8974–8985 *Nucleic Acids Research, 2022, Vol. 50, No. 15* *Published online 10 August 2022*
https://doi.org/10.1093/nar/gkac672

Nonlinear manipulation and analysis of large DNA datasets

Meiying Cui[1,†], Xueping Zhao[2,†], Francesco V. Reddavide[3], Michelle Patino Gaillez[1], Stephan Heiden[3], Luca Mannocci[4], Michael Thompson[3] and Yixin Zhang[1,*]

[1]B CUBE, Center for Molecular Bioengineering, Technische Universität Dresden, Dresden, Germany, [2]School of Mathematical Sciences, Xiamen University, China, [3]DyNAbind GmbH, Dresden, Germany and [4]DECLTech consulting, Switzerland

Received June 25, 2021; Revised June 18, 2022; Editorial Decision July 12, 2022; Accepted August 01, 2022

ABSTRACT

Information processing functions are essential for organisms to perceive and react to their complex environment, and for humans to analyze and rationalize them. While our brain is extraordinary at processing complex information, winner-take-all, as a type of biased competition is one of the simplest models of lateral inhibition and competition among biological neurons. It has been implemented as DNA-based neural networks, for example, to mimic pattern recognition. However, the utility of DNA-based computation in information processing for real biotechnological applications remains to be demonstrated. In this paper, a biased competition method for nonlinear manipulation and analysis of mixtures of DNA sequences was developed. Unlike conventional biological experiments, selected species were not directly subjected to analysis. Instead, parallel computation among a myriad of different DNA sequences was carried out to reduce the information entropy. The method could be used for various oligonucleotide-encoded libraries, as we have demonstrated its application in decoding and data analysis for selection experiments with DNA-encoded chemical libraries against protein targets.

INTRODUCTION

As the medium to encode genetic information, DNA has in recent years found many new applications, e.g. encoding chemical structures of combinatorial libraries, formation of self-assembled nanostructures, molecular computation, and data storage (1–4). Processing DNA-based information with DNA computation can facilitate bioanalyses in a direct and programmable manner. DNA computation methods have been developed for logical computation with synthetic biology systems (5), gaming (6–9), pattern recognition (10), sorting molecules with DNA robots (11), cellular analyses (12–16), modeling complex systems (17) and solving mathematical problems (18). However, DNA computation has rarely been integrated into an established biotechnological application workflow (e.g. drug discovery) to handle large datasets, for which the advantage of DNA-based parallel computation could be particularly attractive. Oligonucleotide libraries are essential for many biotechnologies, including screening of pathogen mutations, aptamer technology, protein/peptide display and DNA-encoded chemical library (DEL). Data analyses have also become increasingly complex with the growing power of next-generation sequencing (NGS). Suppose a DNA sample with complex information can be subjected to DNA-based parallel computation using a biologically relevant algorithm. In that case, the pre-processed data might lead to a more insightful, even direct reporting system. DELs are collections of organic compounds, individually coupled to distinctive DNA sequences as barcodes. DELs have become increasingly used in academia and industry for discovering small molecular compounds (SMC) to protein targets (19–21). In the past decade, the sizes of DELs have been growing dramatically, at a pace even faster than that of NGS. In 2008, when a DEL was analyzed by NGS for the first time, the library had 4000 compounds (1). Now DELs can have trillions of different compounds (22). Unlike the systematic evolution of ligands by exponential enrichment (SELEX) or phage display, amplification of DNA does not lead to the generation of DNA-encoded compounds. Resynthesis after selection is not possible for most of the DELs in practical use. Only a few reported DNA-directed library syntheses (23–26) that involve very sophisticated chemistry allow resynthesis after each round. In general, performing iterative selections is expensive as it requires a high amount of initial input. Each selection round drastically reduces the library available for the subsequent round. Synthesizing a high-quality library

[*]To whom correspondence should be addressed. Tel: +49 351 463 43040; Email: yixin.zhang1@tu-dresden.de
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

is always the most expensive, time-consuming, and labor-intensive part of DEL technology.

Systems with linearity are essential for many technologies, as they simplify the ways how we collect, analyse, process, and respond to different signals. However, there are many more complex processes, which cannot be reduced to a linear model. Winner-take-all (WTA) as a type of biased competition method, is among the simplest and most powerful competitive models for many biological and computational systems, including neural networks. (27,28). DNA computation with a WTA algorithm has found many interesting applications, including pattern recognition (10) and cancer diagnosis (29). While the methods have been designed to select a winner among a few sequences, it remains unknown whether there is a size limit (the number of different DNA species) to perform a WTA competition. If the most abundant species (winners) after selection can be further enriched and the minor members (noise) can be depleted, the downstream selection data analysis will be remarkably simpler. This work aimed to develop a biased competition function for Nonlinear manipulation and Analysis of DNA-Encoded Library (NADEL). A selection output is biased towards the abundant sequences (winners) through parallel molecular computation. Moreover, mathematical simulations demonstrated that a DNA library is converging to the winner sequence after iterative NADEL operations (Winner-take-all).

MATERIALS AND METHODS

Reagents

All oligonucleotides were purchased from IBA life sciences (Göttingen, Germany) and Metabion (Steinkirchen, Germany) in high-performance liquid chromatography (HPLC)-purified grade, molecular biology grade, Next-generation sequencing (NGS) grade or on the controlled pore glass (CPG) solid support based on different applications. Surveyor mutation detection kit was purchased from IDT DNA Technologies (Coralville, Iowa, USA). Building blocks of the libraries were purchased from Sigma-Aldrich (St. Louis, MO, USA), Enamine (Kiev, Ukraine), Alinda Chemical (Moscow, Russia), ChemBridge Corporation (San Diego, CA, USA) and Maybridge Chemical Company (Altrincham, UK). Other reagents were unless otherwise noted in the text, were purchased from Thermo Fisher Scientific (Waltham, MA, USA). T4 DNA ligase and T4 PNK were obtained from New England Biolabs (Massachusetts, USA).

Polymerase chain reaction (PCR) in NADEL

The PCR mixture (50 μl) for all samples contained 10× High GC buffer, primers (500 nM), dNTP mix (0.2 mM), and Phusion high-fidelity polymerase (1 U) and 30 ng of respective template DNA. Samples A-G and 5562-member library were amplified using (Primer A and Primer B), and (Primer C and D), respectively, with the following cycling conditions in PCR thermocycler (VWR, USA): 45 s at 98°C, 30 cycles of 30 s at 98°C, 1 min at 55°C, and 30 s at 72°C, closing the cycle, final extension for 10 min at 72°C, and storing at 4°C.

DNA-encoded chemical library selection output samples were amplified using primers Primer E and Primer F with the cycling conditions: 45 s at 98°C, 30 cycles of 30 s at 98°C, 1 min at 64°C, and 30 s at 72°C, closing the cycle, final extension for 10 min at 72°C, and storing at 4°C.

Overlap extension PCR on a 766 480-member library was performed with primers P and Q (Supplementary Table S6). The cycling condition is 45 s at 98°C, three cycles of 30 s at 98°C, 30 s at 55°C, and 30 s at 72°C and 17 cycles of 30 s at 98°C, 30 s at 64°C and 30 s at 72°C, followed by final extension for 10 min at 72°C.

PCR for Sanger sequencing

Before subjecting to Sanger sequencing, samples A-G and 5562-member library were further amplified with primer pairs (Primer G and Primer H), and (Primer I and Primer J), respectively, which can extend the length of the amplicon and therefore, ensure the read quality of code region of Sanger sequencing. The PCR protocol was as follows. The PCR mixture (50 μl) contained 10× High GC buffer, primers (500 nM), dNTP mix (0.2 mM), and Phusion high-fidelity polymerase (1U) and 1 μl of respective template DNA. The cycling conditions are 30 s at 98°C, then 2 cycles of: 30 s at 98°C, 1 min at 55°C, and 30 s at 72°C, closing the cycle, 18 cycles of: 30 s at 98°C, 1 min at 65°C, and 30 s at 72°C, followed by 10 min at 72°C, and storing at 4°C. Sanger sequencing primers for sample A-G, 5,562-member library, and selection output were Primer K, Primer L, and Primer M, respectively. Sanger sequencing was performed by Eurofins genomics, Germany. Sequences of all primers are listed in Supplementary Table S6.

