Product Note

gSynth™

A highly accurate, enzymatic, *de novo* synthesis and gene assembly technology

Neil M. Bell, Sylwia A. Mankowska, Steven A. Harvey, Derek L. Stemple and Andrew Fraser

- **gSynth™ 300mers yielded, on average, 85.3% full-length sequences**
- **Phosphoramidite 300mers yielded, on average, 22.7% full-length sequences**

INTRODUCTION:

The ability to produce synthetic DNA is essential for the investigation and engineering of biological pathways. For many years phosphoramidite synthesis has been the gold-standard technology for synthetic DNA production, however, it is well known that over long stretches of nucleotides this method is error-prone[1]. With the best nucleotide coupling efficiencies, when phosphoramidite synthesis reaches 200 nucleotides in length, only ~40% of the material is the correct full-length product. These synthesis errors have a dramatic effect on downstream results and, consequently, these technical limitations are holding back many synthetic biology applications.

Recently there has been a drive to create novel enzymatic DNA synthesis technologies but, so far, the developed technologies are far from being commercially viable and have synthesis

300mer Synthesised	gSynth™			Phosphonamidite HAE		
	[a]Alignment (%)	[b]Full-length Reads (%)	[c]Coupling Efficiency (%)	Alignment (%)	Full-length Reads (%)	Coupling Efficiency (%)
SeqG1 & SeqP1 GC content ranging from 40 to 60 %GC	**99.6**	**82.9**	**>99.9**	**97.8**	**28.3**	**99.6**
SeqG2 & SeqP2 Two, 10-nucleotide homopolymer regions.	**99.8**	**89.2**	**>99.9**	**94.4**	**12.5**	**99.3**
SeqG3 & SeqP3 six variable nucleotides N_1 to N_6	**99.6**	**83.8**	**>99.9**	**97.0**	**27.5**	**99.6**
Overall	**99.7 ± 0.1**	**85.3 ± 3.4**	**>99.9**	**96.4 ± 1.7**	**22.7 ± 8.9**	**99.5 ± 0.17**

Table 1. Analysis from 100,000 quality trimmed (Trim_Galore) paired-end reads for each sample SeqG1, SeqG2, SeqG3, SeqP1, SeqP2 & SeqP3. (a) Percent of concordantly aligned sequences, from 100,000 quality trimmed paired-end reads (Bowtie2). (b) Percent of full-length concordantly aligned sequences, 100,000 quality trimmed paired-end reads. (c) Equivalent nucleotide coupling efficiency based on yield of full-length sequences (yield = coupling efficiency^(length-1)).

error-rates higher than phosphoramidite synthesis[2]. To unlock novel synthetic biology applications there is the need for a DNA synthesis technology that is accurate, fast and able to produce complex DNA sequences.

Camena Bioscience has developed a novel enzymatic *de novo* synthesis and gene assembly technology called gSynth™. To highlight the advantages of gSynth™ we have performed benchmarking experiments comparing a series of 300mer fragments produced with gSynth™ and phosphoramidite synthesis. Our results demonstrate that gSynth™ is far more accurate at producing DNA sequences.

MATERIALS & METHODS:

During the benchmarking experiments we synthesised a series of 300mer sequences with gSynth™ and phosphoramidite synthesis. Specifically, these 300mers included sequences that ranged from 40-60% GC (SeqG1 and SeqP1), contained six variable nucleotides N1 to N6 (SeqG2 and SeqP2), or contained T and C homopolymers (SeqG3 and SeqP3).

All synthesised 300mers were subject to next-generation sequencing (2 x 151 bp), providing a detailed analysis of the different technologies. Sequences SeqG1, SeqG2 and SeqG3 were synthesised using gSynth™, SeqP1, SeqP2 and SeqP3 were synthesised as two standard desalted 162 nucleotide phosphoramidite oligos. The two 162 nucleotide oligos were hybridised at a complementary 24 nucleotide region at their 3' ends and then extended with the high-fidelity Q5 DNA polymerase to generate double-stranded DNA (referred to as phosphoramidite HAE). While this approach permits analysis of synthesis quality it is also comparable to polymerase

