# SUNi mutagenesis: scalable and uniform nicking for efficient generation of variant libraries

Taylor L. Mighell[1*], Ignasi Toledano[1], Ben Lehner[1, 2, 3, 4*]

[1] Center for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.

[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain.

[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

[4] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

*Correspondence: taylor.mighell@crg.eu, bl11@sanger.ac.uk

## Abstract

Multiplexed assays of variant effects (MAVEs) have made possible the functional assessment of all possible mutations to genes and regulatory sequences. A core pillar of the approach is generation of variant libraries, but current methods are either difficult to scale or not uniform enough to enable MAVEs at the scale of gene families or beyond. We present an improved method called Scalable and Uniform Nicking (SUNi) mutagenesis that combines massive scalability with high uniformity to enable cost-effective MAVEs of gene families and eventually genomes.

## Background

Massive mutagenesis followed by functional assays, commonly known as MAVEs or deep mutational scanning, is a powerful strategy for understanding the effects of genetic variation[1–3], dissecting and engineering proteins[4–6], and directed evolution[7]. Modern approaches for generating mutagenesis libraries generally fall into two categories. First, synthesis of oligonucleotides containing programmed mutations followed by subcloning, known as cassette or tile mutagenesis[2,8,9]. Second, synthesis of polymerase chain reaction (PCR) primers containing programmed mutations which bind template DNA and extend to form a mutated strand, followed by various means of degrading and resynthesizing the opposite strand to form mutated double-stranded DNA[10–12].

While cassette mutagenesis yields highly uniform libraries, current DNA synthesis technologies can only generate oligonucleotides up to length ~300, with synthesis quality decaying rapidly with increased length[13]. Since many genes exceed this length, it is necessary to generate sub-libraries, which require complex experimental designs that severely limit scalability. The key advantage of primer-based mutagenesis is that it does not have this limitation; in theory, any number of genes of any length can be mutated in a single pot. However, primer-based methods suffer from differences in mutagenesis efficiency between positions, resulting in libraries with highly nonuniform representation of variants[10–12]. Additionally, primer-based methods can generate substantial amounts of wild-type carryover, requiring the use of larger experimental volumes, increased sequencing, and sequencing errors artificially inflating counts for variants[14]. These drawbacks are problematic because they reduce data quality and increase the cost of every step of a MAVE experiment, thereby limiting scalability.

The Atlas of Variant Effects (AVE) Alliance has the goal to quantify the impact of variation in most human genes and regulatory elements using diverse selection assays[15]. With current rates of progress this endeavor is likely to take decades to

achieve[16]. Here we detail a protocol that we term Scalable and Uniform Nicking (SUNi) mutagenesis that represents a two-fold improvement over the existing state of the art method[12] for large variant library construction. SUNi mutagenesis yields highly uniform variant libraries with massive potential scalability.

## Results and Discussion

Nicking mutagenesis generates mutated plasmid in four steps: degradation of one DNA strand; annealing and extension of a mutagenic primer; degradation of the opposite strand; and resynthesis of the opposite strand, incorporating the mutation[12] (Supplementary Fig. 1). Previous data indicated that longer homology arms could improve mutagenesis efficiency[17], and that the melting temperature ($T_m$) of the mutagenic primer was correlated with mutagenesis efficiency[18]. We reasoned that since binding of both homology arms to the template is required for efficient mutagenesis, performance could be improved by optimizing the $T_m$ of both arms of the primer independently. Therefore, we designed a pool of primers (referred to as opt1) where, for each position, the left and right homology arm had the length between 20-40 nucleotides that had the predicted $T_m$ closest to 61°. These primers were designed to target two 40-codon regions of the μ opioid receptor (MOR) which were chosen because of very high or low GC content (MOR2 =65.8% GC, MOR6 =40.8% GC) and so were expected to provide the greatest challenge for the new design. Advances in DNA synthesis have made oligonuceotide pools an affordable, and therefore scalable, option for synthesizing large numbers of sequences. A previous version of nicking mutagenesis synthesized primers as microarray-based oligonucleotide pools, but the quality of these libraries was substantially lower than the original method[18], possibly due to the femtomole-scale yield of microarray synthesis. To maintain the scalability advantage of oligo pools while still maximizing library quality, we synthesized our primers as IDT oPools, which have picomole-scale yield.