Surveyor nuclease treatment

Before treatment, the PCR product was confirmed by gel electrophoresis by loading 2 μl of the reaction mixture on a 2% agarose gel (Bio-Rad, USA). After the product was confirmed to be a single band of the correct size, the remaining PCR product was mixed with 4.8 μl of $MgCl_2$ (150 mM), 1 μl of Surveyor Enhancer S (SES), and 2 μl of Surveyor Nuclease S (SNS) and incubated at 42°C for 1 h in the PCR thermocycler. The reaction was stopped by adding 1/10 volume of the stop solution.

Agarose gel electrophoresis, DNA purification and quantification

10 μl of each reaction mixture was transferred to new tubes for gel imaging. 5 μl of each sample was mixed with 1 μl of 6× agarose loading dye and loaded on a 2% gel. The agarose gel images were taken using ChemiDoc MP System with Image Lab (Bio-Rad, USA).

The rest were loaded on 2% agarose gel to separate full-length DNA from cleaved fragments. A constant voltage of 90 V was applied for 1.5 h. DNA bands were visualized by a UV transilluminator (UVP, Germany). The DNA bands of the correct size in both treated and untreated samples were sliced out and subjected to gel purification by using Nucleospin Gel and PCR clean-up kit (Macherey-Nagel, Germany). The concentration of purified DNA was measured with Nanodrop 2000 Spectrophotometers to detect the absorption at 260 nm. Purified DNA entered the next rounds of Surveyor nuclease treatment as described above.

Illumina high-throughput sequencing

All samples submitted to next-generation sequencing were PCR amplified to attach specific tags required for high-throughput sequencing. PCR product was recovered by gel extraction. PCR protocol was the same as the protocol for Sanger sequencing sample preparation. The concentration of each PCR product was measured by Nanodrop 2000 spectrophotometers, and samples were mixed in the same concentration and subjected to the second PCR amplification with the primers Illumina FOR and Illumina REV to attach adapter sequences compatible with the flow cell. The final amplification product, which contained information on all four samples was subjected to next-generation sequencing. Next-generation sequencing was performed with HiSeq2500 and Novaseq 6000 Next Generation Sequencer (Illumina).

Synthesis of 309 × 18-member DNA library with two code regions

The 309 × 18-member DNA library was generated by ligating two sets of oligonucleotides. The first 309 oligonucleotides were 77 nt long containing 25 nt variable regions, and the second set contained 18 oligonucleotides which were 48 nt long with 21 nt variable regions. The first set of oligonucleotides was pooled (each 1.5 nmol) and split into 18 tubes in equal amounts. Then each phosphorylated oligonucleotide (each 556 nmol) from the second set was added to each tube of the first set of oligonucleotides. Ligation was performed using T4 DNA ligase in 1X NEB2 buffer supplemented with 1 mM ATP in 100 μl volume. Sequence Q was generated by ligating two oligonucleotides, which belong to the two sets of oligonucleotides, respectively. Ligation of the oligonucleotides was performed in the same way as the encoding step in library synthesis. Both ligation products were PCR amplified and purified. The concentration of PCR products was then measured by Nanodrop 2000 Spectrophotometers. Then both samples were mixed accordingly to obtain a library, in which Q accounts for 20% of the total amount.

Selection against target proteins with DNA-encoded chemical libraries L1, L2 and L3

The target protein (P) was immobilized on *N*-hydroxysuccinimide (NHS)-activated Sepharose 4 Fast Flow (GE Healthcare, UK) via amide bond formation according to the manufacturer's instruction. The target protein on the beads was incubated with the library (1 nM per compound) L1 or L2 for 1 h at room temperature. After washing the beads three times with 1× PBS buffer with 0.05% Tween-20, the bound library members were eluted in 100 μl 10 mM Tris buffer with 0.05% Tween-20, pH 8.3 by heat at 95°C for 10 min.

The DNA-encoded library L3 and the selection output K0 were provided by DyNAbind GmbH. The construct of libraries L1, L2, and L3 are listed in Supplementary Table S7.

Mathematical fitting

We obtained the parameter value in the model by minimizing the difference between data and model, i.e. $\sum_{i=0}^{N}(f_i - D_i)^2$ where f_i is the value obtained from our model, D_i is the data from experiments, N is the number of data points in the dataset.

Data analysis

Illumina sequencing raw data was decoded by a custom python script. Briefly, code sequences in the raw sequences in the fastq files were extracted and assigned to the corresponding identity. This count of each code sequence was obtained by looping through the raw fastq file and counting the occurrence of the same code or code combination. Sequence abundance of each sequence was calculated by dividing each count by total counts.

All plots were generated by the software OriginLab 2019.

RESULTS

Computational simulation of sequence propagation at different NADEL efficiencies

The concept of the nonlinear manipulation and analysis of DEL (NADEL) is described in Figure 1. After PCR amplification of a DNA library carrying a diverse coding region, the library is subjected to a fast reshuffling process and kinetically trapped metastable state (Figure 1A). The resulting heteroduplexes are then selectively cleaved by mismatch-specific endonucleases, such as Surveyor nuclease, a widely used enzyme for mutation detection, error correction in synthetic DNA and quantifying genome editing efficiency (30–36). Because the relatively more abundant sequences have higher probabilities of forming homoduplexes during the reshuffling process, they are protected from cleavage. The homoduplexes, as the 'winners' surviving this process, but not the fragments from the cleaved heteroduplexes are further amplified by PCR. The biased sample can be then, either purified and sequenced, or subjected to the next NADEL cycle to further enhance the content of winner(s). Three reshuffling conditions with different ramp were tested and we discovered that a fast reshuffling can introduce the highest degree of heteroduplexes (Supplementary Figure S2).

To build a molecular circuit that exerts a biased competition function, we first simulated the propagation of various mixtures containing ten different DNA sequences (Figures 2, Supplementary Figure S3 and S4). Through analysis on the thermodynamics of DNA duplexes, we found that the formation of heteroduplex and homoduplex are equally probable at the reshuffling condition, as both structures contain stable complementary domains (Supplementary Figure S1). Therefore, to describe the sequence distribution of a DNA library after a NADEL cycle, we utilized the proportion of each sequence to represent the abundance in the system and the efficiency of consumption of heteroduplexes by the mismatch-detecting endonuclease was defined by factor α. The fraction P of the ith member in the mixture after a NADEL cycle is expressed by the following

