Rapid desktop synthesis of degenerate oligonucleotides using Enzymatic DNA Synthesis (EDS)

EDS, on the SYNTAX™ platform, delivers more accurate base distribution at degenerate positions compared to conventional chemical DNA synthesis (CDS).

Synthesis of degenerate sequences constitutes a powerful tool in the rapid generation of genetic diversity, benefiting multiple fields such as protein engineering applied to therapeutic or industrial processes. In this Technical Note, we demonstrate the performance of DNA Script's Enzymatic DNA Synthesis (EDS) in the synthesis of oligonucleotides harboring degenerate sequences. Degenerate oligonucleotides produced by EDS technology resulted in base and codon distributions closely matching the expected theoretical frequencies at degenerate positions and performed better in comparison with oligos produced by conventional Chemical DNA Synthesis (CDS). With its ease of use and 96-well format, DNA Script's SYNTAX platform provides a time-saving and flexible way of synthesizing degenerate oligonucleotides using EDS technology, allowing to speed up mutant enzyme screening processes and other downstream applications requiring high-quality degenerate sequence.

INTRODUCTION

Degenerate oligonucleotides (oligos) contain random bases located at specific sites within a nucleotide sequence. A variety of applications utilize degenerate oligos in their experimental design, particularly in protein engineering studies[1,2] aiming to screen for optimized mutants, but also in the biomedical, research, and diagnostics fields where degenerate sequences are used for sequencing with Unique Molecule Identifiers (UMIs) in liquid biopsy applications or single cell experiments for instance (Figure 1).

In modern protein engineering, the directed evolution of enzymes has emerged as a strategy to optimize specific molecular properties, such as thermostability, enantioselectivity, and catalytic activity.[2-5] These are relevant to multiple fields, such as in bio-catalysis, bioproduction, or the development of therapeutic drugs in the pharmaceutical and research fields. Optimization of enzyme properties

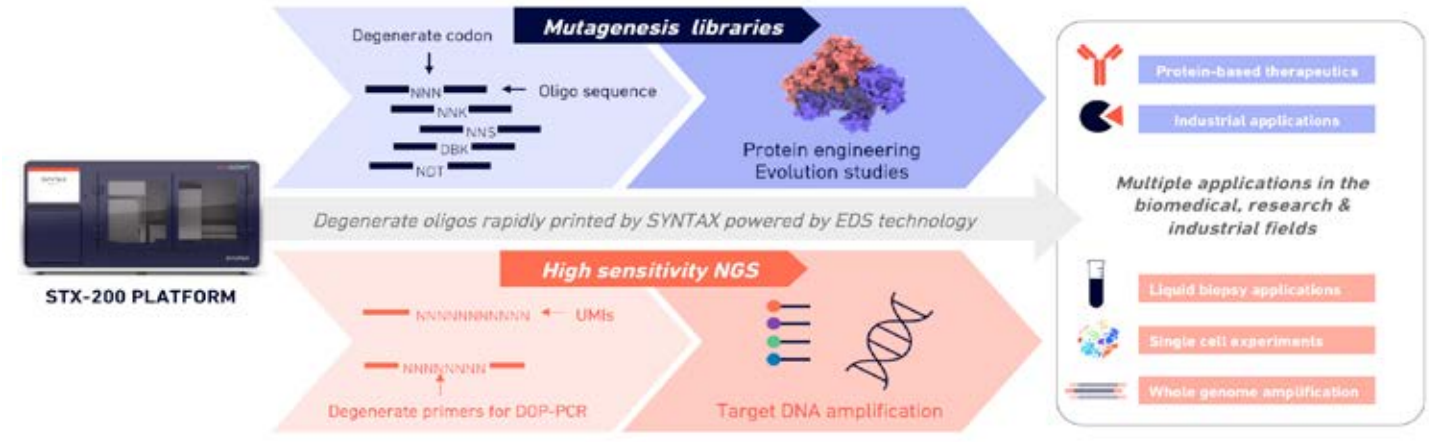


FIGURE 1. DEGENERATE OLIGOGNUCLEOTIDES AND THEIR APPLICATIONS IN THE BIOMEDICAL FIELD

Oligonucleotides containing degenerate bases can be rapidly synthesized with the SYNTAX platform, resulting in highly accurate and uniform base distributions. They can then be used in multiple applications in the biomedical and research fields, including in protein engineering studies where they enable fast generation of mutagenesis libraries for therapeutic and industrial purposes. UMIs: unique molecular identifier; DOP-PCR: degenerate oligonucleotide-primed PCR.