The sequential degradation of each DNA strand of a plasmid is accomplished with the nicking activity (cleavage of only one strand of double stranded DNA) of engineered variants of the BbvCI restriction enzyme. We found that some plasmids containing only one BbvCI site are inefficiently digested in the first nicking step, potentially leading to wild-type carryover. Adding a second BbvCI site to the plasmid improved digestion efficiency (Supp. Fig 2). Therefore, we engineered a plasmid bearing MOR to contain two BbvCI sites, and followed the published nicking protocol with minor modifications (Supplementary Protocol 1). Sequencing of the mutagenesis libraries revealed similar proportions of programmed mutations (63.8 and 58.8% for opt1 versus 65.3 and 64.2% for standard nicking) and slightly increased wild-type percent (26.9 and 33.2% for opt1 versus 23.8 and 23.3% for standard nicking) but with improved uniformity (log difference (LogDiff) between $90^{th}$ and $10^{th}$ percentile of mutants of 0.83 and 0.92 for opt1 libraries versus 1.18 and 0.94 for standard nicking libraries, Fig. 1a,b). While overall uniformity was improved, there was still substantial positional bias (Fig. 1b), which we next sought to understand. However, we found no relationship between mutagenesis frequency (median frequency of all programmed mutations per position) and predicted $T_m$ of left or right mutagenesis primer homology arm, or for minimum, maximum, sum, or difference between left and right $T_m$. We also found no contribution of predicted free energy of secondary structure formation of primers (Supplementary Table 4).

Surprisingly, we did find a significant contribution of GC content of the five 5' terminal bases of the primer. The strongest signal comes when considering GC content of the three 5' terminal bases (Spearman $\rho$=0.56, $p=6.8\times10^{-8}$ Fig. 1c,d). A GC-rich 3' terminus of a primer (also known as "GC clamp") is widely thought to improve priming efficiency, but here we find no contribution of 3' GC clamp (Supplementary Table 4). We divided primers based on the 5' terminus sequence and found that primers with SSS, SWS, or SSW sequence (from 5' to 3', where S =G or C and W =A or T) have the highest median mutagenesis efficiency (Fig. 1e). Conceptually, the importance of a 5'

GC clamp makes sense because the extension step of the mutagenesis PCR is long and at a relatively high temperature (7 minutes at 72°), and if the mutagenic primer terminus is dissociated from the template when the polymerase completes the mutagenic strand, it may polymerize extra bases and make ligation of the mutagenic strand impossible.

We designed a new set of nicking primers (referred to as SUNi), targeting the same regions, and taking advantage of the 5' GC clamp discovery. Briefly, for each position we sought to find a primer that had optimal predicted $T_m$ and also a strong 5' GC clamp (full description in Methods). Further, we reasoned that one contribution to wild-type carryover is NNK primers in which the wild-type codon is encoded by NNK. Since K encodes G and T, for any codon that ends in these bases the wild-type sequence will be present in the NNK pool, and this fully complementary wild-type primer would be expected to outcompete mutation-bearing primers. To minimize this, we used NNK to mutagenize codons that end in A, C, or G, and NNS (S= G or C) if the wild-type codon ended in T. Sequencing of MOR2 and MOR6 SUNi mutagenesis libraries demonstrated increased percentage of programmed mutants (77.5 and 68.9%, respectively) decreased percentage of wild-type (13.9 and 23.7%, respectively), and improved uniformity (LogDiff = 0.65, 0.92, respectively, Fig. 2a).

