Solid-Phase DNA Synthesis Technology Allows Tight Control of Combinatorial Library Quality and Diversity

INTRODUCTION

Combinatorial DNA libraries are essential to modern protein and metabolic engineering research. As modular sets of DNA sequences, they can be arranged and rearranged as needed to achieve novel or improved protein properties. Modern, high-throughput DNA synthesis technologies have lowered the cost and improved the throughput of library synthesis, making them more accessible to both academic and biotech scientists. The utility of combinatorial libraries for discovering desired protein variants, however, depends on library quality and complexity.

In this article, we present the most commonly used approaches and technologies available for combinatorial library synthesis and explain their associated codon bias, amino acid distribution, error rate, yield, library diversity, and quality. We demonstrate that Twist Bioscience's proprietary silicon-based solid-phase technology stands apart with its impressive ability to offer high throughput and low cost while enabling tight control of library quality and diversity.

	DENEGERATE SYNTHESIS				TRIM	SOLID-PHASE SYNTHESIS
	FULL DEGENERACY	PARTIAL DEGENERACY				
NUMBER OF:	NNN	NNT/C	NNG/T	NNT/C/G		
CODONS	64	32	32	48	20	64
AMINO ACIDS	20	15	20	20	20	0
STOP CODONS	3	0	1	1	0	0

Table 1. Comparison of oligonucleotide-based combinatorial variant library generation methods. NNN = full degeneracy at all three nucleotides; NNT/C, NNTG/T and NNT/C/G = partial degeneracy at the third position (T or C; G or T; or T, C or G, respectively). Note that of these three methods, only Twist solid-phase synthesis allows access to the full complement of codons without the introduction of unwanted stop codons.

OLIGONUCLEOTIDE-BASED COMBINATORIAL LIBRARY SYNTHESIS METHODS

The three most commonly used approaches and technologies for combinatorial library synthesis (degenerate synthesis, trimer phosphoramidite [TRIM], and Twist solid-phase synthesis) are summarized in **Table 1** and described in more detail above.

Degenerate Synthesis

The most common approach for building a combinatorial library involves the use of degenerate oligonucleotides to create desired mutations at specified positions in the protein. Though this approach has the lowest upfront cost, it is prone to codon bias and introduces unwanted stop codons (9% of mutated codons are nonsense codons when using full degeneracy).

Two methods of degenerate synthesis favor amino acids coded by more than one codon (leucine, arginine, serine), which represent up to 28% of the codons: full degeneracy (any nucleotide can be incorporated at all three positions of the codon and NNN) and partial degeneracy (incorporation limited to specific nucleotides at third position: T or C; G or T; or T, C, or G). Though partial degeneracy at the third position of the codon can mitigate codon bias and reduce the presence of nonsense codons, it limits the number of possible codons and so may hinder efforts to optimize codons or modify sequences to avoid unwanted restriction sites or higher-level motifs (**Table 1**).

Trimer Phosphoramidite (TRIM) Technology

The use of trimer phosphoramidites (also known as TRIM technology) circumvents the problems encountered with degenerate synthesis. Instead of using single bases, TRIM uses premade trimers representing the codons for all 20 amino acids to synthesize oligonucleotides. In theory, this technique provides control over all codons used to generate the library and prevents the introduction of nonsense codons. In practice, however, researchers have no control over the codon used because only 20 are available, and they encounter difficulties when trying to introduce mutations in multiple distal regions.

Using TRIM to synthesize long oligonucleotides leads to lower sequence fidelity due to deletions, depurination events, and mutations arising from deamination of cytidine (Leproust et al. 2010). This problem can be alleviated to some extent by using antisense trimer phosphoramidites. The reverse complement of canonical sense codons, these trimers can be incorporated in the opposite strand of the mutant gene during gene assembly. The various trimers in the coupling reaction also have innate differences in reactivity (Randolph et al. 2008). To get around this, the concentration (volume) of each trimer must be adjusted according to a known reaction factor such that each trimer has the same chance of being incorporated in the extending chain. As a result, the accuracy of composition depends on the accuracy of liquid handling and is impacted by the sequence-dependent over- and under-incorporation bias, as well as length limitations. Inaccurate codon composition is a significant problem with TRIM if library design includes fine ratio control or length variation.