can be achieved by using degenerate oligos to create enzyme mutant libraries, in which one or multiple amino acid residues are changed into all possible residues. This experimental approach is referred to as saturation mutagenesis and provides an efficient way to rapidly generate large and diverse mutant libraries.

Degeneracy refers to positions within a DNA sequence in which more than one base is incorporated. Different letters indicate different combinations of nucleotides (mix of either 2, 3, or 4 bases), N refers to the combination of the 4 bases A, C, G, and T (Figure 2A).

In single-site saturation mutagenesis, a single codon is substituted into all possible codons by the incorporation of an NNN degenerate codon. At the protein level, this results in all possible amino acid (AA) residues at a specific position (Figure 2B). However, due to the redundancy of the genetic code, an NNN codon degeneracy will result in 64 distinct codon possibilities for only 20 biologically encoded AAs. Therefore, there will be an uneven and biased proportion of AAs generated, and a significant oversampling is required to obtain the desired diversity. To reduce codon redundancy and library screening efforts, multiple approaches utilize degeneracies targeting smaller sets of codons, such as NNK degeneracy (32 possible codons for 20 AAs),[6] the Trick22c[7] (mixture of three degenerate primers encoding 22 unique codons), Tang-20c[8] or other more focused libraries such as NDT degeneracy.[9] Due to the trade-off between reduced redundancy, low complexity, and low cost (only one degenerate primer pair is needed), NNK-based degeneracy remains the conventional approach to generate saturation mutagenesis libraries. Since biases in base incorporation can strongly affect the equal representation of each sequence within a degenerate library and thus have repercussions on the protein engineering application, uniformity of base distribution is critical when synthesizing degenerate oligonucleotides.

DNA Script's proprietary EDS technology[10] enables a rapid, on-demand, in-house, and cost-effective synthesis of DNA oligos on a benchtop instrument. This contrasts with the highly centralized production of oligos using conventional – and harsher – phosphoramidite-based CDS, which requires users to rely on third-party logistics.

This Technical Note describes the ability of EDS technology to synthesize degenerate oligos, testing five commonly used codon degenerate patterns: NNN,

A–Degenerate code

Symbol	Meaning
N	A / C / G /T
V	A / C / G
H	A / C / T
D	A / G / T
B	C / G / T
M	A / C
K	G /T
Y	C / T
R	A / G
S	C / G
W	A / T