We wanted to compare methods using a more comprehensive metric, so we calculated $\text{screening efficiency} = \frac{\%\ \text{programmed}}{10^{\text{LogDiff}}}$, a term which incorporates the fraction of programmed sequences in the library and the uniformity of those sequences, which are both important to determine the efficiency of screening the library. Screening efficiency for both libraries increases from opt1 to SUNi designs, and on average SUNi is twice as efficient as the standard nicking protocol (0.128 versus 0.058, respectively, Fig. 2b,c). We also compared a mutagenesis library made by cassette mutagenesis (b2AR2, 250 nucleotide oligonucleotides introducing mutations at 70 positions). We find that in the best case (MOR2), SUNi screening efficiency approaches that of

cassette mutagenesis (0.173 versus 0.200, respectively, Fig. 2b,c), while requiring substantially less hands-on time and allowing mutagenesis of much larger and many different targets in a single reaction pool. Cassette mutagenesis yields highly uniform libraries, but the percent of programmed mutants is low (Fig. 2c) due to errors in DNA synthesis.

We chose to mutagenize regions with high and low GC content, assuming these would be difficult templates for mutagenesis. However, we didn't anticipate the crucial importance of the 5' GC clamp. The data suggests that the mutagenesis efficiency of SUNi is likely related to GC content, indicating that MOR6 is likely difficult while MOR2 is likely an amenable template. We expect SUNi mutagenesis efficiency for regions with intermediate GC content to be intermediate between the examples shown here.

SUNi mutagenesis has the potential to be massively scaled, as there is no theoretical limit to the length of mutated region or the number of mutated regions in a single reaction. The efficiency of screening a SUNi library is twice that of the standard nicking protocol, meaning that at all steps (library generation, screening, and sequencing), the reagents required, and therefore cost, will be halved. We expect SUNi mutagenesis coupled with a panel of selection assays[16] will allow the rapid and cost-effective generation of variant effect atlases for entire gene families. The bright future of MAVEs is reliant on scalable methods for generating high quality variant libraries, and SUNi mutagenesis represents an important step in that direction.

## Conclusions

More efficient libraries empower more scalable experiments that will be necessary for generating atlases of variant effect at the gene-family or genome scale. In this report, we outline design and experimental improvements that improve the screening efficiency of nicking mutagenesis two-fold.

## Materials & Methods

### Opt1 primer design

Primers were designed to introduce all single amino acid mutations and stop codon (via "NNK" codon mutagenesis) for 80 codons in the μ opioid receptor (MOR). To pick a guide for each position, for each homology arm, we found the candidate between 20 and 40 nucleotides with $T_m$ closest to 61° (calculated with biopython[19] using the Bio.SeqUtils.MeltingTemp.Tm_NN function). The two pools of opt1 primers were ordered as IDT oPools. Sequences reported in Supplementary Table 2.

### SUNi primer design

Like opt1, we designed primers to introduce single amino acid mutations at 80 positions of MOR. For each position, we found the right homology arm in the same way as for Library 1, i.e. the arm between 20 and 40 nucleotides that had predicted $T_m$ closest to 61°. For the left homology arm, we enumerated all arms that had predicted $T_m$ between 59° and 66°. If one or more of these arms had all three 5' terminal nucleotides as S (degenerate codon notation; S=G or C, W=A or T), the shortest of these was chosen. If there were no SSS 5' termini, then we looked for arms with SSW or SWS termini, and if there were one or more, we chose the shortest arm. If there were no suitable homology arms with SSW or SWS termini, we then found the arm closest to 64° irrespective of 5' terminus. Since we would then predict this primer to be suboptimal, we encoded it twice in the oPool. In this library we used NNK as the degenerate mutagenic codon if the WT codon ended in A, C, or G, but we used NNS if the WT codon ended in T. The two pools of SUNi primers were ordered as IDT oPools. Sequences reported in Supplementary Table 2.

### b2AR2 mutagenesis

Oligonucleotides were designed to introduce all possible single amino acid changes, and many double amino acid changes, for a total of 4005 variants. These were synthesized by Twist Bioscience as 250 nucleotide oligos. PCR with primers dialout_tile2_[F/R] (primers used in this study reported in Supplementary Table 1) was

done to amplify these mutagenic oligos. PCR with primers designed to amplify the rest of the vector besides the region to be mutagenized (b2AR_satmut_tile2_[F/R]) was performed to prepare the vector, and then Gibson assembly was used to introduce the mutagenic oligos.