Twist Solid-Phase DNA Synthesis

Another approach to combinatorial mutagenesis is to synthesize all of the mutant sequences to an exact specification. Though this approach would be prohibitively time-consuming and cost-prohibitive with traditional column-based DNA synthesis, modern array-based and massively parallel oligonucleotide synthesis methods (Kosuri et al. 2010; Leproust et al. 2010) have made it an increasingly practical solution.

One such method is Twist Bioscience's high-throughput solid-phase DNA synthesis platform, which overcomes the codon bias observed with degenerate oligonucleotide approaches, but without the limitations of the TRIM technology (Li et al. 2018a; Hoebenreich et al. 2015; Reetz 2016). On a silicon chip with approximately the same footprint as a 96-well plate, Twist technology can precisely synthesize more than a million oligonucleotides in a single run. It is thus possible to design and synthesize all of the oligonucleotides needed to generate a complete combinatorial library without making compromises on the codons used and without extending project timelines. In addition, the Twist synthesis workflow involves a screen of all variants before the combinatorial library is assembled to remove the variants that contain unwanted motifs or restriction sites (**Figure 1**).

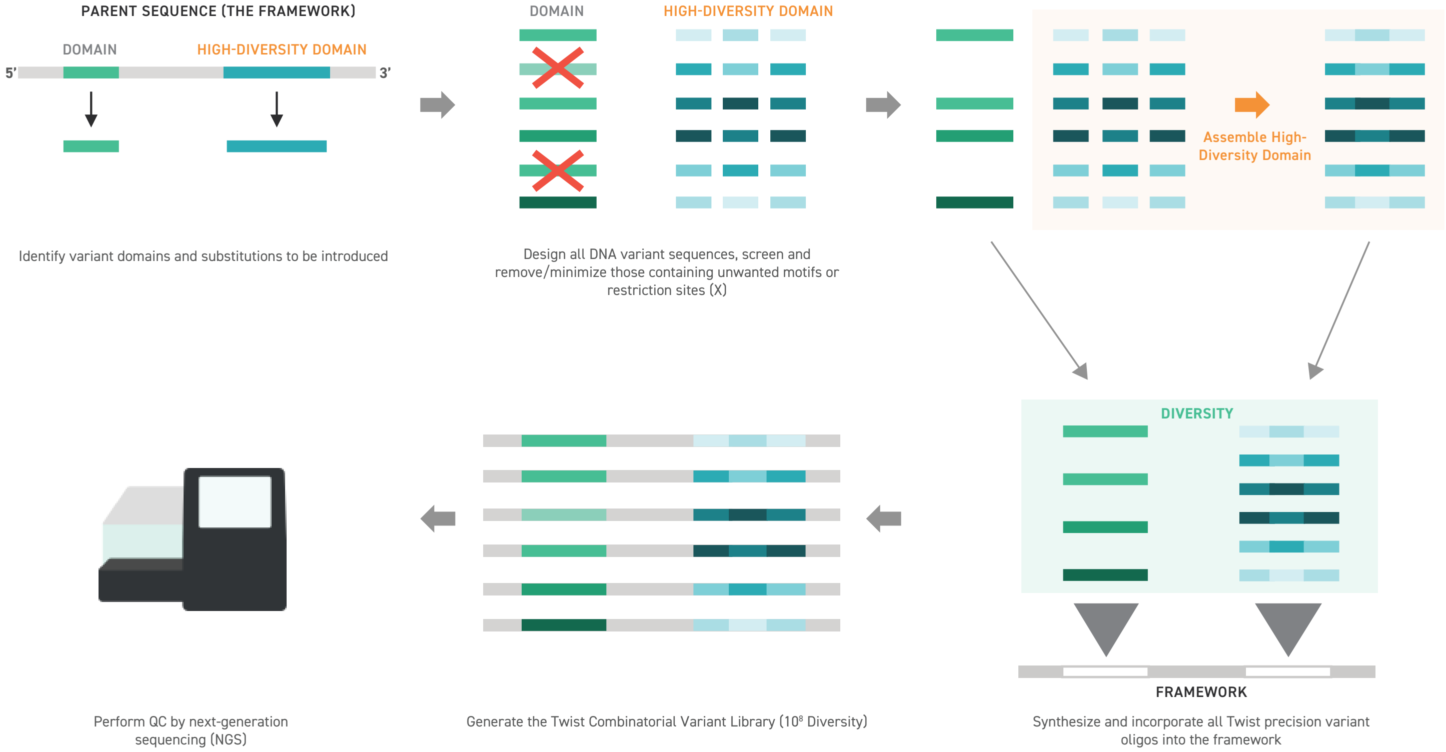
PARENT SEQUENCE (THE FRAMEWORK)
DOMAIN
HIGH-DIVERSITY DOMAIN
5'
3'
Identify variant domains and substitutions to be introduced
DOMAIN
HIGH-DIVERSITY DOMAIN
Design all DNA variant sequences, screen and remove/minimize those containing unwanted motifs or restriction sites (X)
Assemble High-Diversity Domain
DIVERSITY
FRAMEWORK
Synthesize and incorporate all Twist precision variant oligos into the framework
Generate the Twist Combinatorial Variant Library (10⁸ Diversity)
Perform QC by next-generation sequencing (NGS)