B– Generation of site-saturation variant library using NNN codon

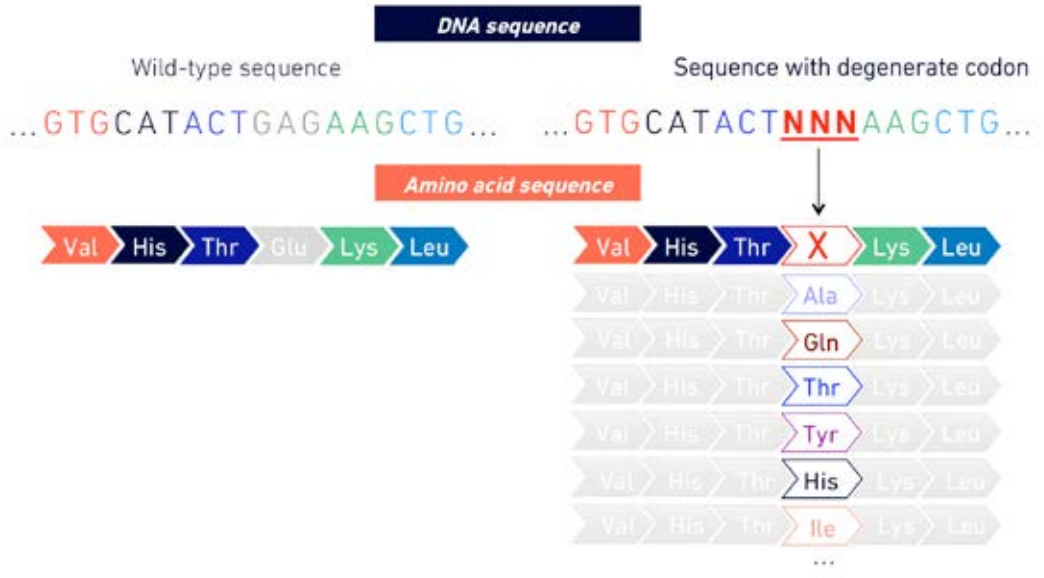


FIGURE 2. DEGENERATE BASES: NOMENCLATURE (A) AND THEIR USE IN THE GENERATION OF SITE-SATURATION VARIANT LIBRARIES (B)

A. Each letter refers to a mixture of either 2, 3, or 4 standard nucleotides (A, C, G, T). **B**. A common application of degenerate bases is the generation of variant libraries for protein engineering studies, by site-saturation mutagenesis. In the example, a DNA sequence containing a NNN degenerate codon may translate into a library containing all amino acid residues (indicated by their 3-letter code on the illustration) at a particular site.

NNK, NNS, NDT, and DBK. Uniformity of distribution at the base and codon levels was assessed and compared to those of the same degenerate oligo sequences ordered from a lead commercial supplier and produced by conventional CDS (CDS oligos). The study confirmed the ability of EDS to generate all degenerate patterns tested in different sequence contexts, outperforming CDS oligos in terms of both base and codon distributions (Figures 3&4). These results highlight the accuracy of EDS technology in the incorporation of degenerate bases at multiple positions, regardless of the degeneracy type (mix of 2, 3, or 4 bases), and thus extend the utility of EDS technology to directed evolution applications.

MATERIALS AND METHODS

Oligo design and synthesis

Degenerate oligos used in this study were either ordered from a leading commercial supplier using conventional chemical DNA synthesis ('CDS oligos') or synthesized using our EDS technology with the commercially available Hi-Fidelity kit ('EDS oligos'). The 30 oligos comprised six different 52-mer sequences and a 13-nt barcode (Table 1) and were designed to incorporate equimolar amounts of mixed bases at three degenerate positions in the middle of their sequence, including five randomization patterns: either NNN (N= mix of A, C, G, and T), NNK (K= G and T), NNS (S= C and G), NDT (D= A, G and T) or DBK (B= C, G and T) .We chose to test these degenerate patterns as they are among the most frequently used ones to generate site-saturation libraries in protein engineering applications. Different barcodes were used to distinguish each degenerate pattern and synthesis method (EDS and CDS) during a demultiplexing step in DNA sequencing.

The six oligonucleotide sequences were designed to test the incorporation of mixed bases in different sequence contexts, including a variation of nucleotide/codon directly upstream and the presence of secondary structures (hairpin structures & G-quadruplexes).

Library Preparation and Next-Generation Sequencing

Barcoded EDS and CDS oligos were pooled together to generate an equimolar mix of all degenerate oligos. Sample preparation for Next Generation Sequencing (NGS) was performed using the xGenssDNA library prep kit (IDT) following the manufacturer's instructions. According to the manufacturer's instructions, the library was sequenced using an iSeq 100 benchtop sequencer (Illumina) at a 2x75 read length.

Bioinformatics pipeline for sequence analysis

Forward and reverse FASTQ files were processed through the following packages: FastQC for sequencing quality control, Trimmomatic to trim Illumina adapters, BBMerge to merge forward and reverse reads and Bowtie 2 to locally align the reads against the expected sequence and to generate BAM alignment files. Finally, the bam-read count was used for variant calling and assignment of degenerate bases, used in a custom Python script to determine

TABLE 1. LIST OF SEQUENCES WITH DEGENERATE CODONS USED IN THIS STUDY, SYNTHESIZED BY EDS TECHNOLOGY, AND SOURCED BY A COMMERCIAL VENDOR USING CONVENTIONAL CDS.