**Sequencing library preparation**

Two stage PCR was performed to amplify each mutated region and append indexed Illumina sequencing adapters. Q5 High Fidelity polymerase (New England Biolabs) was used for all PCRs. For MOR2 and MOR6 regions, primers MOR_nicking_T[2/6]_seq_[F/R] were used in stage1 PCR to amplify the target and append partial Illumina sequencing adapters, with 50 ng of purified plasmid as template. Cycling protocol was 98° for 30s, followed by 17 cycles of [98° for 20s, 55° for 30s, 72° for 30s]. Products were column purified and 0.2% of PCR1 was used as input for PCR2 with primers indexed_i[5/7] and cycled with 98° for 30s, followed by 5 cycles of [98° for 15s, 64° for 30s, 72° for 30s]. Products were column purified and sequenced on Illumina Nextseq 500 or Nextseq 2000 instruments. For b2AR2, 10 ng of purified plasmid was used as input to PCR using primers b2AR_Tile2_PCR1_5N_[F/R] and cycling with 98° for 30s, followed by 12 cycles of [98° for 15s, 66° for 30s, 72° for 30s]. Products were column cleaned and 0.2% of PCR1 was used as input for PCR2 with primers indexed_i[5/7] and cycled with 98° for 30s, followed by 10 cycles of [98° for 15s, 64° for 30s, 72° for 30s]. Products were column cleaned and sequenced on Illumina MiSeq instrument.

**Sequencing data processing**

We obtained raw fastq data from the original nicking paper[12] from the Short Read Archive with accession numbers SRR4105481 and SRR4105482. All fastq data were processed identically: first, read pairs were merged and filtered for reads which contained <0.5 expected errors using vsearch[20]. Then, cutadapt[21] was used to trim adapters and only those reads with matching adapters were retained. Variant counts were enumerated by comparing sequencing reads to expected sequences based on

mutagenesis strategy (i.e. NNN, NNK, or NNS) and counting only perfect matches. Read processing data in Supplementary Table 3.

**Availability of data and materials**

Code to generate SUNi mutagenesis primers is available at https://github.com/lehner-lab/SUNi_mutagenesis. Raw sequencing data produced for this study can be found at the Sequence Read Archive with accession number PRJNA939024.

**Figure 1. Optimization and analysis of nicking mutagenesis primer design**

**a,** Per position mutation frequency presented as fraction of all sequencing reads for standard nicking. Dashed lines indicate 90$^{th}$ and 10$^{th}$ percentile of all mutation frequencies.

**b,** Per position mutation frequency presented as fraction of all sequencing reads for opt1 nicking. Dashed lines indicate 90$^{th}$ and 10$^{th}$ percentile of all mutation frequencies.

**c,** Spearman correlation between GC content of the 5' terminus and mutagenesis efficiency, when considering between one and five terminal bases.

**d,** Mutagenesis frequency of positions with different GC content in the 5' terminal three bases. Spearman $\rho$=0.56, p=6.8x10$^{-8}$.

**e,** Mutagenesis frequency of positions with different SW sequences (S=G or C, W=A or T) in the 5' terminal three bases.

**Figure 2. Performance and comparison of SUNi mutagenesis with other methods**

**a,** Per position mutation frequency presented as fraction of all sequencing reads for SUNi mutagenesis. Dashed lines indicate 90$^{th}$ and 10$^{th}$ percentile of all mutation frequencies.

**b,** Screening efficiency of different mutagenesis methods.

**c,** Screening efficiency of different mutagenesis methods, as a function of uniformity and percent programmed. Colors the same as in **b**.

## References


1. Mighell TL, Evans-Dutson S, O'Roak BJ. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. Am J Hum Genet. 2018;102:943–55.