Figure 1. Combinatorial library design and synthesis using the Twist Bioscience's solid-phase DNA synthesis platform. Sequence domains and amino acids in a protein framework are identified for substitution, and all potential genetic variants within those domains are designed. Next, the variants are screened in silico and those containing unwanted motifs or restriction sites are removed. The remaining variants are then synthesized and incorporated into the framework to generate the Twist combinatorial library, which is analyzed by NGS in a final quality control step.

FACTORS AFFECTING LIBRARY QUALITY

Amino Acid Frequency

A critical step in building a combinatorial library is designing a mutation scheme that maximizes the chances of identifying protein variants with desired properties. The scheme can call for either complete randomness or focused diversity based on structural knowledge, but the library should approximate the intended design as accurately as possible. For example, if a strategy calls for testing all 20 amino acids at one position in a peptide sequence, each amino acid is expected to be present at that position in 5% (1/20) of the mutant molecules in the library. Codon bias has a strong influence on amino acid frequency, with higher bias leading to more significant deviations from expected frequencies. For this reason, libraries prepared using degenerate synthesis tend to require screens of larger numbers of variants than libraries prepared using other methods, and this adds both time and cost to a project.

One assessment of library quality involves comparing the expected and observed amino acid frequencies. This assessment is best performed by next-generation sequencing (NGS) of the library to analyze the diversity at highly randomized positions, where each amino acid is present at a relatively low frequency that is more difficult to reach with precision. **Figure 2** shows an example of codon frequencies in a fully randomized region of seven codons built using Twist's solid-phase DNA synthesis (and analyzed by NGS). In this analysis, cysteine was not included, so 19 amino acids were analyzed at each position (expected frequency of 5.3%). The observed frequency is within 25% of the expected frequency at each position, a result that is common to Twist combinatorial libraries but not to a TRIM combinatorial library (**Figure 3**). These results reflect the fact that the Twist solid-phase DNA platform synthesizes ("writes") libraries according to precise sequence specifications.

A.

AMINO ACID		POSITION 1	POSITION 2	POSITION 3	POSITION 4	POSITION 5	POSITION 6	POSITION 7
	A	5.84	5.58	5.07	5.59	5.35	5.35	5.57
	C	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	D	4.92	4.77	4.48	4.62	4.59	4.59	5.01
	E	4.75	5.13	4.91	5.10	4.99	4.99	4.73
	F	5.75	5.23	5.39	5.56	5.20	5.20	4.93
	G	4.55	5.20	4.47	5.22	4.61	4.61	4.67
	H	5.15	5.16	5.85	5.02	5.98	5.98	6.00
	I	5.33	4.45	4.58	4.91	4.43	4.43	4.79
	K	5.01	4.72	4.66	4.64	4.60	4.60	4.46
	L	5.45	5.67	6.28	5.67	6.14	6.14	5.77
	M	5.41	5.44	4.95	5.22	4.88	4.88	4.71
	N	5.12	4.65	4.81	4.37	4.84	4.84	4.75
	P	5.23	6.00	5.73	6.04	5.87	5.87	6.34
	Q	4.69	4.58	5.20	4.90	5.16	5.16	5.72
	R	5.28	6.25	6.52	6.07	6.44	6.44	6.07
	S	5.11	4.70	4.77	4.92	4.76	4.76	4.90
	T	5.55	5.57	5.47	5.43	5.53	5.53	5.57
	V	5.33	4.98	4.91	5.03	5.23	5.23	4.91
	W	5.32	5.85	5.82	5.63	5.49	5.49	5.25
	Y	5.29	5.32	5.20	5.28	4.98	4.98	4.89

