 2.5-fold with perfect fidelity. The full data and model can be found at https://github.com/FordyceLab/usortm.

Charting the efficient recovery of large and highly skewed libraries with iterative resampling.

For large or highly skewed libraries, sampling redundancy increases substantially as previously isolated variants are re-sampled (**Figure 4C**). As a result, attaining high library coverage via a single round of sampling can require sorting tens of thousands of wells across many dozens of 384-well plates. In these cases, iterative resampling with targeted resynthesis can dramatically increase efficiency. In this scheme, an initial round of modest oversampling (e.g., 2-fold) is applied to efficiently sample unique variants in a regime where the rate of unique variant recovery per sampled well is nearly linear. After this initial stage, the missing variants are then resynthesized as a new oligo pool and used for a second round of sampling. For a 1,000-member library with 2-fold skew and 30% off-target variation (**Figure 4D**), a single-pass approach would require sorting 8,569 wells (~8.5-fold oversampling) to achieve 95% coverage. By contrast, one round of resampling at 2-fold coverage would achieve the same coverage with 4,743 total sorted wells—a 45% reduction. Incorporating a resynthesis step into our Monte Carlo simulation, we found an overall cost of ~$6,000 at 2-fold resampling (compared with $6,600 for a single-pass approach). By simulating resynthesis across various initial sampling depths, we identified approximately 2-fold oversampling as the optimal trigger point for resynthesis of large libraries (**Supporting Information, Figure S9**). This yields considerable savings for large libraries: for a 3,000-

member library, resynthesizing the library at 2-fold oversampling yields more than 20% cost savings and 10,300 fewer total wells sorted compared to a single-pass workflow.

Conclusions

uSort-M provides a scalable and cost-efficient path toward generating large libraries of isolated genes. Pairing widely accessible FACS instruments for clone isolation with commercial long-read nanopore sequencing bypasses the need for specialized robotics and circumvents amplicon length and composition limitations inherent to short-read sequencing-by-synthesis. Evaluating uSort-M on a 328-member scanning mutagenesis library yielded a recovery rate of 96% with both MiSeq- and Nanopore-based sequencing pipelines, demonstrating that long-read sequencing offers comparable accuracy for deep amplicon sequencing at lower cost and with fewer library preparation constraints. uSort-M is fully agnostic to the composition of the input pool: any method that yields transformable plasmid DNA (e.g., degenerate codon synthesis,[1,19] assembly of tiled variable regions or entire genes[5,20,21], pooled site-drected mutagenesis[4], error prone PCR[22]) can be parsed using this workflow.

Sampling simulations revealed that library skew exerts the strongest influence on recovery efficiency, as substantially greater oversampling is required to achieve high coverage of rare variants. Pooled oligonucleotide synthesis from commercial providers (e.g., Twist Biosciences or IDT) is commonly limited to sequences shorter than 350 bp, although recent advances in pooled assembly are extending accessible fragment lengths by engineering orthogonal, pairwise interactions in diverse pools.[5] Libraries comprised of shorter sequences typically have low skew (e.g., < 4-fold guaranteed) and are commercially available from pooled oligo synthesis providers. The library skew for the hAcyP2 library was ~2-fold, which enhanced the efficiency of variant recovery (see **Supplemental Information, Figure S8**). While libraries of larger fragments offered by commercial providers or produced via fragment assembly expand the range of sequences that can be explored, they can have skews of 10-fold or more.[5,23] For either large or highly skewed libraries, iterative resampling with targeted resynthesis reduces total sorting effort by ~50% compared to single-pass, exhaustive oversampling approaches, with 2-fold oversampling as the optimal resynthesis point. This iterative strategy trades increased synthesis costs against reduced sorting and barcoding effort, which may be preferable for applications constrained by instrument time or plate storage capacity. The simulation framework and demultiplexing tools provided at https://github.com/FordyceLab/usortm allow users to predict the required effort and expected coverage for a given library composition before beginning experiments, enabling optimization of the trade-off between sorting effort and resynthesis cost.

Overall, uSort-M greatly reduces the cost and time of parsed library generation by leveraging the accessibility of single-cell *E. coli* sorting and long-read sequencing. We anticipate that this workflow will be broadly useful for generating defined collections of protein variants for downstream biochemical characterization, accelerating efforts to link sequence to function at scale.