2. Weile J, Sun S, Cote AG, Knapp J, Verby M, Mellor JC, et al. A framework for exhaustively mapping functional missense variants. Mol Syst Biol. 2017;13:957.

3. Amorosi CJ, Chiasson MA, McDonald MG, Wong LH, Sitko KA, Boyle G, et al. Massively parallel characterization of CYP2C9 variant enzyme activity and abundance. Am J Hum Genet. 2021;108:1735–51.

4. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. Nature. 2022;604:175–83.

5. Penn WD, McKee AG, Kuntz CP, Woods H, Nash V, Gruenhagen TC, et al. Probing biophysical sequence constraints within the transmembrane domains of rhodopsin by deep mutational scanning. Sci Adv. American Association for the Advancement of Science; 2020;6:eaay7505.

6. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, et al. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol. 2012;30:543–8.

7. Acevedo-Rocha CG, Ferla M, Reetz MT. Directed Evolution of Proteins Based on Mutational Scanning. Methods Mol Biol Clifton NJ. 2018;1685:87–128.

8. Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res. 2014;42:e112.

9. Macdonald CB, Nedrud D, Grimes PR, Trinidad D, Fraser JS, Coyote-Maestas W. DIMPLE: deep insertion, deletion, and missense mutation libraries for exploring protein variation in evolution, disease, and biology. Genome Biol. 2023;24:36.

10. Kitzman JO, Starita LM, Lo RS, Fields S, Shendure J. Massively parallel single-amino-acid mutagenesis. Nat Methods. Nature Publishing Group; 2015;12:203–6.

11. Firnberg E, Ostermeier M. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. PLOS ONE. Public Library of Science; 2012;7:e52031.

12. Wrenbeck EE, Klesmith JR, Stapleton JA, Adeniran A, Tyo KEJ, Whitehead TA. Plasmid-based one-pot saturation mutagenesis. Nat Methods. Nature Publishing Group; 2016;13:928–30.

13. Lietard J, Leger A, Erlich Y, Sadowski N, Timp W, Somoza MM. Chemical and photochemical error rates in light-directed synthesis of complex DNA libraries. Nucleic Acids Res. 2021;49:6687–701.

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

15. AVE Alliance Founding Members. The Atlas of Variant Effects (AVE) Alliance: understanding genetic variation at nucleotide resolution [Internet]. Zenodo; 2021 Mar. Available from: https://zenodo.org/record/7508716

16. Tabet D, Parikh V, Mali P, Roth FP, Claussnitzer M. Scalable Functional Assays for the Interpretation of Human Genetic Variation. Annu Rev Genet. 2022;56:441–65.

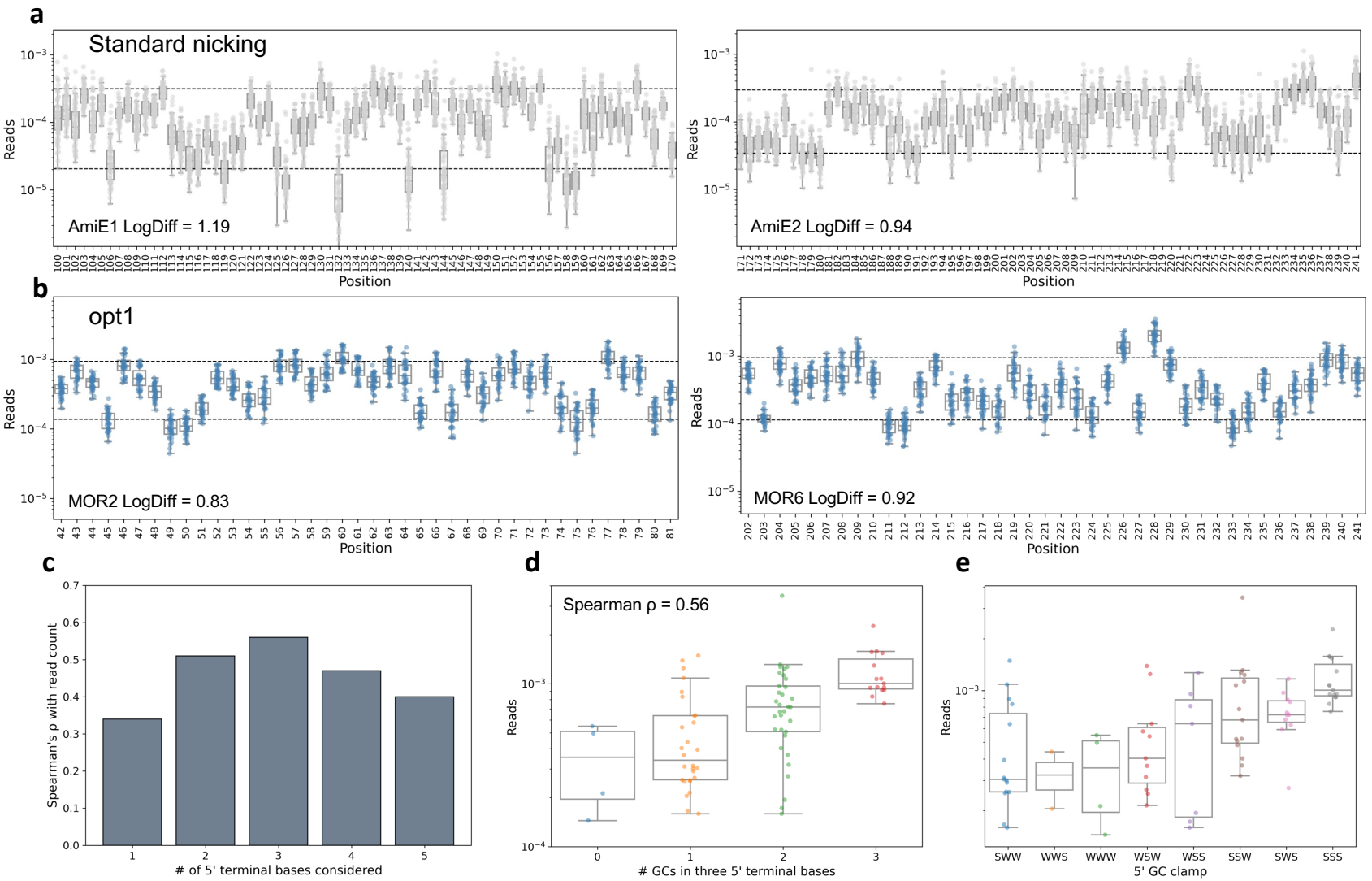
17. Kirby MB, Medina-Cucurella AV, Baumer ZT, Whitehead TA. Optimization of multi-site nicking mutagenesis for generation of large, user-defined combinatorial libraries. Protein Eng Des Sel. 2021;34:gzab017.

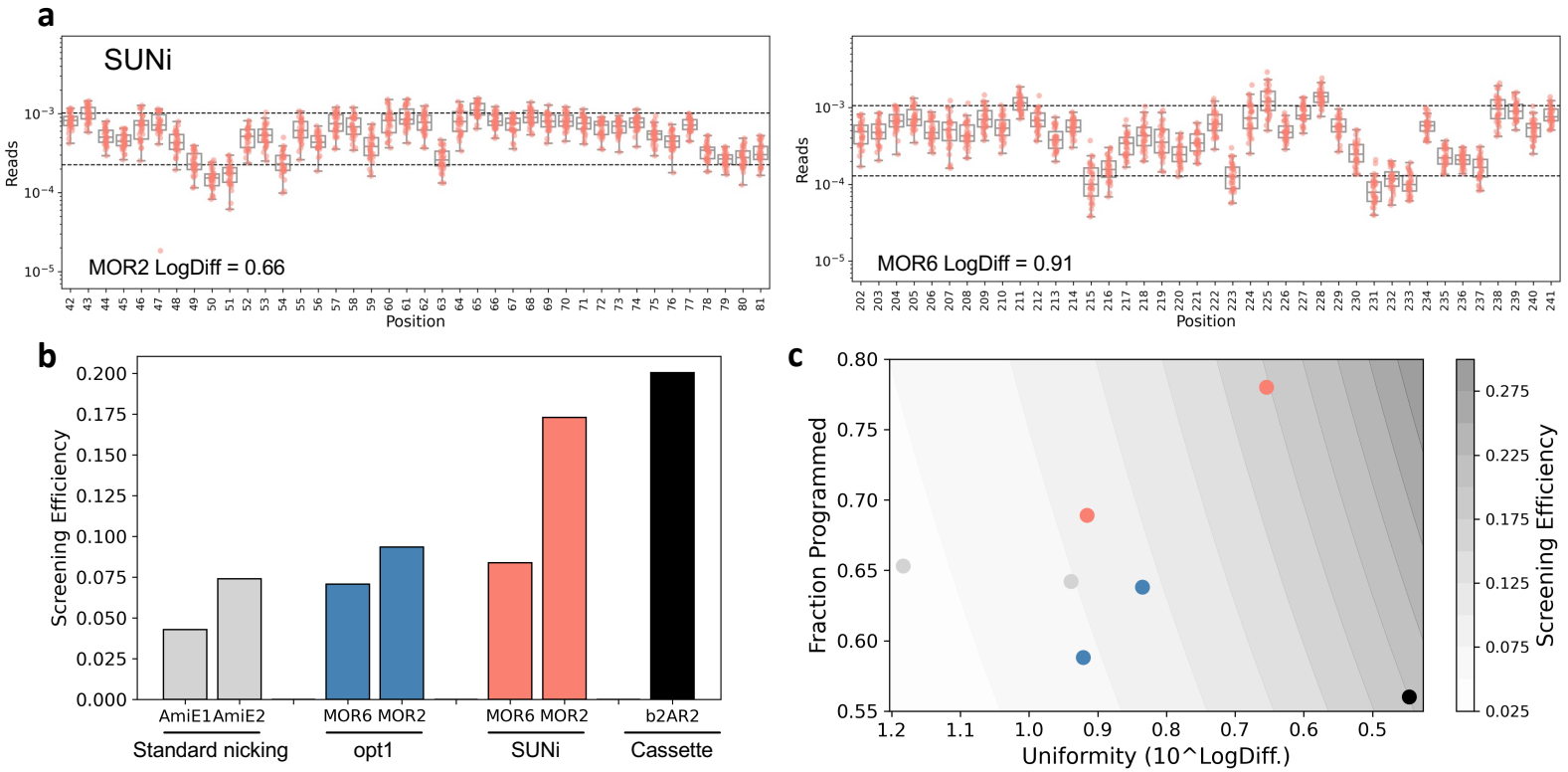
18. Medina-Cucurella AV, Steiner PJ, Faber MS, Beltrán J, Borelli AN, Kirby MB, et al. User-defined single pot mutagenesis using unamplified oligo pools. Protein Eng Des Sel. 2019;32:41–5.

19. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.

20. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. PeerJ Inc.; 2016;4:e2584.

21. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.


a
Standard nicking
AmiE1 LogDiff = 1.19
AmiE2 LogDiff = 0.94
Reads
Position
b
opt1
MOR2 LogDiff = 0.83
MOR6 LogDiff = 0.92
c
Spearman's ρ with read count
# of 5' terminal bases considered
d
Spearman ρ = 0.56
# GCs in three 5' terminal bases
e
SWW
WWS
WWW
WSW
WSS
SSW
SWS
SSS
5' GC clamp


a
SUNi
Reads
MOR2 LogDiff = 0.66
MOR6 LogDiff = 0.91
Position
b
Screening Efficiency
AmiE1 AmiE2
MOR6 MOR2
MOR6 MOR2
b2AR2
Standard nicking
opt1
SUNi
Cassette
c
Fraction Programmed
Uniformity (10^LogDiff.)
Screening Efficiency