🟩 0–15%, ▢ 15–20%, □ 20–25%, 🟥 25%+

B.

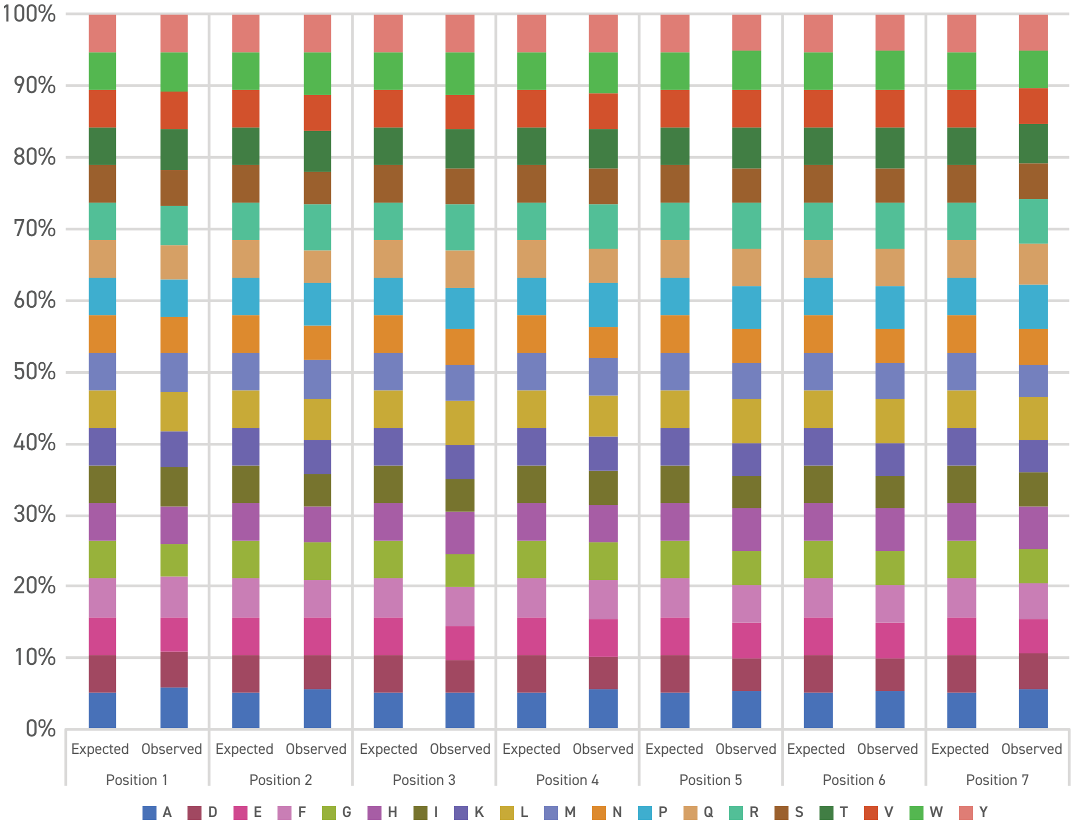

Figure 2. Amino acid frequency (%) at seven sites of a mutagenized region synthesized using Twist solid-phase technology. Variants in seven sequential amino acid positions were generated with 19 amino acid residues (cysteine was omitted) in the first seven sites. **A**, tabulated frequency data obtained from NGS, with the shade of green indicating deviation from the expected value. All expected variants were present at all positions and their observed frequency was within 25% of the expected value (specification) at 5.3%. **B**, the same data plotted next to expected frequencies to illustrate the tight control of amino acid frequency at each position.

AMINO ACID DISTRIBUTION: OBSERVED VS. EXPECTED

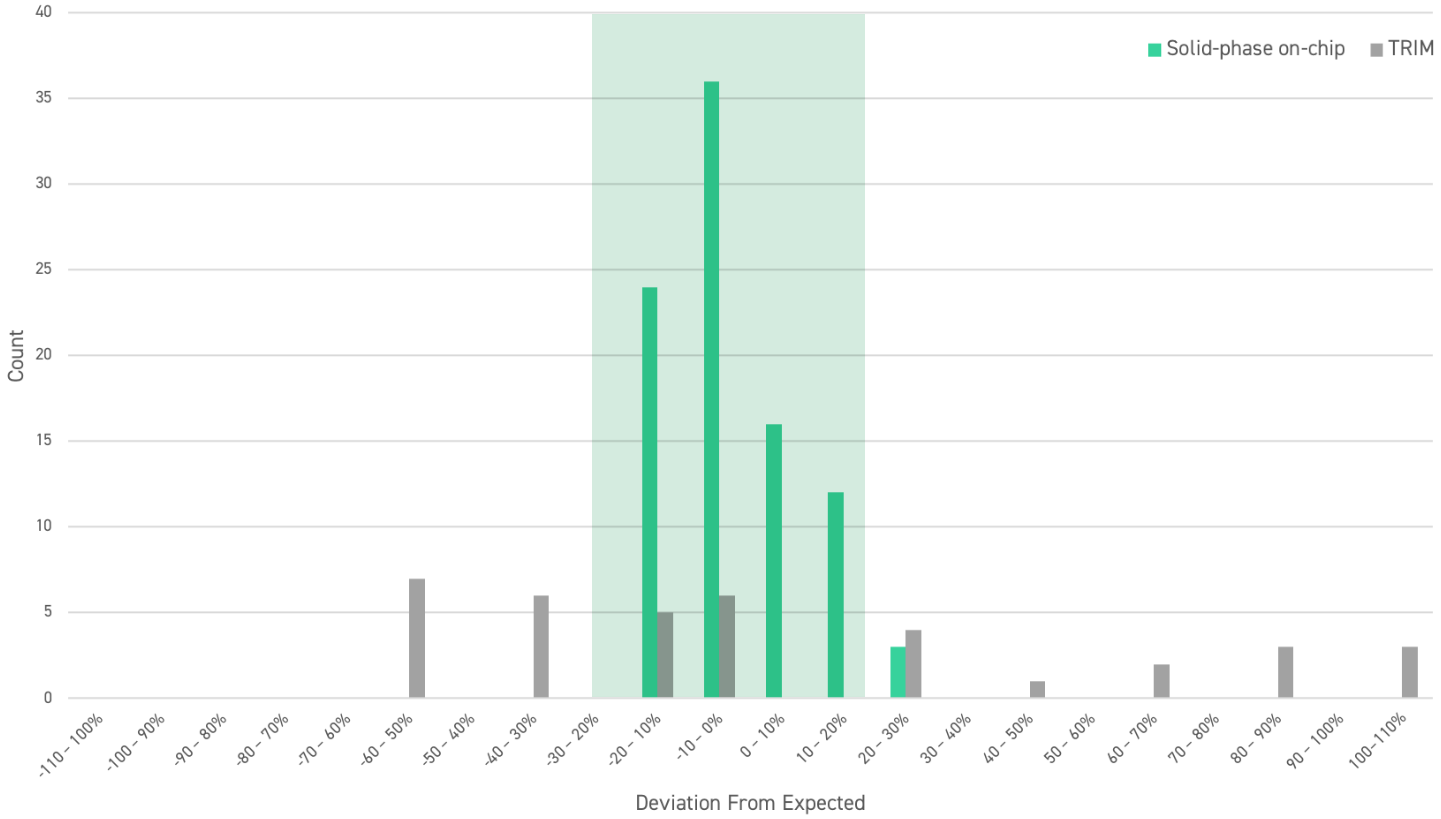

Figure 3. Observed vs expected amino acid frequency: comparison of TRIM and Twist solid-phase DNA synthesis. Whereas more than 96% of the observed values were within 25% of expected values (specification, light green shaded area) when using Twist's solid-phase DNA synthesis platform (green bars), only ~30% fell within this range when using TRIM technology (gray bars).