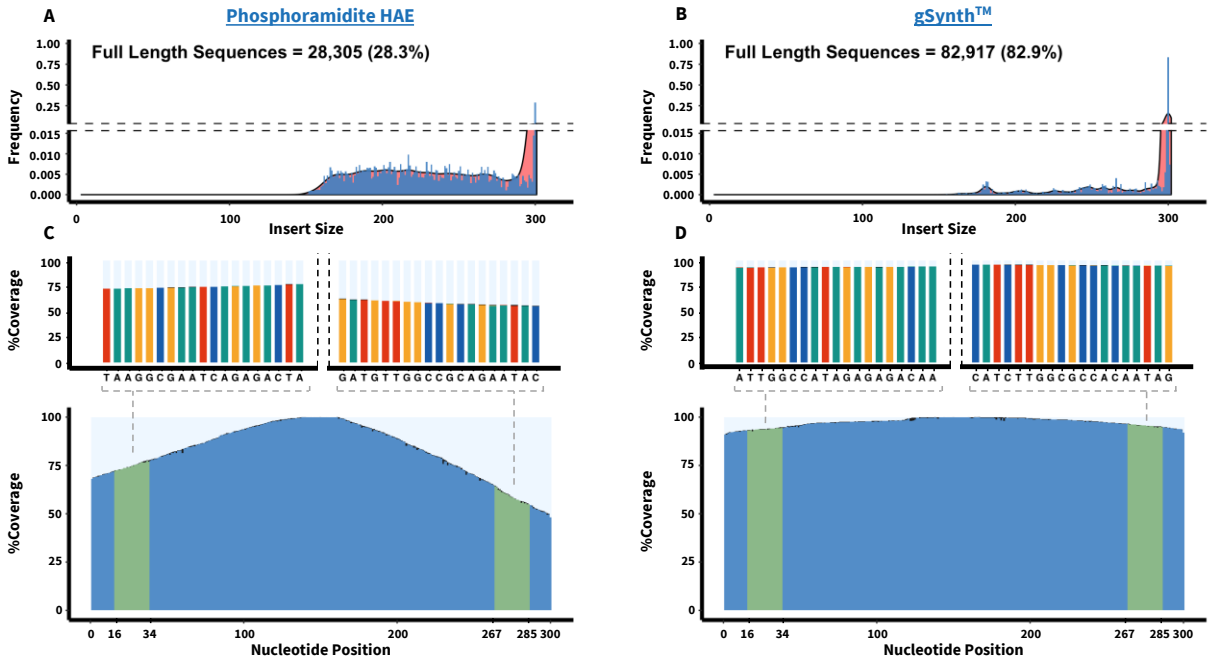

Figure 1. Analysis of 300mers SeqP1 and SeqG1, which varied from 40-60% in GC content. Histogram of the insert size for the phosphoramidite HAE product (A) and gSynth™ product (B). Graphs showing the sequence coverage throughout the 300mers for the phosphoramidite HAE product (C) and gSynth™ product (D). The green section in the coverage plots (C & D) are highlighted at greater resolution above the coverage plots, to show specific nucleotides (A = green, C = blue, G = yellow, T = red and insertions/deletions = black).

cycling assembly (PCA), which is a commonly used gene assembly method[3].

Informatics was performed by sampling 100,000, quality trimmed[4], paired-end reads for each 300mer and mapping to the reference sequences[5]. Overlapping regions from the aligned paired-end reads were removed[6] before synthesis accuracy was determined.

RESULTS:

To determine differences between DNA produced by the two different methods, we performed paired-end sequence analysis of the products. From 100,000 paired-end reads per sample, on average, 99.7 ± 0.1% of the gSynth™ reads aligned to their reference sequences (SeqG1, SeqG2 and SeqG3; Table 1), while an average of 96.4 ± 1.7% of the phosphoramidite HAE reads aligned (SeqP1, SeqP2 and SeqP3; Table 1). Furthermore, 85.3 ± 3.4% of the gSynth™ reads were the correct full-length, while only 22.7 ± 8.9% of the phosphoramidite HAE reads were the full-length (Table 1). Yields of 85.3% and 22.7% full-length product indicated a coupling efficiency of >99.9% for gSynth™ and 99.5% for phosphoramidite HAE. A nucleotide coupling efficiency of 99.5%, is consistent with known efficiencies and demonstrates our analysis was robust.