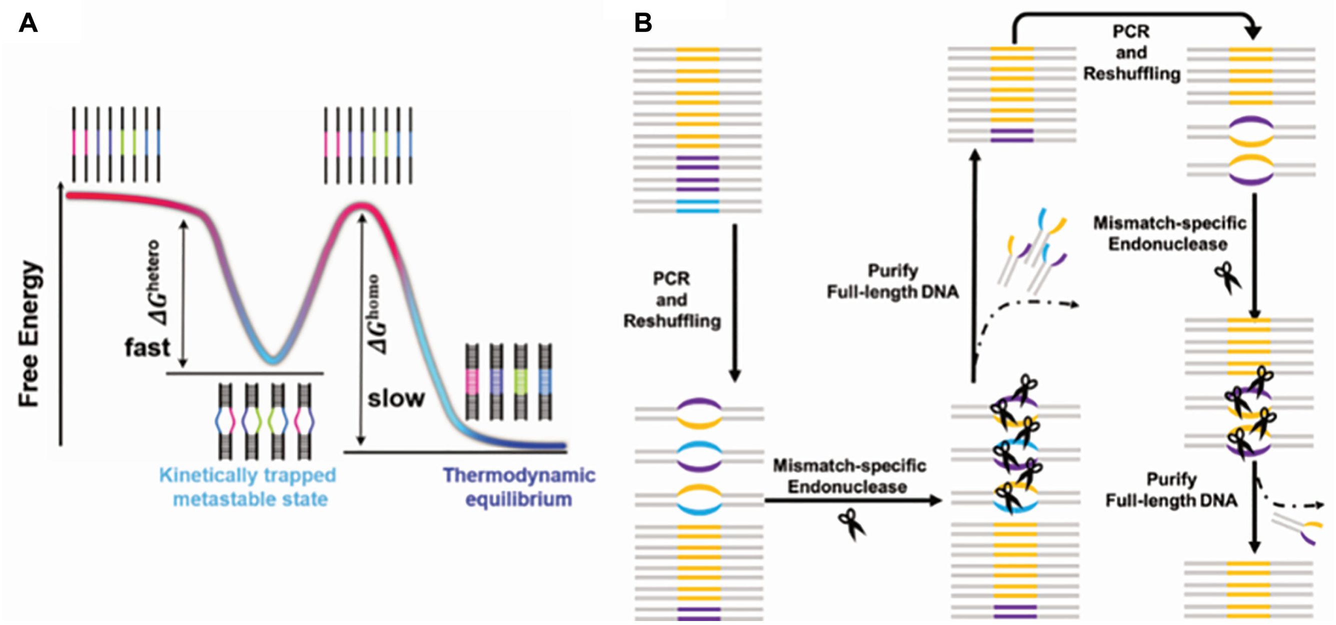


Figure 1. Schematic illustration of Nonlinear Manipulation and Analysis of DEL (NADEL). (**A**) Energy landscape in DNA hybridization among different sequences sharing universal primer sequences. Rapid cooling and the strong specific interactions between the complementary primer regions can lead to the formation of mismatched heteroduplexes trapped in a deep local minimum of free energy. The mismatched duplexes are targeted by the mismatch-detecting endonuclease. (**B**) After amplification by PCR, the library is subjected to fast reshuffling process to generate pool of homo-and hetero-duplexes. Resulting heteroduplexes can be digested by mismatch-specific endonuclease. Then only full-length DNA can be purified. Iteration of the process leads to nonlinear propagation of 'winner' DNA (orange code).

equation:

$$P(i) = \frac{p_i^2 + (1-\alpha)\, p_i\,(1-p_i)}{\sum_{\mathrm{j}=1}^{N} p_j^2 + (1-\alpha)\sum_{j=1}^{N} p_j\left(1-p_j\right)} \tag{1}$$

(Supplementary Figure S1). For a NADEL cycle at an ideal condition, α is 1, indicating all the mismatches generated are consumed by the nuclease, whereas α is 0 for an unbiased competition (i.e. in the absence of the nuclease). As NADEL is designed as an operation for various mixtures of DNA as datasets, it is essential to identify a deterministic complexity measure that can reliably describe a DNA library in terms of library diversity and sequence distribution. We found that information entropy (IE) a good descriptor to reflect a DNA population, as it measures the probability distribution quantitatively (37). Moreover, the calculation of IE includes all data points. (Equation 2).

$$\mathrm{IE}(X) = -\sum_{i=1}^{n} P(X_i)\log_2 P(X_i) \tag{2}$$

In our system, the probability of drawing a certain sequence equals to the sequence's abundance. IE is large when the library shows a uniform distribution of sequences, such as a library before selection. Like selection, NADEL reduces IE, pushing the population away from a uniform distribution. As shown in Figures 2, Supplementary Figure S3 and S4, the effect of NADEL was simulated in terms of sequence abundance, IE, and signal-to-noise ratio (SNR) with three α values (0, 0.5 and 1). For a mixture with only one member in large excess at the initial condition, iterative NADEL cycles lead to 100% sequence abundance of the winner; thus, IE converges to zero. However, when two sequences are equally abundant over others, the library converges to two winners.

Experimental validation of NADEL with 10-member DEL

To test NADEL experimentally, we designed a small library of 10 DNA sequences with 20 nt codes (Figure 3A and Supplementary Table S1), flanked by primers A and B. To ensure high sequence diversity of the small library, the difference between any two sequences was designed to be over 15. The ten sequences were mixed in seven different ratios (Figure 3A), mimicking various mixture compositions before and after DEL selections. For example, mixing ten sequences in equal amounts (sample A_0) represents an ideal library before selection. Sample B_0 represents the situation of one strong binder with some weak binders, while sample F_0 mimics a selected mixture with more than one moderate binder. As a representative example, after one cycle of NADEL, sample A showed a remarkable decrease of the 60 bp band on the gel, compared to the sample without treatment. The 60 bp DNA band was separated, purified, and submitted to the next NADEL cycles. To demonstrate the possibility of analyzing the NADEL products by Sanger sequencing, samples (B_0 – G_0) were subjected to NADEL cycles until the sequencing profile were dominated by the signals from the most abundant sequence with low ambiguity (Figure 3). As expected, the NADEL sequencing profiles of A_0, A_1, A_2 and A_3, where the subscript number is the number of NADEL cycles, are as noisy as the original untreated A_0 sample (Supplementary Figure S5A). For sam-

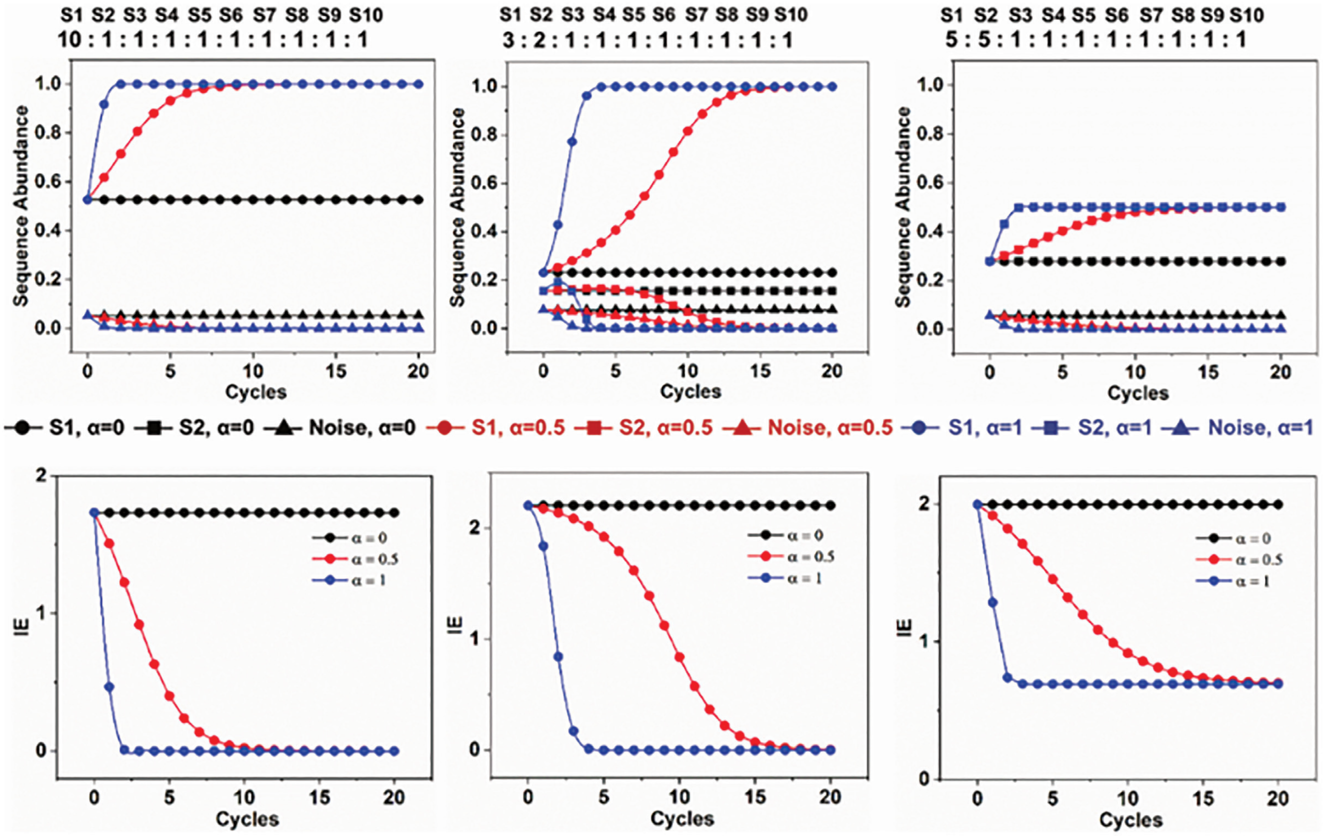


Figure 2. Simulated NADEL propagation of 10-member libraries of different initial ratios with three α values (α = 0, 0.5 and 1). The effect of NADEL was monitored from the aspects of sequence abundance (top panel) and information entropy (IE) (bottom). Any sequence from S3 to S10 (in triangles) is considered as noise.