Repetitive Yield

The coupling reactions used in DNA synthesis are not 100% efficient, and the extent of inefficiency varies between the technologies used. This is another factor to consider when constructing libraries because inefficiencies in coupling not only affect yields, but they can also create truncations in the oligos and thus generate unintended frameshifts in the final library.

Repetitive yield (RY) is the measure of the efficiency of oligonucleotide coupling reactions at each cycle. It determines the final yield (FY) of the product according to the relationship:

> $FY= RY^{n-1}$, where n is the number of cycles performed (the first nucleotide is not counted)

As shown in **Figure 4**, even a small decrease in coupling efficiency can significantly decrease the final yield at the end of the synthesis cycles, affecting the quality of the library.

Twist's solid-phase synthesis technology can achieve an RY of 99.6%, far superior to the reported 90–98.3% RY achieved with TRIM (Kayushin et al., 1996; Vargeese et al., 1998). Though TRIM technology requires only a third of the cycles required by Twist's technology to synthesize a given oligonucleotide, TRIM is nevertheless less efficient due to the size of the triplet coupled at each round of synthesis.

EFFECT OF REPETITIVE YIELD ON FINAL OLIGONUCLEOTIDE YIELD

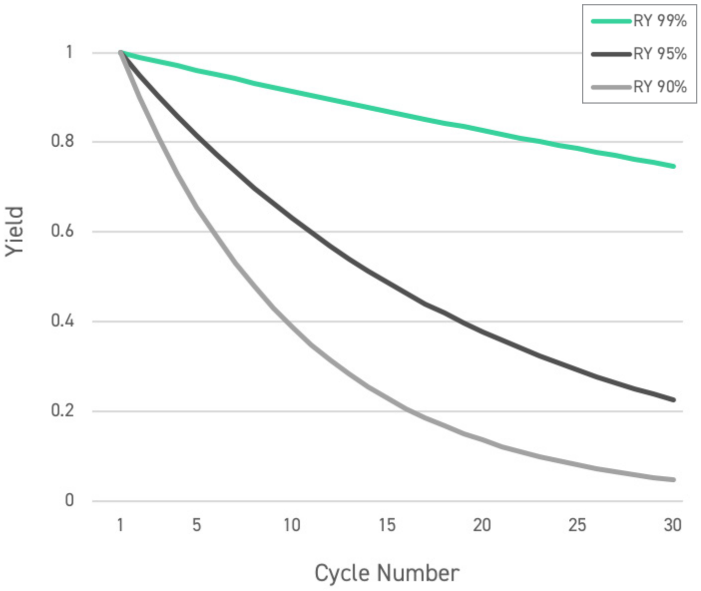

Figure 4. Effect of repetitive yield (RY) on final yield (FY). The final yield of coupling reaction cycles is significantly affected by repetitive yield.

Sequence Errors

Oligonucleotide synthesis can also introduce errors in the sequences that affect library quality. Due to its highly efficient in situ chemistry, however, Twist's silicon-based solid-phase on-chip gene synthesis technology offers industry-leading low error rates of around 1/1000 or even lower. Such low error rates in the oligonucleotide synthesis step help generate high-quality libraries that increase the chances of finding expected mutants.

ASSESSMENT OF LIBRARY DIVERSITY

Library diversity, the relative frequency (representation) of the different intended variants, directly impacts the screening effort needed to find desired mutants: libraries with higher diversity require less screening. Though theoretical library diversity can be easily estimated from the number of sites mutated and the number of amino acids tested at each site, this theoretical value is rarely obtained in practice. Errors that occur during oligonucleotide synthesis, gene assembly, and PCR amplification introduce codon bias, spurious nonsense codons, sequence errors, and incomplete sequences (often referred to as infidelity), as well as unintended variants, which all impact library diversity.

The only reliable way to assess library diversity is by sequencing. When Sanger sequencing is used, however, cost constraints generally limit analysis to only a subset of 96 clones from a library. Using NGS, on the other hand, allows analysis of the entire library to provide a much larger data set (106 reads or more). This allows derivation of statistically significant measurements of library diversity, and NGS also provides information on other factors that affect diversity — sequence duplications, incorrect sequences (sequence errors, deletions), the presence of wild-type/parental sequences, non-designed codons, and yield—with greater accuracy than Sanger sequencing of a small subset (Li et al. 2018a, 2018b). Furthermore, the quality of the data can be improved by adjusting the oversampling factor (Li et al. 2018a, 2018b).

Combinatorial library synthesis using the Twist solid-phase platform includes full NGS analysis as an important quality control step (**Figure 1**). This ensures all errors are detected before the library is used in downstream analysis and allows researchers to know precisely which variants are in their library. Any negative results, then, can be attributed directly to variants, and those variants can be omitted from subsequent library designs to improve efficiency.

CONCLUSION

In choosing an approach for constructing oligonucleotide-based combinatorial libraries, a range of factors impact library quality and diversity to influence the overall time and cost associated with protein engineering projects. We have described a number of those factors and have demonstrated why Twist Combinatorial Libraries are the ideal choice when tightly controlled diversity is required.

Twist's silicon-based solid-phase DNA synthesis technology allows the effective design and meticulous synthesis of each variant in a combinatorial library — base by base in a high-throughput manner, with a low error rate of 1/1000 or less, and at a low cost per base. This allows complete control over which codons and which combinations of codons can be incorporated (single, double, triple substitutions etc.), as well as the flexibility to explore variations in length. Moreover, Twist screens all variants for unwanted motifs and restriction sites before they are synthesized. Putting all of this together, Twist offers high-diversity libraries without unwanted bias or loss of complexity and can confidently supply libraries requiring even the most precise synthesis.

Finally, inasmuch as library diversity determines the screening effort needed to find desired mutants, obtaining the most accurate assessment of library quality and diversity is of critical importance. Using NGS — as is done in the Twist synthesis pipeline — offers the most reliable quality analysis. NGS generates useful insights into true library diversity and allows more efficient planning of screening experiments. Researchers using Twist combinatorial libraries, therefore, know exactly which variants are in their libraries and can apply this information to interpreting their data and designing subsequent libraries. These critical steps enable focused functional screens that save time and money and amplify the efficiency of protein engineering projects.

REFERENCES

Hoebenreich S, Zilly FE, Acevedo-Rocha CG, Zilly M, Reetz MT (2015) Speeding up directed evolution: combining the advantages of solid-phase combinatorial gene synthesis with statistically guided reduction of screening effort. ACS Synth Biol 4: 317–331

Kayushin AL, Korosteleva MD, Miroshnikov AI, Kosch W, Zubov D, Kosch W, Zubov D, Piel N (1996) A convenient approach to the synthesis of trinucleotide phosphoramidites—synthons for the generation of oligonucleotide/peptide libraries. Nucleic Acids Res 24: 3748–3755

Kosuri S, Eroshenko N, Leproust EM, Super M, Way J, Li JB, Church GM (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat Biotechnol 28: 1295

Leproust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, Caruthers MH (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Res 38: 2522–2540

Li A, Acevedo-Rocha Carlos G, Sun Z, Cox T, Xu Jia L, Reetz MT (2018a) Beating bias in the directed evolution of proteins: combining high-fidelity on-chip solid-phase gene synthesis with efficient gene assembly for combinatorial library construction. ChemBioChem 19: 221–228

Li A, Sun Z, Reetz MT (2018b) Solid-phase gene synthesis for mutant library construction: the future of directed evolution? Chem BioChem19: 2023–2032

Randolph J, Yagodkin A, Lamaitre M, Azhayev A, Mackie H (2008) Codon based mutagenesis using trimer phosphoramidites. Nucleic Acids Symp Ser 52: 479–479

Reetz MT (2016) Directed evolution of selective enzymes: catalysts for organic chemistry and biotechnology (Weinheim: Wiley-VCH)

Vargeese C, Carter J, Yegge J, Krivjansky S, Settle A, Kropp E, Peterson K, Pieken W (1998) Efficient activation of nucleoside phosphoramidites with 4,5-dicyanoimidazole during oligonucleotide synthesis. Nucleic Acids Res 26: 1046–1050

These Products are subject to certain use restrictions as set forth in Twist's Supply Terms and Conditions
www.twistbioscience.com/supply-terms-and-conditions