XXX represents the degenerate positions at which a mixture of multiple nucleotides was incorporated in equimolar amounts during synthesis. For each of the 6 sequences, 5 degenerate patterns were synthesized: either NNN, NNK, NNS, NDT, or DBK (total of 30 oligos).

Sequence 1	GTCTCTGCGGAGGAAGACACTTCGGCT**XXX**CGGAATCGACTATCAGGCGGGG[13nt-barcode]
Sequence 2	TGGCGCTGTTTCGCGTGACATTCTAAATAC**XXX**TGTGGCATCCGACTGGGGG[13nt-barcode]
Sequence 3	GTATGGCGCGATGACTCGCGCACGCTA**XXX**ATTCACTTGCTAAATATCACGG[13nt-barcode]
Sequence 4	GATTGCGGGCGATGGGTGAGGCTA**XXX**CAGCGATACCTTGTCAATGACTTGG[13nt-barcode]
Sequence 5	TAGTAGCAAGAGCACCTCCTTCTAG**XXX**TTACCCGACTGATAACCGCGACGG[13nt-barcode]
Sequence 6	TGATGGCGGTTCAGGCACACAGGG**XXX**CGTGCGGTTCCGCACAGGGCAATGG[13nt-barcode]

degenerate base distribution. For triplet (codon) distribution analysis at degenerate positions, the Python module Pysam was used. The number of sequencing cycles being higher than the length of the oligonucleotides, forward and reverse reads fully overlap in the oligos regions. The merging step thus enabled the filtering of reads containing sequencing errors by using the BBMerge parameter 'pfilter=1' (which prevents the merging in case of any mismatch between the forward and reverse reads in the overlap region).

RESULTS AND DISCUSSION

We evaluated the performance of EDS compared to conventional CDS in the incorporation of mixed bases at degenerate positions by synthesizing six DNA oligos harboring one of five degenerate patterns in the middle of their sequence: NNN, NNK, NNS, DBK, or NDT. We assessed the uniformity of distribution at (i) the base level and (ii) the codon level and compared them to theoretical distributions in a perfectly equimolar context.

Base distribution

We first assessed the uniformity of distribution at the base level by determining the average frequency of occurrence of each base (% of A, C, G, and T) at the degenerate positions (NNN, NNK, NNK, DBK, and NDT). Results are shown in Figure 3. As shown, the overall base distribution of EDS oligos at the three N degenerate positions within the NNN degenerate codon is very close to the theoretical equimolar mix of the 4 bases (expected frequency for A, C, G, and T of 25% each), outperforming CDS oligos for which the T base is over-represented at the expense of C. Furthermore, for both EDS and CDS oligos, the position of the N degenerate base (N_1, N_2, or N_3) does not have any significant effect on base distribution. Comparable results were obtained for degenerate patterns including a mix of two bases, for which the incorporation of K (K = G/T) or S (S= G/C) resulted in a highly balanced ratio of bases for EDS NNK and NNS oligos: the base distribution was close to the 50:50 expected ratio for K and S bases, and close to the 25:25:25:25 ratio for N bases (Figure 4A). CDS oligos, on the contrary, resulted in a significant bias towards the T base at the K position of NNK oligos, and in the same T bias at the N positions of NNK and NNS oligos. The same could be observed for D and B degenerate bases (mix of 3 bases), where the deviation from the target 33% frequency does not exceed 6% for EDS oligos, vs. 13% for CDS oligos (Figure 4A).

Summarizing these findings, Figure 4B presents the maximum deviation from theory in the percentage frequency of the 4 bases for each degenerate position, comparing EDS to CDS oligos. For EDS, all degenerate patterns result in a base distribution +/-6% away at its maximum from the theoretical equimolar distribution, vs. +/- 14% for CDS oligos.