For 300mers the ranged from 40-60% GC content (SeqG1, SeqP1) 28.3% of the phosphoramidite HAE synthesised product was full-length, whereas 82.9% of the gSynth™ product was the correct length (Figure 1A-B). Furthermore, plots of sequence coverage versus sequence position highlighted that, for the phosphoramidite HAE product, the greatest coverage was at the centre of the 300mer and gradually tailed off toward the ends (Figure 1C, SeqP1). These results were

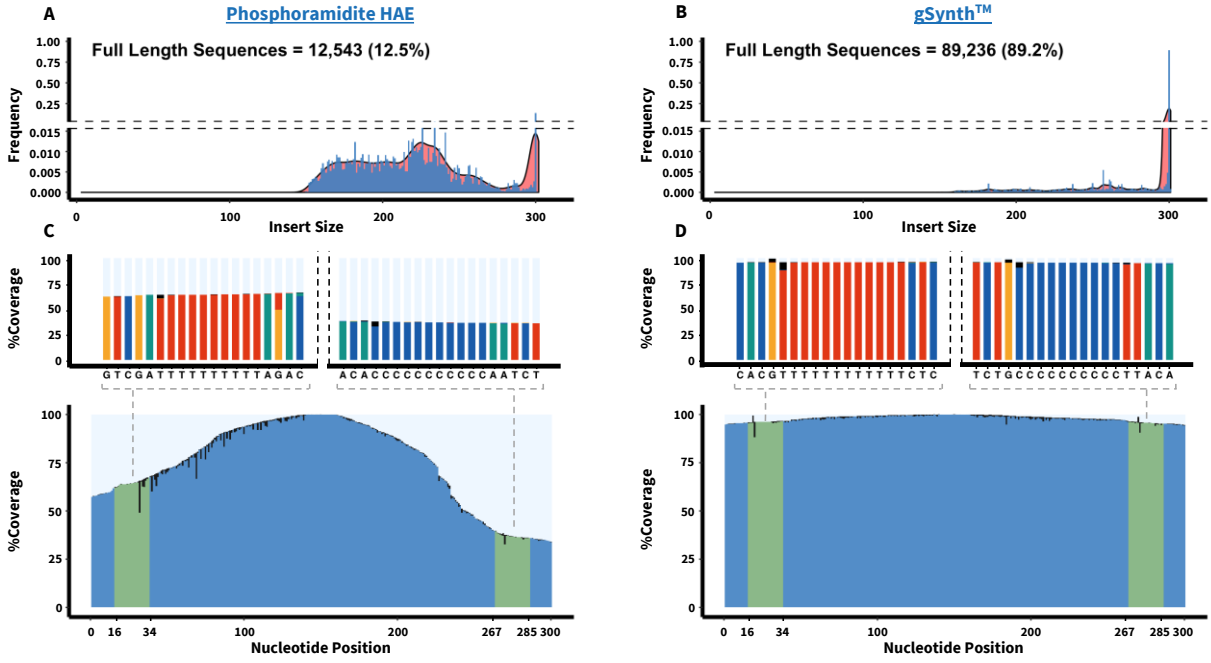

Figure 2. Analysis of 300mers SeqP2 and SeqG2, which contained 10 nucleotide T and C homopolymeric regions. Histogram of the insert size for the phosphoramidite HAE product (A) and gSynth™ product (B). Graphs showing the sequence coverage throughout the 300mers for the phosphoramidite HAE product (C) and gSynth™ product (D). The green section in the coverage plots (C & D) are highlighted at greater resolution above the coverage plots, to show specific nucleotides (A = green, C = blue, G = yellow, T = red and insertions/deletions = black).

expected as the central position is the 3' end of the two different phosphoramidite oligos, which is the most accurate. The gradual decrease in coverage reflects phosphoramidite synthesis errors and the accumulation of truncated sequences. In contrast, the sequence coverage for the gSynth™ 300mer (Figure 1D, SeqG1) remained high throughout all positions, which is consisted with a higher accuracy and coupling efficiency

Analysis of 300mers containing 10 nucleotide T and C homopolymeric regions (SeqG2, SeqP2) revealed an even more significant difference in accuracy between gSynth™ and phosphoramidite HAE (Figure 2). For the phosphoramidite HAE only 12.5% of the molecules were full-length (Figure 2A), compared to 89.2% for gSynth™ (Figure 2B). Sequence coverage of the phosphoramidite HAE homopolymeric 300mer (Figure 2C, SeqP2) was significantly lower than the sequence coverage of the phosphoramidite HAE 300mer with 40-60% GC (Figure 1C, SeqP1). This difference is consistent with the known difficulties in synthesising homopolymers. In contrast, sequence coverage obtained from the gSynth™ product remained high across all positions (Fig. 2D) demonstrating that gSynth™ can also accurately produce problematic sequences.