Methods

Cost analysis

Costs for traditional gene synthesis were calculated using pricing from three commercial gene fragment products: IDT gBlocks ($0.09/bp), IDT eBlocks ($0.07/bp), and Twist Gene Fragments ($0.07/bp). Costs for the full uSort-M workflow were calculated as the sum of pooled synthesis, cloning, sorting, barcoding, and sequencing. Costs are listed in **Supporting Information, Table S1**. Unless otherwise noted, pooled library inputs were synthesized using Twist Oligo Pools (pricing available at https://www.twistbioscience.com/products/oligopools and in **Supporting Information, Table S2**). Cloning costs were calculated assuming a single-insert, 100-µL Golden Gate Assembly reaction into the user's destination vector, followed by transformation into chemically competent cells. Assembly reactions were projected to consume 50 µL of 2X NEBridge Golden Gate Assembly Master Mix (NEB E1601L), corresponding to $53.60 in reagent costs. Transformation reagent costs constituted $3.13 for 25 µL of NEB 5-alpha Ultracompetent cells (NEB C2987H). Sorting costs were calculated based on instrument

and operator time at a combined core facility rate of $135/hr ($70/hr for instrument time, $65/hr for operator time). At approximately 5 minutes per 384-well plate, sorting cost $11.25 per plate. Sorted cells were collected in 384-well plates (Greiner 781091). Barcoding costs, including primers and PCR reagents, totaled $97.73 per 384-well plate. Sequencing costs were calculated using Plasmidsaurus Custom Sequencing pricing, which starts at $500 for 1 gigabase of sequencing and increases by $50 for each added gigabase (as of 2025; https://www.plasmidsaurus.com/custom). Anticipated sequencing depth was fixed at 100-fold coverage of the amplicon library; this target depth was used to compute the required sequencing volume for each library size.

The minimum cost per variant ($2.58) represents an optimized scenario using short oligonucleotide substitutions (30 bp) assembled into a 1,500 bp wild-type ORF, assuming minimal library skew (≤2-fold), high sorting efficiency (≥90%), and modest oversampling (2–3x). Actual costs vary with library size, fragment length, and complexity (see **Figure 1C** and **Supporting Information, Supplemental Methods** for detailed justification and cost projections).

Scanning library design and synthesis

Scanning libraries were based on a sequence of human acylphosphatase 2 (hAcyP2; UniProt ID: P14621) that was codon-optimized for expression in *E. coli*. Each in-frame codon was mutated to code for Gly (`GGC`), Ala (`GCG`), Phe (`TTT`), or the amber stop codon (`TAG`), excluding synonymous mutations. A single-stranded oligonucleotide pool containing each sequence flanked by BsaI restriction sequences and terminal primer adapter sequences was ordered from Twist Biosciences. The WT hAcyP2 DNA sequence and destination vector sequence are available in the **Supporting Information**.

AcyP library cloning and transformation

The oligonucleotide pool was resuspended to 20 ng/µL in DI water. To convert the library to dsDNA, we used 0.5 µL of the sample to perform PCR using a Roche KAPA HiFi HotStart ReadyMix PCR Kit (Roche KK2602; final DNA concentration 0.4 ng/µL) and subjected to the following thermal cycle: 3 min at 95 °C, then 12 cycles of 20 sec at 98 °C, 15 sec at 62 °C, and 15 sec at 72 °C, then a final step of 72 °C for 60 sec.

Amplicon products were analyzed and isolated by gel electrophoresis and recovery (Zymo D4001). Purified DNA was assembled into a plasmid destination vector (see **Supporting Information**) via a Golden Gate reaction using NEBridge BsaI-HFv2 Golden Gate master mix (NEB E1601S). Assembly reactions were used directly to transform NEB 5-alpha chemically competent *E. coli* (NEB C2987H). 25 µL of thawed cells were incubated on ice with the 3 µL of the assembly reaction for 30 minutes. After a 30-second incubation at 42 °C and a 5-minute incubation on ice, 1 µL of the transformation reaction was removed for transformant scale determination as described below. The remaining volume of cells was cultured in LB media containing 100 µg/mL carbenicillin (LB_{carb}) overnight. The following day, a glycerol stock was prepared by combining 500 µL of culture with 500 µL of filtered 50% glycerol.

Determination of transformation scale

The transformation reaction prepared above was diluted at 1:20 and 1:100 in LB_{carb} and 20 µL was streaked onto solid media of LB-agar supplemented with 100 µg/mL carbenicillin. For each dilution the colony forming units (CFUs) were determined and used to estimate the number of transformants in the 20 µL used to inoculate the overnight culture, yielding 5,250 cells and a 16-fold oversampling of the 328-member library.

Initial estimation of library coverage from sampling statistics

The expected fractional coverage of a library (*F*) in each sample is given by

$$F = C/V = 1 - e^{-L/V},$$

where C = unique variants sampled, V = total number of unique variants in the library, and L = total number of samples taken (e.g., number of transformants or number of sorted wells).[2,3] The level of "oversampling" (e.g., transformation scale) is represented by the ratio of L/V in the exponent. This equation, derived from the Poisson distribution which approximates this sampling process, assumes sampling with replacement and an unbiased distribution of variant abundance in the population.

Optimized single-cell *E. coli* FACS protocol

Twelve hours before sorting, the *E. coli* pool was used to inoculate a culture containing LB_{carb} and grown with shaking at 37 °C. A 100-µL volume of the saturated cell culture was diluted into 900 µL of 1X phosphate-buffered saline (PBS, Thermo 10010023). Cells were washed twice to remove residual media by pelleting via centrifugation at 10,000×*g*, removing the supernatant, and gently resuspending the cells in 1 mL of new PBS. Sorting was performed on either a Sony SH800S or BD Aria FACS instrument. Prior to sorting the cell sample, a sample of the PBS used for dilution was analyzed to observe the distribution of scattering signal for non-cell events. This was used to gate against the cell sample to reduce the frequency of false positives (see **Figure 3**). Sorted cells were isolated within the wells of transparent-bottom 384-well plates (Greiner 781091) that were pre-filled with 40 µL of LB_{carb}. (*Note*: The transparent plate bottom is critical for optical density (OD) measurements in the steps to follow.) Sorted plates were subjected to centrifugation at 4,000×*g* and then tightly covered with adhesive foil seals (Thermo AB0626) and incubated at 37 °C for 24 hours without shaking.

After 24 hours, single colonies were present at the bottom of wells. (*Note*: at this stage images can be obtained to identify doublets.) Colonies were resuspended via pipetting using a 96-channel Liquidator96 (Rainin 17010335) and 1 µL was used in the subsequent barcoding PCRs (below). Glycerol stock plates were prepared simultaneously by thoroughly mixing 10 µL of the culture with 10 µL of 50% (v/v) glycerol in a second 384-well plate and stored at -80 °C. The OD_{600} of each well was measured by a plate reader (Tecan) and used to determine sorting efficiency.

evSeq and Nextera barcoding PCRs

Well-based barcoding of the sorted AcyP scanning library followed the evSeq protocol to create the initial, well-specific inner barcoded amplicon. Briefly, 1 µL of resuspended cell culture from sorted plates was added to 7 µL of evSeq PCR master mix containing site-specific inner primers, Taq polymerase, DMSO, dNTPs, and ThermoPol buffer (**Supporting Information, Table S6** and ref. [6]). An initial PCR was performed by heating to 95 °C for 5 min, then 10 cycles of 95 °C for 20 sec, a touchdown annealing step starting at 63 °C and decreasing 1 °C every cycle (ending at 53 °C) for 20 sec, and a 68 °C extension step.

Following this, well-specific barcodes were added according to the following scheme, which is graphically outlined in **Supporting Information, Figure S10**: for odd-numbered plates (1, 3, 5, 7), dual-index barcoding plates DI01–DI04 (96 unique barcode pairs each) were used to barcode each well in the plate. The 384-well plates were treated as four unique interleaved 96-well plates, with each one mapping to alternating wells originating from A1, A2, B1, or B2. (For example, A1 wells are the union of columns 1, 3, …, 24 and rows A, C, E, …, P.) DI01 was added to wells associated with A1, DI02 with A2, DI03 with B1, and DI04 with B2. For even-numbered plates (2, 4, 6, 8), DI05–DI08 were used and similarly associated with A1, A2, B1, and B2, respectively. After addition of 2 µL of the DI plate barcodes, the PCR was continued for 25 cycles of 95 °C for 20 sec, annealing 53 °C for 20 sec, and extension at 68 °C, followed by a final 5-min extension at 68 °C.

At the completion of the PCR, wells for each unique DI plate were combined into a single mixture containing a final concentration of 20 mM EDTA, pH 8 to inhibit further polymerase activity. (This is performed efficiently by using a multichannel pipette to transfer 5 µL from each well into a strip of 12 PCR tubes pre-filled with 10 µL of 100 mM EDTA, pH 8. Then the mixture from each PCR tube is combined into a single well of a deep-well 96-well plate.) Reactions originating from the same origin 384-well plates were then further combined into microcentrifuge tubes. These eight mixtures were then analyzed and purified by gel electrophoresis using 2% agarose.

Purified DNA was then subjected to a short secondary PCR that appends the plate-specific, outer Nextera barcodes onto each amplicon (**Supporting Information, Table S4**). Addition of these barcodes fully differentiates plates that otherwise share the same set of well-specific barcode sets (e.g., odd and even-numbered plates) into fully unique DNA samples that can be mixed and sequenced together.

Pooled DNA from each plate was amplified with KAPA HiFi HotStart polymerase at a scale of 1 µL DNA at a concentration of 1 ng/µL, 12.5 µL of 2X KAPA HiFi HotStart ReadyMix (Roche KK2602), 10 µL of H_2O, 0.75 µL of 10 µM forward primer, and 0.75 µL of 10 µM reverse primer (see **DNA Sequences** in **Supporting Information**). Thermal cycling conditions were as follows: 95 °C, 5 min; 10 × [98 °C, 20 s; 60 °C, 15 s; 72 °C, 2 min]; 72 °C, 2 min.

Illumina MiSeq sequencing and analysis

Library preparation. Pooled amplicons were purified using Ampure XP beads (Beckman Coulter A63881) at a 1.8x ratio according to the manufacturer's protocol. Purified library concentration was quantified using a Qubit HS Assay Kit (Thermo Fisher Q32854) and fragment size was verified by eGel. The final pooled library concentration was 12.25 ng/*µ*L with an average fragment size of 590 bp.

Sequencing. Libraries were diluted to 4 nM and denatured with 0.2 N NaOH for 5 minutes at room temperature. Denatured libraries were diluted to 9 pM in Illumina Hybridization Buffer and pooled with 30% PhiX Sequencing Control v3 (Illumina FC-110-3001). A 600 µL aliquot of the library was loaded onto a MiSeq Reagent Kit v3 (600-cycle) cartridge (Illumina MS-102-3003) and sequenced with 300 cycles on R1 and 300 cycles on R2.

Basecalling and demultiplexing. Basecalling was performed using bcl2fastq (2.20) on the Stanford Sherlock HPC cluster. Individual 384-well plates were demultiplexed by their 12-bp Nextera indices (see **Supporting Information Table S4** for barcode sequences), then further demultiplexed to well positions using the evSeq software[6] with default parameters and flags `--keep_parsed_fastqs` and `--return_alignments` (See **Supporting Information** for exact commands, **Tables S6–7** for example refseq files). Total read depth and composition for each well was determined with the well-specific FASTQ files generated from the previous step.

Nanopore Long-read sequencing and analysis

Sample submission. From the Illumina library preparation, 40 µL of pooled amplicons at 6 ng/µL were submitted to Plasmidsaurus for Custom DNA Sequencing with a target yield of 3 Gb (https://www.plasmidsaurus.com/custom).

Basecalling and demultiplexing. Basecalling was performed by Plasmidsaurus using Dorado (4.3.0), and FASTQ files were downloaded for downstream analysis. Reads were demultiplexed using Oxford Nanopore's Dorado demux tool (1.1.1; see **Supporting Information** for parameters). The same 12-bp Nextera and 7-bp evSeq barcode sequences used for Illumina demultiplexing were supplied as a custom barcode arrangement file (see **Supporting Information** for configuration details).

Variant calling. Demultiplexed reads for each well were aligned to the hAcyP2 reference sequence using minimap2 (2.28) with parameters `-ax map-ont`. Consensus sequences were generated using samtools consensus (1.21) with parameters `-f fasta`. Variants were called by comparing each consensus sequence to the reference; wells were assigned to library members when ≥90% of reads supported a single variant call. Total read depth and composition for each well were determined from the well-specific FASTQ files.

Hit picking

Clone selection. A map linking each desired variant to its source plate and well position was compiled from the sequencing analysis. This map was reformatted as an input file for an Integra ASSIST PLUS pipetting robot (see **Supporting Information** for file format).

Liquid handling. Glycerol stock plates were thawed on ice for 30 minutes. Using the ASSIST PLUS, 5 µL of each identified clone was transferred from the source glycerol stock into 100 µL of LB in a destination 96-well plate. Destination plates were incubated overnight at 37 °C with shaking at 250 rpm. Following overnight growth, cultures were mixed 1:1 with 50% filtered glycerol to create 25% glycerol stocks for long-term storage at −80 °C.

For downstream applications, glycerol stocks were used to inoculate expression cultures or subjected to plasmid miniprep for sequence verification.

Acknowledgements

We thank Forrest L. Zepezauer for assistance with initial cloning experiments, and Michael T. Montgomery and Benjamin R. Doughty for preparing sequencing libraries and performing Illumina sequencing. We thank the Stanford Shared FACS Facility for instrument access and technical support. This work was supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under grant number R01GM064798 awarded to P.M.F. Individual support comes from the NIGMS of the NIH (F32GM156066 to P.J.A.) and the National Human Genome Research Institute (NHGRI) of the NIH (F31HG013267 to M.B.O.). P.M.F is a Chan Zuckerberg Biohub San Francisco Investigator. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author Contributions

M.B.O., P.J.A., and P.M.F conceived the study. M.B.O. and P.J.A. developed the methodology, performed experiments, analyzed data, developed software, and wrote the manuscript with input from all authors. P.J.A. and M.B.O. developed simulations and modeling. J.D.S. performed initial feasibility testing for single-cell sorting of *E. coli*. P.M.F. supervised the research, provided funding, and edited the manuscript.

References

(1) Kille, S.; Acevedo-Rocha, C. G.; Parra, L. P.; Zhang, Z.-G.; Opperman, D. J.; Reetz, M. T.; Acevedo, J. P. Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis. *ACS Synth. Biol.* **2013**, *2* (2), 83–92. https://doi.org/10.1021/sb300037w.

(2) Patrick, W. M.; Firth, A. E.; Blackburn, J. M. User-friendly Algorithms for Estimating Completeness and Diversity in Randomized Protein-encoding Libraries Wayne M. Patrick and Andrew E. Firth Contributed Equally to This Work. *Protein Eng.* **2003**, *16* (6), 451–457. https://doi.org/10.1093/protein/gzg057.

(3) Bosley, A. D.; Ostermeier, M. Mathematical Expressions Useful in the Construction, Description and Evaluation of Protein Libraries. *Biomol. Eng.* **2005**, *22* (1), 57–61. https://doi.org/10.1016/j.bioeng.2004.11.002.

(4) Appel, M. J.; Longwell, S. A.; Morri, M.; Neff, N.; Herschlag, D.; Fordyce, P. M. uPIC–M: Efficient and Scalable Preparation of Clonal Single Mutant Libraries for High-Throughput Protein Biochemistry. *ACS Omega* **2021**, *6* (45), 30542–30554. https://doi.org/10.1021/acsomega.1c04180.

(5) Freschlin, C. R.; Yang, K. K.; Romero, P. A. Scalable and Cost-Efficient Custom Gene Library Assembly from Oligopools. bioRxiv April 15, 2025, p 2025.03.22.644747. https://doi.org/10.1101/2025.03.22.644747.

(6) Wittmann, B. J.; Johnston, K. E.; Almhjell, P. J.; Arnold, F. H. evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library. *ACS Synth. Biol.* **2022**, *11* (3), 1313–1324. https://doi.org/10.1021/acssynbio.1c00592.

(7) Long, Y.; Mora, A.; Li, F.-Z.; Gürsoy, E.; Johnston, K. E.; Arnold, F. H. LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution and Machine Learning. *ACS Synth. Biol.* **2025**, *14* (1), 230–238. https://doi.org/10.1021/acssynbio.4c00625.
(8) Jain, M.; Koren, S.; Miga, K. H.; Quick, J.; Rand, A. C.; Sasani, T. A.; Tyson, J. R.; Beggs, A. D.; Dilthey, A. T.; Fiddes, I. T.; Malla, S.; Marriott, H.; Nieto, T.; O'Grady, J.; Olsen, H. E.; Pedersen, B. S.; Rhie, A.; Richardson, H.; Quinlan, A. R.; Snutch, T. P.; Tee, L.; Paten, B.; Phillippy, A. M.; Simpson, J. T.; Loman, N. J.; Loose, M. Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. *Nat. Biotechnol.* **2018**, *36* (4), 338–345. https://doi.org/10.1038/nbt.4060.
(9) Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K. F. Nanopore Sequencing Technology, Bioinformatics and Applications. *Nat. Biotechnol.* **2021**, *39* (11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.
(10) Stefani, M.; Taddei, N.; Ramponi, G. Insights into Acylphosphatase Structure and Catalytic Mechanism. *Cell. Mol. Life Sci. CMLS* **1997**, *53* (2), 141–151. https://doi.org/10.1007/PL00000585.
(11) Engler, C.; Marillonnet, S. Golden Gate Cloning. In *DNA Cloning and Assembly Methods*; Valla, S., Lale, R., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2014; pp 119–131.
(12) Byrne, A.; Beaudin, A. E.; Olsen, H. E.; Jain, M.; Cole, C.; Palmer, T.; DuBois, R. M.; Forsberg, E. C.; Akeson, M.; Vollmers, C. Nanopore Long-Read RNAseq Reveals Widespread Transcriptional Variation among the Surface Receptors of Individual B Cells. *Nat. Commun.* **2017**, *8* (1), 16027. https://doi.org/10.1038/ncomms16027.
(13) Ni, Y.; Liu, X.; Simeneh, Z. M.; Yang, M.; Li, R. Benchmarking of Nanopore R10.4 and R9.4.1 Flow Cells in Single-Cell Whole-Genome Amplification and Whole-Genome Shotgun Sequencing. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 2352–2364. https://doi.org/10.1016/j.csbj.2023.03.038.
(14) *GitHub - nanoporetech/dorado: Oxford Nanopore's Basecaller.* https://github.com/nanoporetech/dorado/.
(15) Hu, T.; Chitnis, N.; Monos, D.; Dinh, A. Next-Generation Sequencing Technologies: An Overview. *Hum. Immunol.* **2021**, *82* (11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012.
(16) Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **2018**, *34* (18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
(17) Danecek, P.; Bonfield, J. K.; Liddle, J.; Marshall, J.; Ohan, V.; Pollard, M. O.; Whitwham, A.; Keane, T.; McCarthy, S. A.; Davies, R. M.; Li, H. Twelve Years of SAMtools and BCFtools. *GigaScience* **2021**, *10* (2), giab008. https://doi.org/10.1093/gigascience/giab008.
(18) Ferrari, B. C.; Oregaard, G.; Sørensen, S. J. Recovery of GFP-Labeled Bacteria for Culturing and Molecular Analysis after Cell Sorting Using a Benchtop Flow Cytometer. *Microb. Ecol.* **2004**, *48* (2), 239–245. https://doi.org/10.1007/s00248-003-1069-9.
(19) Tang, L.; Gao, H.; Zhu, X.; Wang, X.; Zhou, M.; Jiang, R. Construction of "Small-Intelligent" Focused Mutagenesis Libraries Using Well-Designed Combinatorial Degenerate Primers. *BioTechniques* **2012**, *52* (3), 149–158. https://doi.org/10.2144/000113820.
(20) Casini, A.; MacDonald, J. T.; Jonghe, J. D.; Christodoulou, G.; Freemont, P. S.; Baldwin, G. S.; Ellis, T. One-Pot DNA Construction for Synthetic Biology: The Modular Overlap-Directed Assembly with Linkers (MODAL) Strategy. *Nucleic Acids Res.* **2014**, *42* (1), e7. https://doi.org/10.1093/nar/gkt915.
(21) Plesa, C.; Sidore, A. M.; Lubock, N. B.; Zhang, D.; Kosuri, S. Multiplexed Gene Synthesis in Emulsions for Exploring Protein Functional Landscapes. *Science* **2018**, *359* (6373), 343–347. https://doi.org/10.1126/science.aao5167.
(22) Reetz, M. T.; Carballeira, J. D. Iterative Saturation Mutagenesis (ISM) for Rapid Directed Evolution of Functional Enzymes. *Nat. Protoc.* **2007**, *2* (4), 891–903. https://doi.org/10.1038/nprot.2007.72.
(23) Instance Bio. Instance DNA Pool Specifications. https://instance.bio/ (accessed 2026-01-09).