ples, whose relative abundances of sequences are varied to mimic different selection results (samples B to G), the noise of sequencing profiles decreased after NADEL cycles. For B_0, C_0, D_0 and E_0, mimicking the presence of one major binder after the selection process, one to two NADEL cycles were sufficient to read the most abundant sequence S_1 without error (Figures 3C, D, and Supplementary Figure S5). The more challenging samples are G_0 and F_0, in which two sequences (S_1 and S_2) were in comparable excess (i.e. mimicking the presence of multiple similarly strong binders after library selection). G_0 could be particularly difficult because both sequences are in equal amount. Interestingly, a single NADEL cycle of F_0 led to the unambiguous identification of S_1 (Supplementary Figure S5B). Differently, after the first NADEL cycle, the sequencing profile of G_1 remained very noisy (Supplementary Figure S5F), while the noise has markedly decreased following three NADEL cycles (G_3, Figure 3D). Eventually, by analyzing the major peaks in the chromatogram, both sequences of S_1 and S_2 could be fully retrieved (Figure 3D). When there is more than one major species in relatively similar amount, NADEL cycles may lead to more than one major species in the final mixture in good agreement with the simulation (Figure 2). As we will show later, when analyzing the samples from real selection experiments with large DELs using NGS, multiple ‘winner’ sequences can be co-evolved against other ‘loser’ sequences.

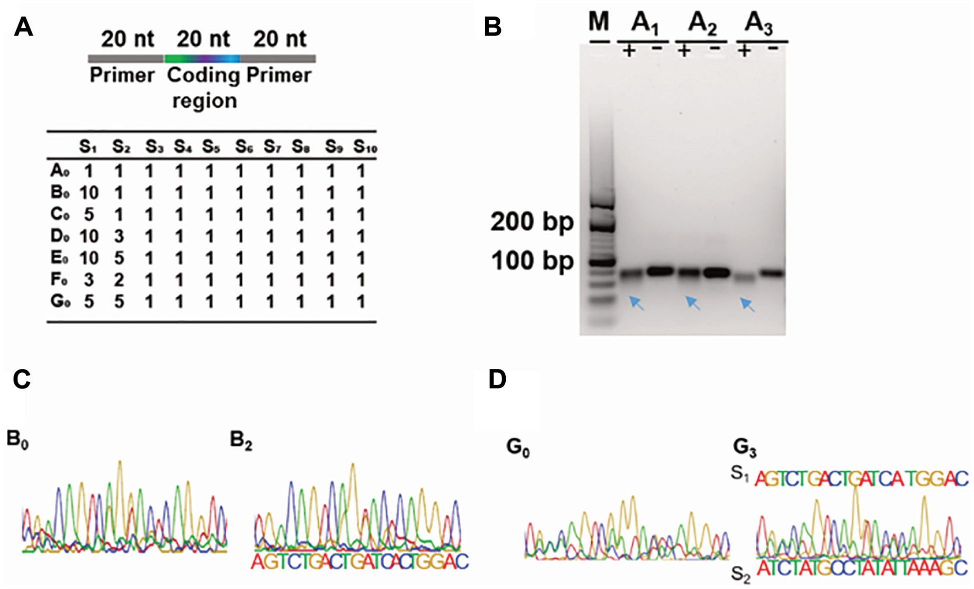
Experimental validation of NADEL with two-code DEL

Next, we tested the NADEL method with a DNA library containing two codes (Figure 4A). Codes 1 and 2 contain 309 and 18 different code sequences of 25 nt and 21 nt, respectively. The 5562-member library contains a sequence **Q**, accounting for 20% of the total library (Figure 4A).

The mole ratio of code 1 of **Q** to every other code 1 sequence is 77:1, and the ratio of code 2 of **Q** to every other code 2 sequence is 4.25:1. This design allowed us to test the NADEL method in a combinatorial library setup and evaluate its efficacy upon treating different diversities (e.g. code 1 309-member versus code 2 18-member). Through the four NADEL cycles, the noise of sequencing profiles reduced gradually. Interestingly, after four cycles, the sequencing profile of code region 1could be unambiguously assigned to code 1 of **Q** with zero error, whereas that of code region 2 contained six errors (Figures 4B and Supplementary Figure S6; errors in lower case). Nevertheless, the obtained sequence could still be assigned to code 2 of **Q**, as they show the least hamming distance, compared to the other library members (Supplementary Table S2).

Selection and NADEL with a 274-member DEL

To test the utility of NADEL with a real DEL library, we performed a selection experiment against protein target P



	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
A_0	1	1	1	1	1	1	1	1	1	1
B_0	10	1	1	1	1	1	1	1	1	1
C_0	5	1	1	1	1	1	1	1	1	1
D_0	10	3	1	1	1	1	1	1	1	1
E_0	10	5	1	1	1	1	1	1	1	1
F_0	3	2	1	1	1	1	1	1	1	1
G_0	5	5	1	1	1	1	1	1	1	1

Figure 3. NADEL with 10-member DNA libraries. (**A**) Scheme of libraries with 10 DNA sequences (S_1–S_{10}) in 7 different ratios (A_0–G_0). (**B**) Representative agarose gel electrophoresis of library A. **A_1+**: Sample digested by endonuclease in the first NADEL round. **A_1-**: Sample without endonuclease treatment. **A_2+**: Sample digested by endonuclease treatment in the second NADEL round. **A_2-**: Sample without endonuclease treatment. **A_3+**: Sample digested by endonuclease in the third NADEL round. **A_3-**: Sample without endonuclease treatment. Arrows indicate smear from cleavage. (**C**) Sanger sequencing chromatograms of library B before NADEL (B_0) and after two rounds of NADEL (B_2). (**D**) Sanger sequencing chromatograms of library G before NADEL (G_0) and after three rounds of NADEL (G_3).

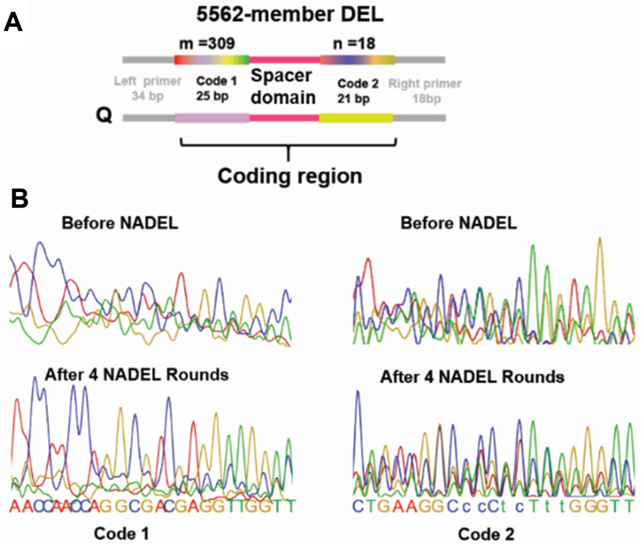


Figure 4. NADEL with a 5562-member DEL. (**A**) Scheme of a 309 × 18-member library with two code regions. Sequence Q accounts for 20% of the mixture. The spacer domain is 27 bp long. (**B**) Sanger sequencing chromatogram of code 1 and code 2 before NADEL and after four rounds of NADEL. Upper caps are the bases correctly read by Sanger sequencing, whereas lower caps are the bases incorrectly read by Sanger sequencing.

using a DEL of 274 compounds with one 20 nt coding region (L1) (Figure 5A) (the identity of the target protein is not relevant to this study). The selection output P0 and NADEL-treated samples P1, P2 and P3 were analyzed by Sanger sequencing. The sequencing profile of the sample after three rounds of NADEL (P3) showed a significant reduction of noise (Figure 5B). Intriguingly, Sanger sequencing chromatogram of P3 could be deconvoluted into two codes, X and Y, from major peaks and secondary signals respectively. The samples were subjected to NGS (Illumina) to confirm this result. Figures 5C and 5D show the relative abundances calculated from NGS results for all codes. As expected, the top two of the three codes, which became dominant after NADEL cycles (Figure 5D), were found to be X and Y, in good agreement with the Sanger sequencing chromatogram. Moreover, their relative abundances increased over the cycles, accounting for more than 87% of total reads after the third NADEL cycle, while the rest of the codes displayed declining sequence abundance.

We calculated the efficiency parameter α by fitting the NGS data with our mathematical model and obtained $\alpha = 0.76$. Comparison of experimental and the simulated results with $\alpha = 0.76$ in terms of IE and sequence abundance of each code in P1 to P3 are shown in Figures 5E and F–H, respectively. The experimental results were in high consistency with simulations.

We applied α value of 0.76 and simulated the Sanger sequencing chromatograms of the ten-member libraries with the calculated ratio of each sequence (Supplementary Figure S7–S10). In parallel, SNR was calculated from the experimentally obtained Sanger sequencing results and compared with the simulated results (Supplementary Figure S11). Both analyses show that Sanger sequencing could not reliably reflect the sequence distribution of a library. Because Sanger sequencing is mainly used to sequence one DNA species, it can only serve as a qualitative measurement to trace the SNR changes, rather than quantitatively analyzing the sequence distribution. Therefore, we utilized next-generation sequencing data to calculate the value of α.

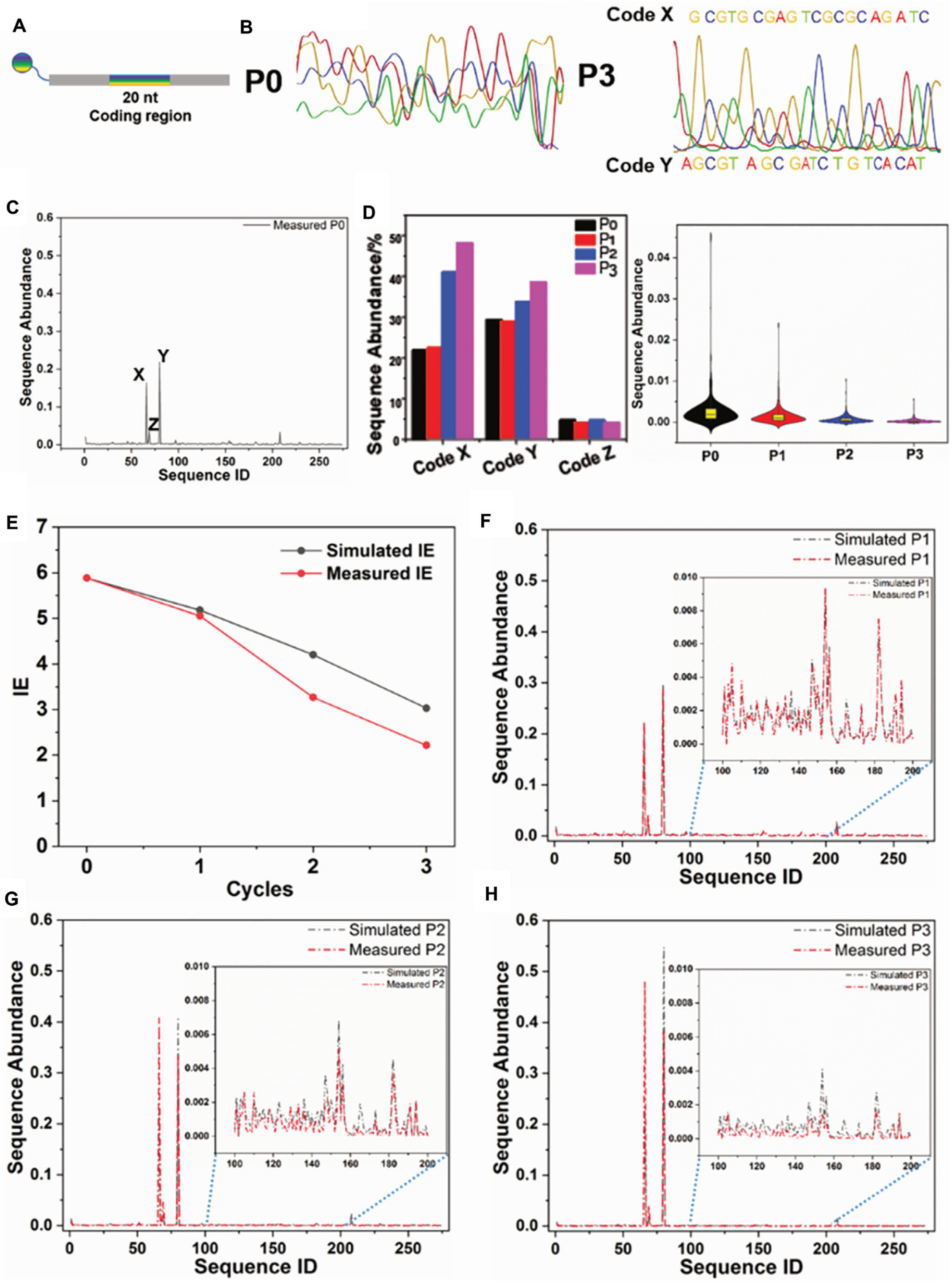


Figure 5. Selection and NADEL with a 274-member DEL. (**A**) Scheme of the 274-member DEL. (**B**) Sanger sequencing chromatogram of selection output before NADEL and after three rounds of NADEL. (**C**) Sequence abundance of library members from selection (P0). (**D**) Sequence abundance of code X, Y and Z before and over three NADEL cycles, as revealed by NGS (left). Violin plot representing sequence abundance of the rest 271 codes before and after NADEL (right) The yellow box inside of each distribution ranges from 25% to 75% of the data and black horizontal line indicates the median. (**E**) Simulated IE with the mathematical model *vs* experimentally observed IE. (**F**–**H**) Comparison between measured sequence abundance (red dashed line) and simulated result at each NADEL cycle (black dashed line). Simulation of IE and sequence abundance was obtained by applying α = 0.76 to the Equation (1).

Selection and NADEL with large DELs

We then tested NADEL with a 232 320-member two-building block library (640 × 363) (L2), synthesized by the split-and-pool method with an amino acid as the first building block to conjugate with amine-functionalized DNA tag and the second building block with carboxyl functionality that can react with the amino group from the first building block. The DNA tag structure is identical to the 5562-member DNA library (Figure 4A). The library was again selected against target protein P, and three rounds of NADEL were performed with the output. NGS monitored each round to trace the change in sequence distribution (Figure 6). The complexity of the sample was greatly reduced by eliminating sequences of low abundance (shaded region in Figure 5A and Supplementary Table S3). However, the IE change was not as remarkable as that of L1. We predicted the IE values of L2 with efficiency parameter α value fitted from L1 and found a good agreement with experimental results (Figure 6B), indicating that efficiency of enzymatic cleavage is similar in both libraries. Further, by in-depth simulation on libraries of varying sizes, we discovered that the large library size represents an important parameter explaining the creeping changes in IE and the abundance of the winner sequence (Figures 6C, D, and Supplementary Figure S12).

Because the enzyme used in our study is often employed for reliably detecting genome editing efficiency or mutation ratio (34–36), we tested if enlarging the DNA construct can improve the efficiency of enzymatic cleavage and therefore improve NADEL efficiency. Since DEL has a relatively small DNA construct compared to a gene, we wondered if increasing the construct length would enhance the efficiency. Thus, we chose a 766 480-member (880 × 871) DNA-encoded library (L3), with the largest construct and the largest library size currently available for us (Supplementary Figure S13). The library was selected against protein K. We increased the size of DNA of the selection output K0 from 140 bp to 227 bp by increasing the length of primer via overlap extension PCR. NADEL was carried out on the extended construct for three cycles. Surprisingly, NADEL efficiency was improved significantly and fitting the experimental data to the model revealed $\alpha > 0.99$. As demonstrated in Figures 5E, F and Supplementary Table S4, NADEL eliminated the low-abundance sequences tremendously, and IE displayed step-by-step decrease over three cycles. Further, on the same ranges of library sizes as in Figure 5C and D, we performed computational simulation using the calculated α (0.9994) from L3. With high NADEL efficiency, the changes in IE and abundance of the winner were remarkably accelerated (Figure 6G and H).

To confirm the fitted value of α experimentally, we measured the fold change of the amount upon enzyme treatment by quantitative PCR (qPCR). Remarkably, the experimentally obtained α was in good agreement with the fitted value (Supplementary Figure S14). Moreover, the effect of overlap extension PCR was also observed with L1, as reflected by the increased value of α (Supplementary Figure S14).

In addition, we examined if the length of spacer domain (SD) influences NADEL efficiency by altering the size of mismatched loop. Therefore, we prepared four mock libraries with different SD sizes. The scheme of the DNA construct is shown in Supplementary Figure S15, and the sequences are listed in Supplementary Table S5. In each mock library, four sequences (S1, S2, S3 and S4) were mixed with a partially degenerate library (SN) of 4^{12} diversity in the abundance of 20%, 5%, 5%, 5% and 65%, respectively. By NGS, we traced the abundance of the four sequences over three NADEL cycles. We calculated α value for four libraries from the experimental results and did not observe a clear correlation between the SD length and NADEL efficacy (Supplementary Figure S15). In the future, studies on different DEL constructs will help us to design libraries with high NADEL efficacy.

DISCUSSION

Through biased molecular competition, the NADEL approach allowed us to reduce the complexity of DNA libraries of different scales, from ten to sub-million members. As DEL is the only large dataset with individually designed and synthesized DNA codes available to perform NADEL, in this work, we have investigated the utility of NADEL in DEL technology. Unlike silicon-based computers, the massive parallelism associated with NADEL is especially attractive for large datasets. When applied to a DEL of 200 000 compounds, NADEL involves competitions among 20 billion different DNA duplexes. Current cutting-edge NGS platforms (e.g. Illumina Novaseq 6000) can provide up to 20 billion reads per sequencing run. To analyze the sequencing reads confidently, it is desired to oversample the library by a factor of 10 (38,39). However, it has become common to index each selection and pool multiple selections together in one sequencing run to reduce cost and time. As a result, total sequence counts for each selection in this work and in reported literature is often in the range of sub-million to a few millions regardless of library size (40–42), covering only a fraction of the library. Thus, current sequencing depth cannot cover the libraries of billions to trillions size, resulting in zero-copy counts for a large portion of the library members. While the NADEL method remains to be further developed, it will never replace the standard DEL analysis protocol (38). Instead, it can provide an alternative way to investigate the selection data, complementary to the direct sampling approach, by analyzing a pool of DNA preprocessed by molecular computation. NADEL can also be used for image processing to enhance the contrast. When random subsets of the L2 and L3 libraries are presented 100 × 100 gray scale pictures, with each unique sequence representing a pixel and its abundance representing the pixel value, NADEL can help increase the contrast (Figure 7 and Supplementary Figure S16). In the future, studying and optimizing NADEL efficiency on more diverse DNA constructs as well as larger libraries (i.e. three-building block DELs) would deepen our understanding on this nonlinear process and bring its applications into drug discovery as well as other fields. Another very important direction for future development is the automation of NADEL, for

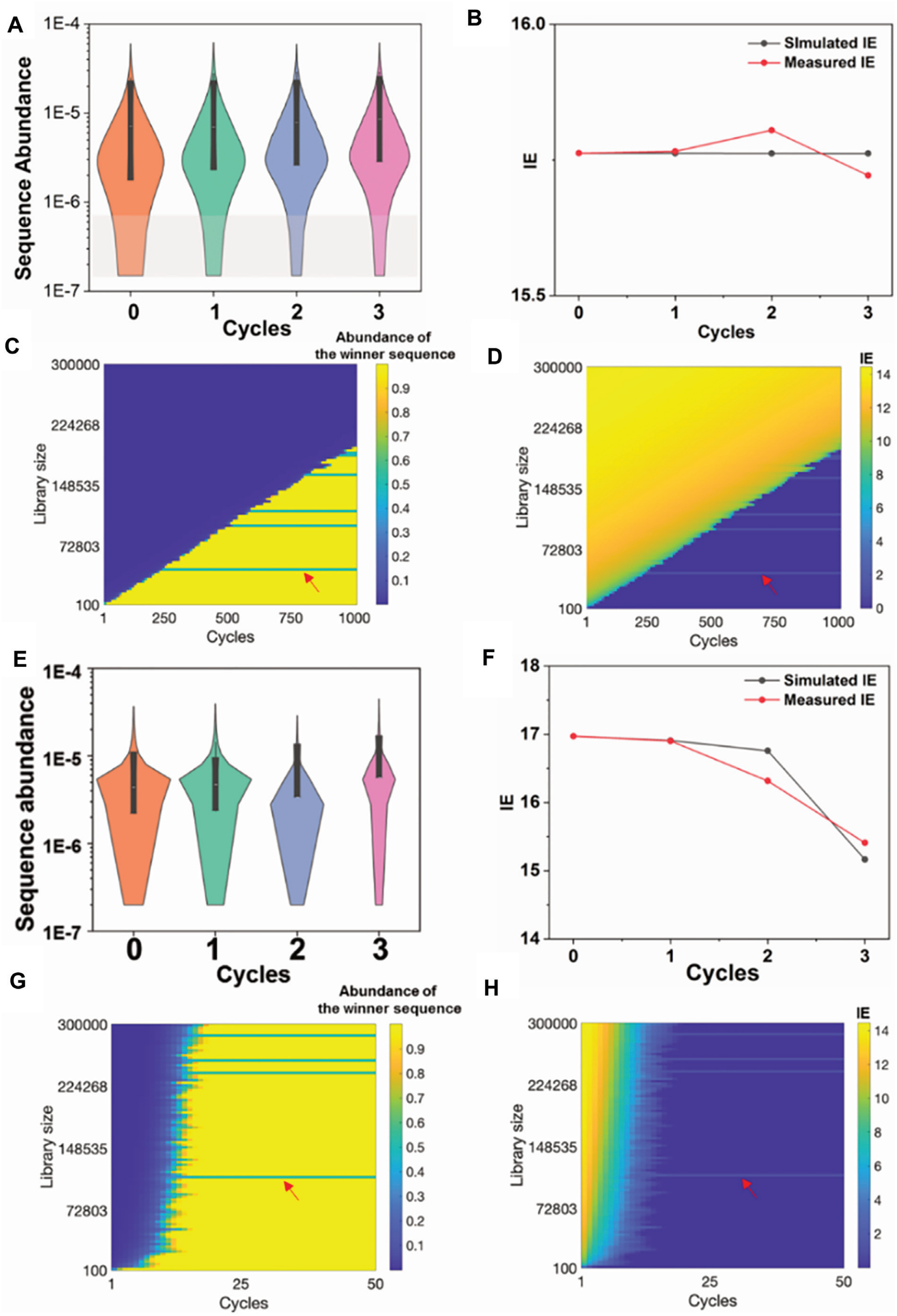


Figure 6. Selection and NADEL with large DELs. (**A**) Violin plots of sequence abundance from 232 320-member library before and after NADEL. The shaded region indicates that NADEL can remove many low-abundance noise sequences. The box inside of each distribution ranges from 10% to 90% of the population. (**B**) Comparison between simulated IE (black) and experimentally obtained IE (red) over three NADEL cycles. IE was calculated with α = 0.76. Simulated sequence abundance of winner (**C**) and IE (**D**) of libraries from 100 members to 300 000 members (α = 0.76). (**E**) Violin plots of sequence abundance from 766,480-member library before and after NADEL. The box inside of each distribution ranges from 10% to 90% of the population. (**F**) Comparison between simulated IE (black) and experimentally obtained IE (red) over three NADEL cycles. IE was calculated with α = 0.9994. Simulated sequence abundance of winner (**G**) and IE (**H**) of libraries from 100 members to 300,000 members (α = 0.9994). Horizontal lines in the heatmaps (e.g. those pointed by red arrows) represent situations where more than one winner sequences are of the same abundance in the randomly generated libraries. Therefore, NADEL can never produce a single winner from these libraries.

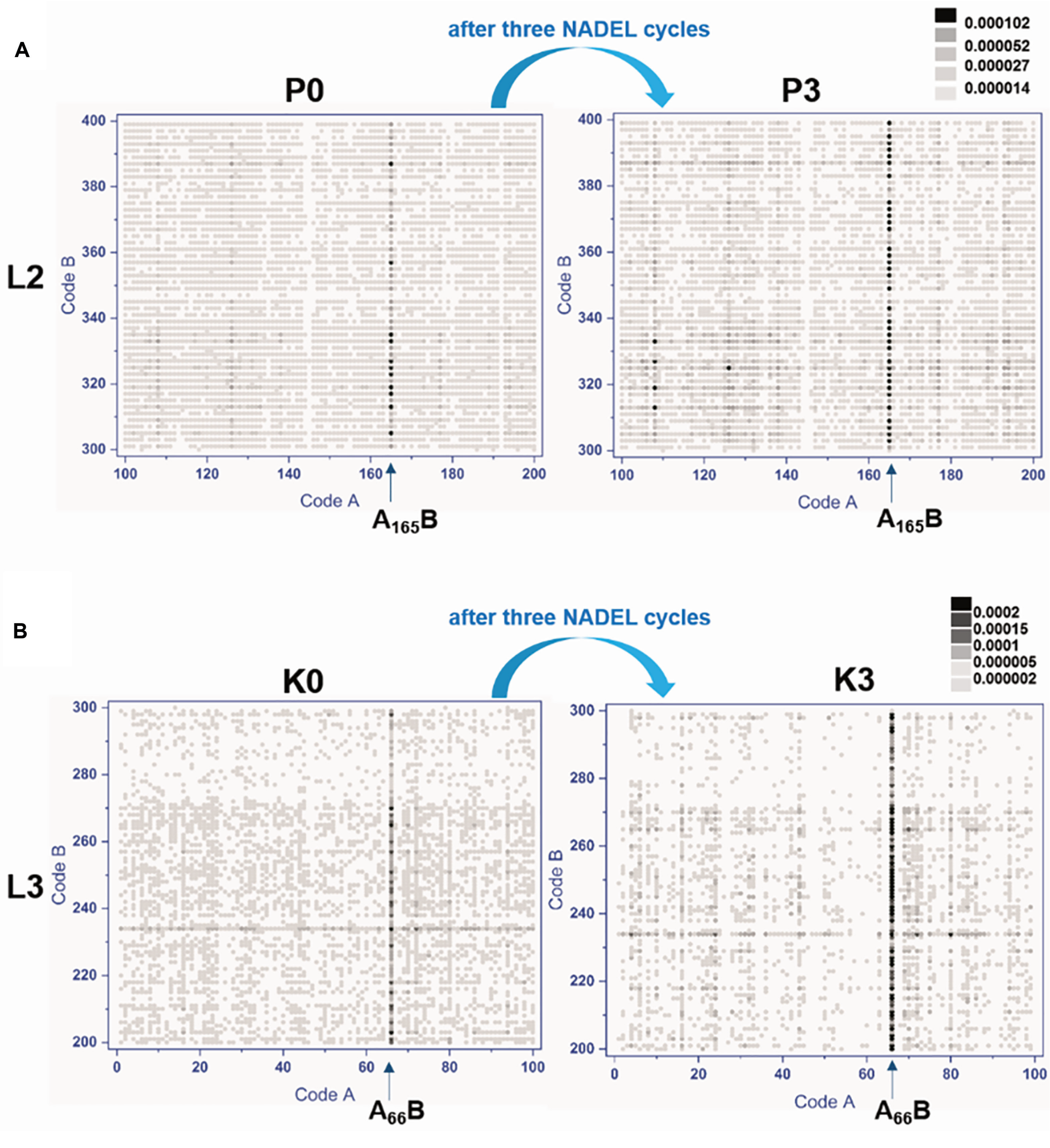


Figure 7. Enhancing image contrast by NADEL. (**A**) A 100 × 100 grey scale image as a subset of 640 × 363 grey scale images (L2) before and after three cycles of NADEL treatment. (**B**) A 100 × 100 grey scale image as a subset of the 880 × 871 grey scale images (library L3) before and after three cycles of NADEL treatment. Full images of L2 and L3 are shown in Supplementary Figure S15.

example, by building a microfluidics device. It will allow us to perform tens of NADEL cycles readily, to demonstrate and realize the nonlinear effect predicted by the simulations.

DATA AVAILABILITY

Next generation sequencing raw data are available at SRA database with the accession number PRJNA807148 and the link https://www.ncbi.nlm.nih.gov/sra/PRJNA807148.

CODE AVAILABILITY

All codes for mathematical fitting and decoding for NGS data are available at https://github.com/Sarah0802/Biased-Competition-NADEL.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Ulrike Hofmann for technical support and Dr. Kathrin Lang from DKMS Life Science Lab (Germany) and Dr. Elena Ferrari and Meng Chen from Novogene (UK) for supporting on Next-Generation Sequencing.

FUNDING

German Federal Ministry of Research and Education (BMBF) [03Z2EN12 and 03Z2E511]. Funding for open access charge: German Federal Ministry of Research and Education (BMBF) [03Z2EN12 and 03Z2E511] and SLUB/TU Dresden.

Conflict of interest statement. DyNAbind GmbH is commercializing the DNA-Encoded Library technology reported in this publication. F.V. Reddavide, Y. Zhang, L. Mannocci and M. Thompson are shareholders of DyNAbind GmbH.

REFERENCES

1. Mannocci,L., Zhang,Y., Scheuermann,J., Leimbacher,M., De Bellis,G., Rizzi,E., Dumelin,C., Melkko,S. and Neri,D. (2008) High-throughput sequencing allows the identification of binding molecules isolated from DNA-encoded chemical libraries. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 17670–17675.
2. Seelig,G., Soloveichik,D., Zhang,D.Y. and Winfree,E. (2006) Enzyme-free nucleic acid logic circuits. *Science*, **314**, 1585–1589.
3. Rothemund,P.W.K. (2006) Folding DNA to create nanoscale shapes and patterns. *Nature*, **440**, 297–302.
4. Organick,L., Ang,S.D., Chen,Y.J., Lopez,R., Yekhanin,S., Makarychev,K., Racz,M.Z., Kamath,G., Gopalan,P., Nguyen,B. *et al.* (2018) Random access in large-scale DNA data storage. *Nat. Biotechnol.*, **36**, 242–248.
5. Joesaar,A., Yang,S., Bögels,B., van der Linden,A., Pieters,P., Kumar,B.P., Dalchau,N., Phillips,A., Mann,S. and de Greef,T.F. (2019) DNA-based communication in populations of synthetic protocells. *Nat. Nanotechnol.*, **14**, 369–378.
6. Qian,L., Winfree,E. and Bruck,J. (2011) Neural network computation with DNA strand displacement cascades. *Nature*, **475**, 368–372.
7. Pei,R., Matamoros,E., Liu,M., Stefanovic,D. and Stojanovic,M.N. (2010) Training a molecular automaton to play a game. *Nat. Nanotechnol.*, **5**, 773–777.
8. Chao,J., Wang,J., Wang,F., Ouyang,X., Kopperger,E., Liu,H., Li,Q., Shi,J., Wang,L., Hu,J. *et al.* (2019) Solving mazes with single-molecule DNA navigators. *Nat. Mater.*, **18**, 273–279.
9. Song,J., Li,Z., Wang,P., Meyer,T., Mao,C. and Ke,Y. (2017) Reconfiguration of DNA molecular arrays driven by information relay. *Science*, **357**, eaan3377.
10. Cherry,K.M. and Qian,L. (2018) Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature*, **559**, 370–376.
11. Srinivas,N., Parkin,J., Seelig,G., Winfree,E. and Soloveichik,D. (2017) Enzyme-free nucleic acid dynamical systems. *Science*, **358**, eaal2052.
12. Han,D., Zhu,G., Wu,C., Zhu,Z., Chen,T., Zhang,X. and Tan,W. (2013) Engineering a cell-surface aptamer circuit for targeted and amplified photodynamic cancer therapy. *ACS Nano*, **7**, 2312–2319.
13. You,M., Zhu,G., Chen,T., Donovan,M.J. and Tan,W. (2015) Programmable and multiparameter DNA-based logic platform for cancer recognition and targeted therapy. *J. Am. Chem. Soc.*, **137**, 667–674.
14. Rudchenko,M., Taylor,S., Pallavi,P., Dechkovskaia,A., Khan,S., Butler,V.P. Jr, Rudchenko,S. and Stojanovic,M.N. (2013) Autonomous molecular cascades for evaluation of cell surfaces. *Nat. Nanotechnol.*, **8**, 580–586.
15. Chang,X., Zhang,C., Lv,C., Sun,Y., Zhang,M., Zhao,Y., Yang,L., Han,D. and Tan,W. (2019) Construction of a multiple-aptamer-based DNA logic device on live cell membranes via associative toehold activation for accurate cancer cell identification. *J. Am. Chem. Soc.*, **141**, 12738–12743.
16. Douglas,S.M., Bachelet,I. and Church,G.M. (2012) A logic-gated nanorobot for targeted transport of molecular payloads. *Science*, **335**, 831–834.
17. Han,D., Wu,C., You,M., Zhang,T., Wan,S., Chen,T., Qiu,L., Zheng,Z., Liang,H. and Tan,W. (2015) A cascade reaction network mimicking the basic functional steps of adaptive immune response. *Nat. Chem.*, **7**, 835–841.
18. Adelman Leonard,M. (1994) Molecular computations of solutions to solve combinatorial problems. *Science*, **266**, 1021–1024.
19. Neri,D. and Lerner,R.A. (2018) DNA-encoded chemical libraries: a selection system based on endowing organic compounds with amplifiable information. *Annu. Rev. Biochem.*, **87**, 479–502.
20. Zhao,G., Huang,Y., Zhou,Y., Li,Y. and Li,X. (2019) Future challenges with DNA-encoded chemical libraries in the drug discovery domain. *Expert Opin. Drug Discov.*, **14**, 735–753.
21. Song,M. and Hwang,G.T. (2020) DNA-encoded library screening as a core platform technology in drug discovery: its synthetic method development and applications in DEL synthesis. *J. Med. Chem.*, **63**, 6578–6599.
22. Halford,B. (2017) How DNA-encoded libraries are revolutionizing drug discovery. *C&EN Global Enterp.*, **95**, 28–33.
23. Gartner,Z.J., Tse,B.N., Grubina,R., Doyon,J.B., Snyder,T.M. and Liu,D.R. (2004) DNA-templated organic synthesis and selection of a library of macrocycles. *Science*, **30**, 1601–1605.
24. Hansen,M.H., Blakskjaer,P., Petersen,L.K., Hansen,T.H., Hoøjfeldt,J.W., Gothelf,K.V. and Hansen,N.J.V.(2009) A yoctoliter-scale DNA reactor for small-molecule evolution. *J. Am. Chem. Soc.*, **131**, 1322–1327.
25. Halpin,D.R. and Harbury,P.B. (2004) DNA display i. Sequence-encoded routing of DNA populations. *PLoS Biol.*, **2**, 1015–1021.
26. Vummidi,B.R., Farrera-Soler,L., Daguer,J., Dockerill,M., Barluenga,S. and Winssinger,N. (2022) A mating mechanism to generate diversity for the darwinian selection of DNA-encoded synthetic molecules. *Nat. Chem.*, **14**, 141–152.
27. Kaski,S. and Kohonen,T. (1994) Winner-take-all networks for physiological models of competitive learning. *Neural Netw.*, **7**, 973–984.
28. Mori,K., Nagao,H. and Yoshihara,Y. (1999) The olfactory bulb: coding and processing of odor molecule information. *Science*, **286**, 711–715.
29. Zhang,C., Zhao,Y., Xu,X., Xu,R., Li,H., Teng,X., Du,Y., Miao,Y., Lin,H.C. and Han,D. (2020) Cancer diagnosis with DNA molecular computation. *Nat. Nanotechnol.*, **15**, 709–715.
30. Hughes,R.A., Miklos,A.E. and Ellington,A.D. (2012) Enrichment of error-free synthetic DNA sequences by CEL i nuclease. *Curr. Protoc. Mol. Biol.*, https://doi.org/10.1002/0471142727.mb0324s99.

31. Saaem,I., Ma,S., Quan,J. and Tian,J. (2012) Error correction of microchip synthesized genes using surveyor nuclease. *Nucleic Acids Res.*, **40**, e23.
32. Ma,S., Saaem,I. and Tian,J. (2012) Error correction in gene synthesis technology. *Trends Biotechnol.*, **30**, 147–154.
33. Woo,J.W., Kim,J., Kwon,S.Il, Corvalán,C., Cho,S.W., Kim,H., Kim,S.G., Kim,S.T., Choe,S. and Kim,J.S. (2015) DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nat. Biotechnol.*, **33**, 1162–1164.
34. Bai,Y., He,L., Li,P., Xu,K., Shao,S., Ren,C., Liu,Z., Wei,Z. and Zhang,Z. (2016) Efficient genome editing in chicken DF-1 cells using the CRISPR/Cas9 system. *G3 Genes Genomes Genet.*, **6**, 917–923
35. Wolfs,J.M., Hamilton,T.A., Lant,J.T., Laforet,M., Zhang,J., Salemi,L.M., Gloor,G.B., Schild-Poulter,C. and Edgell,D.R. (2016) Biasing genome-editing events toward precise length deletions with an RNA-guided tevcas9 dual nuclease. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 14988–14993.
36. Yu,C., Zhang,Y., Yao,S. and Wei,Y. (2014) A PCR based protocol for detecting indel mutations induced by TALENs and CRISPR/Cas9 in zebrafish. *PLoS One*, **9**, e98282.
37. Jaynes,E.T. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 620.
38. Decurtins,W., Wichert,M., Franzini,R.M., Buller,F., Stravs,M.A., Zhang,Y., Neri,D. and Scheuermann,J. (2016) Automated screening for small organic ligands using DNA-encoded chemical libraries. *Nat. Protoc.*, **11**, 764–780.
39. Kuai,L., O'Keeffe,T. and Arico-Muendel,C. (2018) Randomness in DNA encoded library selection data can be modeled for more reliable enrichment calculation. *SLAS DIiscov.*, **23**, 405–416.
40. Li,Y., De Luca,R., Cazzamalli,S., Pretto,F., Bajic,D., Scheuermann,J. and Neri,D. (2018) Versatile protein recognition by the encoded display of multiple chemical elements on a constant macrocyclic scaffold. *Nat. Chem.*, **10**, 441–448.
41. Bassi,G., Favalli,N., Vuk,M., Catalano,M., Martinelli,A., Trenner,A., Porro,A., Yang,S., Tham,C.L. *et al.* (2020) A single-stranded DNA-Encoded chemical library based on a stereoisomeric scaffold enables ligand discovery by modular assembly of building blocks. *Adv. Sci.*, **7**, 2001970.
42. Favalli,N., Bassi,G., Pellegrino,C., Millul,J., De Luca,R., Cazzamalli,S., Yang,S., Trenner,A., Mozaffari,N.L., Myburgh,R. *et al.* (2021) Stereo-and regiodefined DNA-encoded chemical libraries enable efficient tumour-targeting applications. *Nat. Chem.*, **13**, 540–548.