Finally, we analysed 300mers where six variable nucleotides N1 to N6 were incorporated into the sequence at set locations (SeqP3 and SeqG3). gSynth™ again showed that the number of full-length fragments was higher compared to phosphoramidite HAE, 27.5% versus 83.8%, respectively (Figure 3A-B). Also, gSynth™ consistently demonstrated, a greater and more even distribution of A, C,

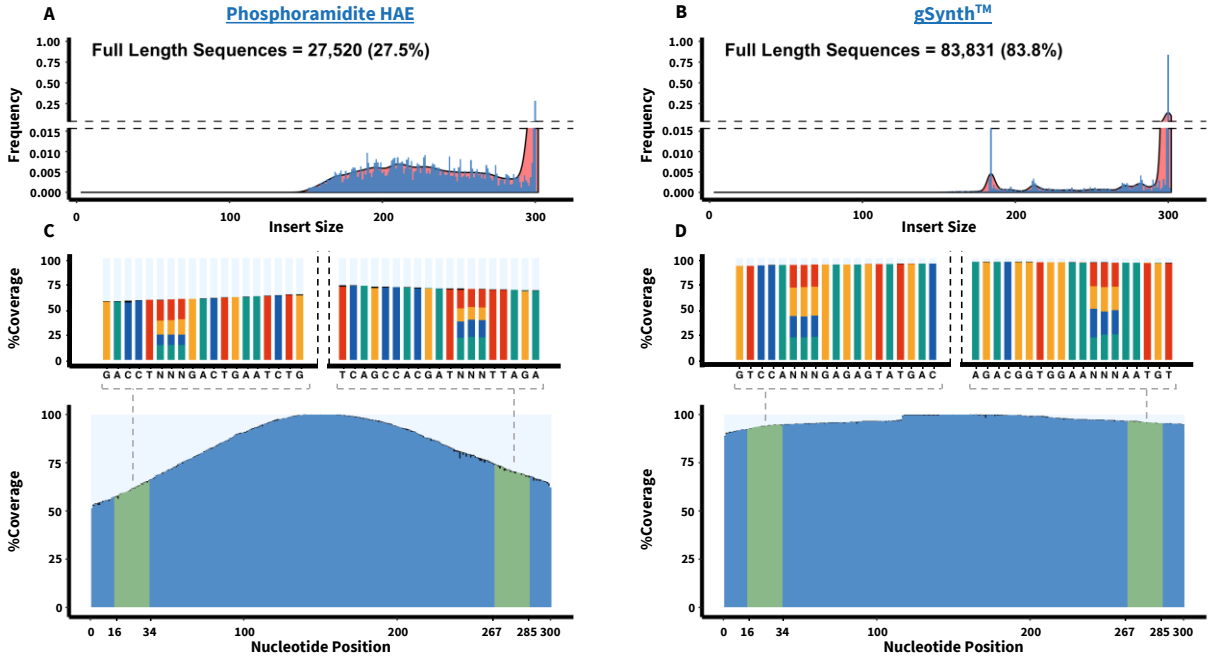

Figure 3. Analysis of 300mers SeqP3 and SeqG3, which contain the variable nucleotides N_1 to N_6. Histogram of the insert size for the phosphoramidite HAE product (A) and gSynth™ product (B). C-D, Graphs showing the sequence coverage throughout the 300mers for the phosphoramidite HAE product (C) and gSynth™ product (D). The green section in the coverage plots (C & D) are highlighted at greater resolution above the coverage plots, to show specific nucleotides (A = green, C = blue, G = yellow, T = red and insertions/deletions = black). At these specific locations the gSynth™ product (D) had balanced representation of nucleotides at all six positions. In contrast the phosphoramidite HAE product (C) has skewed representation of nucleotides.