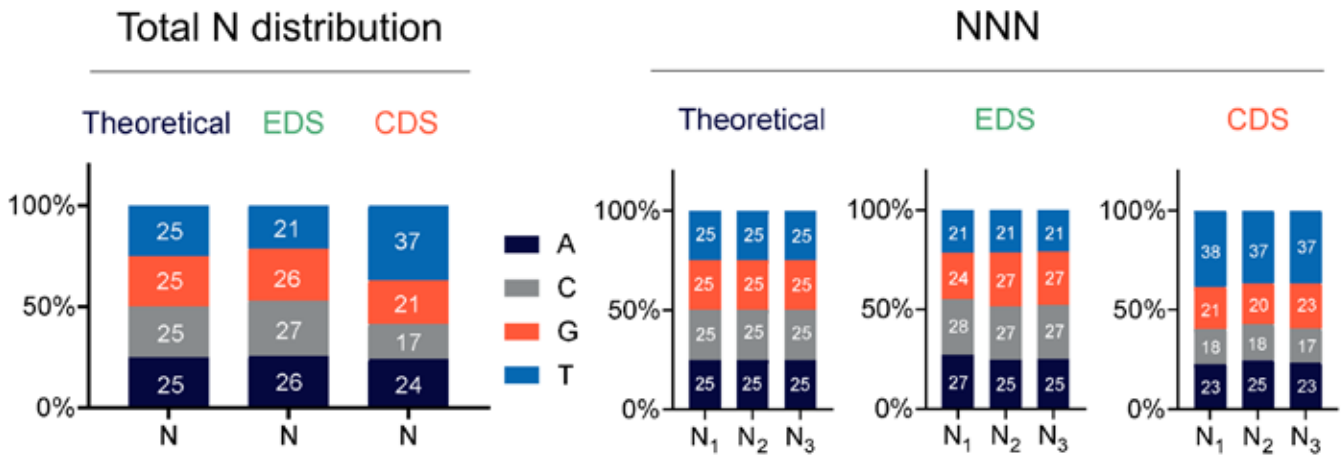
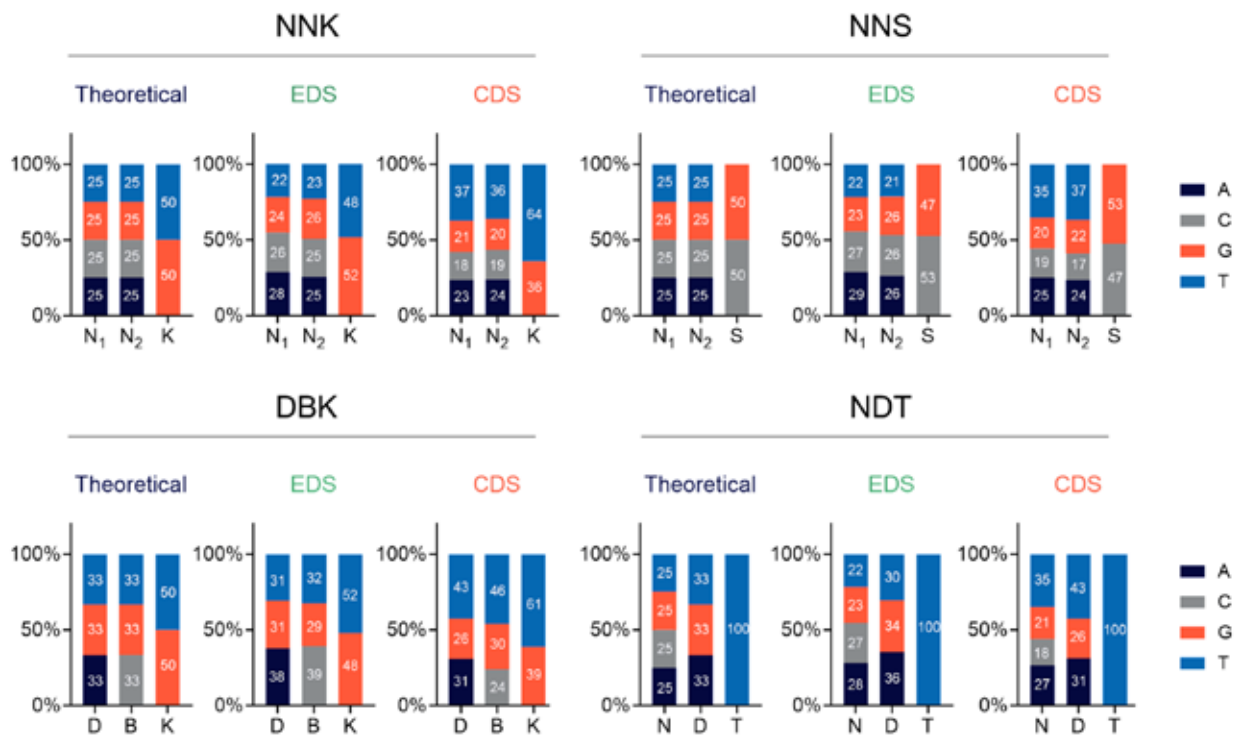


FIGURE 3. BASE DISTRIBUTION AT N DEGENERATE POSITIONS COMPARING EDS AND CDS OLIGOS

The percentage frequency of A, C, G, and T bases at the degenerate base position N (N= equimolar mix of A, C, G, and T) is written within each bar. EDS and CDS oligos are compared to the theoretical distribution (perfectly equimolar insertion of the 4 bases). The "Total N distribution" graph (left) represents the average base distribution across all individual N positions The NNN graph (right) gives the specific base distribution at each position (N_1, N_2, N_3) of the NNN degenerate codon.

A



B

	N				K		S		D			B		
	A	C	G	T	G	T	C	G	A	G	T	C	G	T
Expected frequency	25%	25%	25%	25%	50%	50%	50%	50%	33%	33%	33%	33%	33%	33%
EDS oligos	4%	3%	-2%	-4%	2%	-2%	3%	-3%	5%	-2%	-2%	6%	-4%	-1%
CDS oligos	-2%	-7%	-5%	12%	-14%	14%	-3%	3%	-2%	-7%	10%	-9%	-3%	13%

FIGURE 4. BASE DISTRIBUTION IN DIFFERENT DEGENERATE CODONS COMPARING EDS AND CDS OLIGOS

5 degenerate patterns were synthesized within 6 different 52-mer oligos: NNN (Figure 2), NNK, NNS, DBK, and NDT. **A:** Percentage frequency of A, C, G, and T bases at the degenerate base position N, K, S, D, or B. EDS and CDS oligos are compared to the theoretical distribution (perfectly equimolar insertion of 2-4 bases depending on the degenerate base). **B:** Table summarizing the maximum percent deviation from theoretical base distribution for each degenerate base, based on results shown in A. The expected frequencies are shown for reference, and the largest difference (either positive or negative) from this theoretical frequency is indicated for each nucleotide at the degenerate position, for both EDS and CDS oligos (color code: green to red= from smallest to largest difference in percent frequency).

Codon diversity and distribution

Tightly linked to the base distributions described above, diversity and distribution obtained at the codon level are additional indicators of degenerate incorporation efficiency. This is particularly significant for applications such as site saturation variant libraries in the directed evolution of proteins, as an equimolar codon distribution will result in the maximum diversity and uniform distribution of AAs. Figure 5 summarizes the codon diversity and distribution for the five degenerate patterns tested, comparing EDS to CDS oligos.

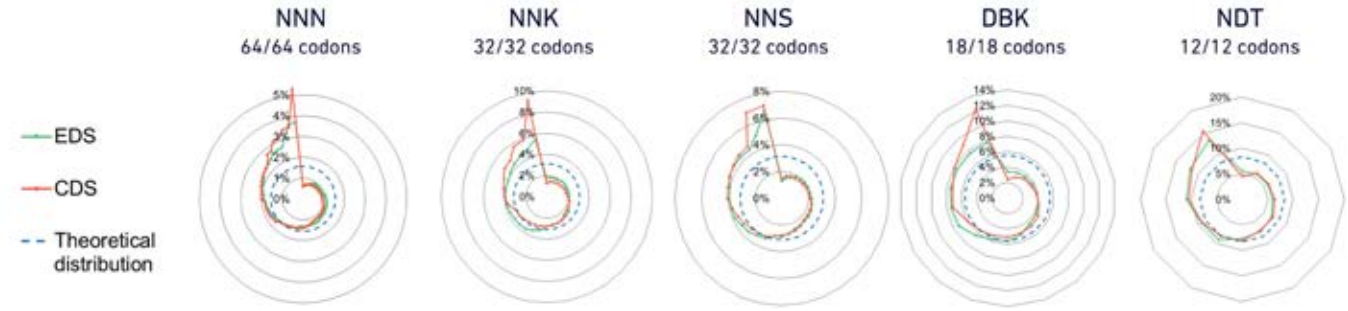
All expected codon types were obtained across all five tested degenerate patterns (NNN: 64 expected codon possibilities, NNK: 32, NNS: 32, DBK: 18, NDT: 12), using both EDS and CDS oligonucleotides. For all five degenerate patterns, the distribution of codon frequencies closely matches the theoretical distribution in an equimolar context (Figure 5A), with narrower ranges of frequency for EDS oligos compared to CDS oligos (Figure 5A&C), indicating a better uniformity of distribution between codons in EDS degenerate oligos. This can also be observed in Figure 5B, where the interquartile range and whiskers are narrower in all five degenerate patterns for EDS, centered around the theoretical distribution. While for EDS, codons combining C and G in their first two bases seem to be slightly over-represented at the expense of codons combining T and G bases, for CDS oligos codons harboring at least two T nucleotides are predominant in all codon degeneracies, resulting

in a more uneven representation of each codon (Figure 5B).

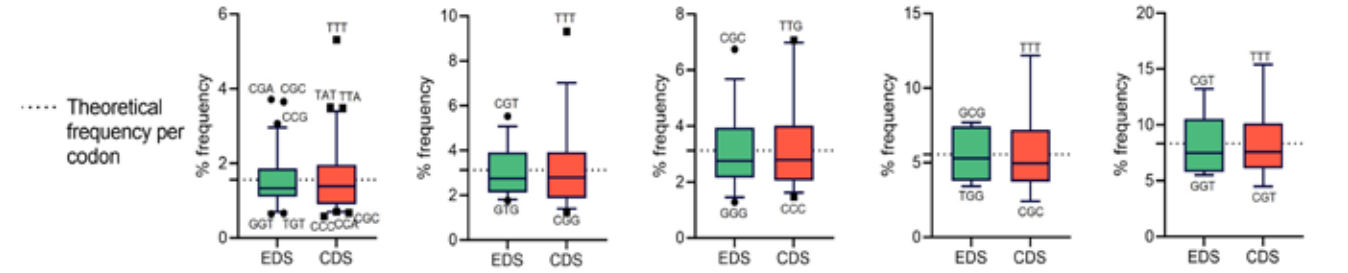
Taken together, both EDS and CDS syntheses of degenerate sequences were able to generate all expected codons, with occurrence frequencies not deviating by more than 4.9% from the expected frequency for EDS, vs. 7.1% for CDS oligos (Figure 5C). In terms of frequency range from the most to least represented codon, taking the widely used NNK codon as an example, a 3.8% range (EDS) compared to an 8.1% range (CDS) may initially seem like a minor difference but may have significant repercussions for variant library screening and colony picking in mutagenesis experiments. Indeed, it implies that the least represented codon will have double the chance of being observed with EDS more balanced distribution compared to CDS, and this codon bias will increase along with the number of degenerate codons incorporated.

We thus demonstrated at both the nucleotide and the codon level the successful insertion of degeneracies in several 52-mers sequences using EDS technology, testing five different patterns (NNN, NNK, NNS, DBK, and NDT) and equimolar mixes of 2, 3, or 4 different bases. The minimal deviations from theoretical frequencies at the base level result in a highly uniform and even distribution at the codon level, suggesting that EDS technology could generate saturation variant libraries at the expected diversity, requiring similar subsampling compared to CDS degenerate oligos.

A



B



C

		NNN	NNK	NNS	DBK	NDT
Expected frequency		1.56%	3.13%	3.13%	5.56%	8.33%
Range	EDS	3.07%	3.77%	5.46%	4.27%	7.72%
	CDS	4.73%	8.06%	5.6%	9.76%	10.9%
Maximum % deviation from expected frequency	EDS	2.15%	2.40%	3.61%	-2.14%	4.89%
	CDS	3.75%	6.18%	3.94%	6.62%	7.06%

FIGURE 5. CODON OCCURRENCE AND DISTRIBUTION FOR FIVE DEGENERATE PATTERNS

A. Radar plots showing the frequency distribution of codons at the degenerate position, from the least represented codon (center of the plot) to the most represented (towards outer circles). Dots represent individual codons, superimposing EDS oligos (orange) with CDS oligos (green). The theoretical codon frequencies are represented by a dashed blue circle. The number of codons (obtained and theoretical) per degenerate pattern is indicated below each pattern. **B.** Boxplots representing individual codon frequencies for each pattern compared to theory (black dashed line). Boxes show the median and 25th and 75th percentile; whiskers extend to the 5th and 95th percentiles. Outliers are represented as black dots, and most and least represented codons are indicated. **C.** Table summarizing the maximum percent deviation from theoretical codon frequency for each degenerate pattern, based on results in B. The expected frequencies per codon are shown for reference, and the largest difference (either positive or negative) from this theoretical frequency is indicated for each codon (color code: green to red= from smallest to largest difference in percent frequency). The frequency range is also indicated, showing narrower windows for EDS compared to CDS syntheses.

CONCLUSION

Synthesis of oligos harboring degenerate positions provides a straightforward and rapid tool to generate genetic diversity at specifically targeted sites or residues, with numerous applications in the biotechnological and pharmaceutical fields, such as the directed evolution of enzymes in protein engineering studies, or in sequencing applications for biomedical research and diagnostics purposes. The data presented here demonstrate the capability of EDS technology to rapidly generate high-quality degenerate oligonucleotides. We show that the five tested degenerate codons (NNN, NNK, NNS, NDT, and DBK) are successfully incorporated in EDS oligos of various sequence contexts. The uniformity of codon and base distribution matched (and sometimes outperformed) those obtained with commercially sourced oligos produced by CDS. This highlights the capability of EDS technology to generate a variety of degenerate sequences with high accuracy and low bias in base incorporation, close to the theoretical distribution.

Providing a significant advantage over centralized CDS oligo production and third-party logistics, our EDS technology, therefore, offers the unique ability to conduct iterative cycles of degenerate oligo syntheses rapidly and easily. This represents an opportunity to accelerate screening processes in enzyme-directed evolution studies by synthesizing a wide range of degenerate sequences, with multiple synthesis replicates, on a single 96-well plate. With much shorter turnaround times from degenerate oligo design to synthesis, the speed of identification of optimized mutants through site-saturation variant libraries using degenerate oligos can be greatly enhanced. In this perspective, future work will involve testing EDS degenerate oligos in a protein engineering application by generating site-saturation variant libraries.

REFERENCES

1. Miyazaki, Arnold et al 1999, J Mol Evol, doi : 10.1007/pl00006593
2. Turner 2009, Nat Chem Biol, doi :10.1038/nchembio.203
3. Reetz 2011, Angew. Chem, doi : https://doi.org/10.1002/anie.201000826
4. Alfaro-Chávez, Liu et al 2019, Protein Eng Des Sel, doi : https://doi.org/10.1016/S1367-5931(98)80044-6
5. Matsumoto, Yamada et al 2019, World J Microbiol Biotechnol, doi : https://doi.org/10.1007/s11274-019-2777-8
6. Belsare, Andorfer et al 2017, ACS Synth Biol, doi : https://doi.org/10.1021/acssynbio.6b00297
7. Kille, Acevedo-Rocha et al 2012, ACS Synth Biol, doi : 10.1021/sb300037w
8. Tang, Gao et al 2012, BiotTechniques 2012, doi : 10.2144/000113820.
9. Reetz, Kahakeaw et al 2008, doi : 10.1002/cbic.200800298
10. DNA Script. Enzymatic DNA Synthesis Technical Note (TN900100) : https://www.dnascript.com/wp-content/uploads/2021/04/DNAScript_EDS_3Pillars_TN900100.pdf

Technical Support
support@DNAScript.com
USA & Canada: +1 888-693-4022
Europe: +33 1 84 88 02 42