Method	Nucleotide	[c]N_1	N_2	N_3	N_4	N_5	N_6
gSynth™	[a]Overall Coverage	93,156	93,264	93,409	95,568	95,529	95,479
	(%)	(100)	(100)	(100)	(100)	(100)	(100)
	[b]A Coverage	22,002	21,984	22,262	22,171	24,801	25,171
	(%)	(23.6)	(23.6)	(23.8)	(23.2)	(26.0)	(26.4)
	C Coverage	21,059	20,445	21,022	27,911	22,544	23,487
	(%)	(22.6)	(21.9)	(22.5)	(29.2)	(23.6)	(24.6)
	G Coverage	27,716	28,691	28,564	21,938	23,944	23,086
	(%)	(29.8)	(30.8)	(30.6)	(23.0)	(25.1)	(24.2)
	T Coverage	22,307	22,132	21,555	23,280	24,221	23,729
	(%)	(23.9)	(23.7)	(23.1)	(14.4)	(25.4)	(24.9)
	Ins/Del Coverage	72	12	6	268	19	6
	(%)	(0.1)	(0.0)	(0.0)	(0.3)	(0.0)	(0.0)
Phosphonamidite HAE	Overall Coverage	57,380	57,633	57,968	67,826	67,584	67,337
	(%)	(100)	(100)	(100)	(100)	(100)	(100)
	A Coverage	14,135	14,225	14,259	20,923	21,605	21,595
	(%)	(24.6)	(24.7)	(24.6)	(30.8)	(32.0)	(32.1)
	C Coverage	9,878	9,908	9,832	15,509	16,416	16,228
	(%)	(17.2)	(17.2)	(17.0)	(22.9)	(24.3)	(24.1)
	G Coverage	13,376	13,525	14,630	12,609	12,258	12,217
	(%)	(23.3)	(23.5)	(25.2)	(18.6)	(18.1)	(18.1)
	T Coverage	19,293	19,923	19,195	17,425	17,165	17,208
	(%)	(33.6)	(24.6)	(33.1)	(25.7)	(25.4)	(25.6)
	Ins/Del Coverage	698	52	52	1,360	140	89
	(%)	(1.2)	(0.1)	(0.1)	(2.0)	(0.2)	(0.1)

Table 2. Analysis of variable nucleotides N_1 to N_6. For the phosphoramidite HAE products 59.3-70.1% of all reads covered the variable nucleotides, for gSynth™ the percentage of reads covering the variable nucleotides was higher, ranging from 96.3-98.8%. Consistently the representation of each nucleotide was more balanced. (a) Number of times nucleotide N_1 to N_6 were covered from the initial 100,000 quality trimmed paired-ends reads. (b) From the overall coverage, how many times each nucleotide was called (and their corresponding percent). (c) Nucleotide positions - gSynth™ | N_1 = SeqG2:21; N_2 = SeqG2:22; N_3 = SeqG2:23, N_4 = SeqG2:21:278; N_5 = SeqG2:279 & N_6 = SeqG2:280. Phosphonamidite HAE | N_1 = SeqP2:21; N_2 = SeqP2:22; N_3 = SeqP2:23, N_4 = SeqP2:21:278; N_5 = SeqP2:279 & N_6 = SeqP2:280.

G and T nucleotides at the degenerate N1 to N6 nucleotide positions compared to phosphoramidite HAE (Figure 3C-D, Table 2).

DISCUSSION:

DNA synthesis is a cornerstone of synthetic biology and to develop critically important synthetic biology products, we must be able to accurately produce any DNA sequence of interest. Phosphoramidite synthesis has been the gold-standard DNA synthesis technology for many years. However, over long distances it is error-prone and consequently holding back many applications. For example, many problematic sequences, which may contain homopolymers or have a high percentage of GC content, are very difficult or impossible to produce. Additionally, for protein engineering, it is critical to generate DNA sequences with accurately synthesised variable nucleotides at specific locations.

In this application note, we have benchmarked phosphoramidite synthesis accuracy against gSynth™, Camena Bioscience's enzymatic *de novo* synthesis and gene assembly method. Our results show that gSynth™ consistently had superior accuracy across a variety of 300mers. This accuracy is highlighted by the balanced incorporation of variable nucleotides at set locations within the 300mers, which will be critically helpful for protein engineering. With improved

accuracy and the ability to produce problematic sequences, gSynth™ will enable synthetic biologists to develop novel synthetic biology applications.

REFERENCES:

1. Lubock, N.B., et al., "A systematic comparison of error correction enzymes by next-generation sequencing", Nucleic Acid Research, 2017, 45(15), 9206, doi: 10.1093/nar/gkx691.

2. Perkely, J.M., et al., "The race for enzymatic DNA synthesis heats up", Nature, 2019, 566(7745), 565, doi:10.1038/d41586-019-00682-0.

3. Stemmer; W.P., et al. (1995). "Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribo-nucleotides", Gene, 1995; 164(1): 49, DOI: 10.1016/0378-1119(95)00511-4.

4. Trim_Galore - https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

5. Bowtie2 - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

6. clipOverlap - https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap