Supplementary Materials for

Expanding the Molecular Alphabet of DNA-Based Data Storage Systems with Neural Network Nanopore Readout Processing

S. Kasra Tabatabaei, Bach Pham, Chao Pan, Jingqian Liu, Shubham Chandak, Spencer A. Shorkey, Alvaro G. Hernandez, Aleksei Aksimentiev, Min Chen, Charles M. Schroeder, Olgica Milenkovic

This file contains:

Materials and Methods

Oligo design and synthesis. All oligos tested are of fixed length 40nt and synthesized by Integrated DNA Technologies (IDT). For MspA experiments, the content of the oligos was chosen to include two polyT sequences at locations 1-12 and 17-40, and a chemically modified tetramer at positions 13-16. All oligos were biotinylated at the 5' end.

PCR Amplification. DNA amplification was performed via PCR using Q5 DNA polymerase, 5× Q5 buffer and pUC19 plasmid as template (New England Biolabs) in 50 µl. The 1.4kb sequence is:

5'CGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTT
GCAGCACATCCCCCTTTCGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATC
GCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCCTGATGCGGTATTT
TCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATGGTGCACTCTCAGTACAA
TCTGCTCTGATGCCGCATAGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGA
CGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACC
GTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGA
GACGAAAGGGCCTCGTGATACGCCTATTTTTATAGGTTAATGTCATGATAATAATG
GTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTG
TTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAA
ATGCTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCC
CTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTG
GTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAAC
TGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCA
ATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGC
CGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAG
TACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATG
CAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGA
TCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAAC
TCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGT
GACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCG

AACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAA
GTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATA
AATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAG
ATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTAT
GGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGT
A3'.

All primers were purchased from Integrated DNA Technologies (IDT). Both B1 and B2 were purchased from TriLink Biotechnologies in form of triphosphates (https://www.trilinkbiotech.com/2-amino-2-deoxyadenosine-5-triphosphate-n-2003.html and https://www.trilinkbiotech.com/5-hydroxymethyl-2-deoxycytidine-5-triphosphate.html). All natural and chemically modified nucleotides were added in equimolar ratios in all PCR reactions.

MD Simulations. The molecular mechanics models of modified nucleotides B1, B3, B4, B5 and B6, including their topology and force field parameter files, were generated using the CHARMM General Force Field (CGenFF) (1). The charge of the atom connecting to the sugar was adjusted so that the total charge of the base is zero, which is the case for all the natural nucleotides in CHARMM36. The parameters for B2 were adopted from a previous study (2). Eight systems each containing a modified Dickerson dodecamers (CGCGAATTCGCG)(3) were created starting from a B-DNA conformation to contain two different pairs of modified and natural bases while all other bases remained as in the original sequence. Each DNA duplex was immersed in a 75 Å x 75 Å x 75 Å volume of 1M KCl solution. After 2000 steps of energy minimization, the systems were equilibrated with the DNA backbone phosphate atoms restrained $(k_s = 1\text{kcal/mol/Å}^2)$ for the first 10ns. Each system contains approximately 39,000 atoms. Additional restrains were applied to enforce the expected hydrogen bonds between the modified and natural nucleotides for the first 20 ns. The systems were simulated for 350 ns in the absence of any restrains in the constant number of particles, pressure (1 atm) and temperature (295 K) ensemble using NAMD2 (4). If prominent structural disruptions had developed in both base pairs surrounding the modified nucleotide base pair, the simulation was terminated. Specifically, the simulation of the systems containing the B4 nucleotide lasted only 250

ns. Simulations of all the systems were performed using periodic boundary conditions. The simulations employed the particle mesh Ewald (PME) algorithm (5) to calculate long-range electrostatic interaction over a $1\,\text{Å}$-spaced grid. RATTLE (6) and SETTLE (7) algorithms were adopted to constrain all covalent bonds involving hydrogen atoms, allowing 2-fs time step integration used in the simulations. van der Waals interactions were calculated using a smooth $10-12\,\text{Å}$ cutoff. The NPT ensembles used the Nosé-Hoover Langevin piston pressure control (8), which maintained a constant pressure by adjusting system's dimension. Simultaneously, Langevin thermostat (25) was adopted for temperature control, with damping coefficient of $0.5\ \text{ps}^{-1}$ applied to all heavy atoms in the systems. CHARMM36 (9), output of CGenFF (1), TIP3P water model (10) as long as custom NBFIX corrections to nonbonded interactions (11) were employed as the parameter set of the simulation. The hydrogen bonds occupancy, the distances between hydrogen bond donors and acceptors as well as the short/long axis lengths of bases are calculated from the well equilibrated last 100 ns fragment of the trajectory using VMD (12). The hydrogen bonds were defined to have the donor-accepter interaction distance of less than 3Å and the cutoff angle of $20°$. Given the largely planar shape of the bases, their short/long were determined by first computing the three principal axes of the bases and then choosing the largest two values. Simulations/analysis of the B4 pairing with natural bases in longer DNA strands were conducted using the same methodology, but with only one modified base contained in the dodecamer. Besides, extra bonds were applied to the donor(N1) and accepter(N3) atoms on the terminal pairs to prevent the ends from fraying in these simulations to adapt the situation of long DNA strands. These simulations ran 550ns except if unstable configurations were observed.

MspA nanopores and purification of M2-NNN MspA. All chemicals were purchased from Fisher Scientific unless stated otherwise. Streptavidin was ordered from EMD Millipore (Burlington, MA) (Catalog # 189730). Phenylmethylsulfonyl fluoride (PMSF) was ordered from GoldBio (St. Louis, MO) (Catalog # P-470). DNA of M2-NNN MspA construct(13) was a gift from Dr. Giovanni Maglia (University of Groningen, Netherlands). The pT7-M2-NNN-MspA was transformed into BL21 (DE3) pLyss cells and grown in LB medium at 37°C until the OD600 reached 0.5-0.6. The cells were then induced with 0.5 mM isopropyl β–D-1-thiogalactopyranoside (IPTG) and continued to grow at 16°C for 16

hours. Cells were harvested and centrifuged at 19,000 x g for 30 min at 4°C. Cells were resuspended in the lysis buffer containing 100 mM Na_2HPO_4/NaH_2PO_4, 1 mM ethylenediaminetetraacetic acid (EDTA), 150 mM NaCl, 1 mM phenylmethylsulfonyl fluoride (PMSF) pH 6.5, before heating at 60℃ for 10 minutes. The cells were sonicated by using VWR Scientific Branson 450 sonicator (duty cycle of 20% and output control of 2) for 8 minutes. The lysate was centrifuged at 19,000 x g for 30 min and the supernatant was discarded. The pellet was resuspended in the solubilization buffer containing 100 mM Na_2HPO_4/NaH_2PO_4, 1 mM EDTA, 150 mM NaCl, 0.5% (v/v) Genapol X – 80, pH 6.5. After completely resuspending the pellet, it was centrifuged at 19,000 x g for 30 min. The supernatant, containing solubilized membrane extract, was collected for Ni-NTA purification. MspA was further purified using a 5 mL HisPur™ Ni-NTA resin (GE Healthcare) and eluted in a buffer of 0.5 M NaCl, 20 mM HEPES, 0.5% (v/v) Genapol X – 80, pH 8.0 by applying an imidazole gradient. MspA oligomers were further purified by SDS-PAGE gel extraction. The purified MspA protein was run in 7.5% SDS-PAGE gel. The band of MspA oligomer was cut from the gel and extracted in the extraction buffer containing 50 mM Tris-HCl, 150 mM NaCl, 0.5% Genapol X – 80, pH 7.5. The protein was extracted at room temperature (23°C) for 6 hours before centrifuged at 9,000 x g for 30 min to collect the protein solution. The purified MspA oligomer was fast frozen and stored at -80°C for further use.

Single-channel recording using MspA. The experiments were performed in a device containing two chambers separated by a 25 μm thick polytetrafluoroethylene film (Goodfellow) with an aperture of approximately 100 μm diameter located at the center. A hexadecane/pentane (10% v/v) solution was first added to cover both sides of the aperture. After the pentane evaporated, each chamber was then filled with buffer containing 1 M KCl 10 mM HEPES pH 8.0. 1, 2-diphytanoyl-sn-glycero-3-phosphocholine (DPhPC) dissolved in pentane (10 mg/mL) was dropped on the surface of the buffer in both chambers. After the pentane evaporated, the lipid bilayer was formed by pipetting the solution in both chambers below the aperture several times. An Ag/AgCl electrode was immersed in each chamber with the *cis* side grounded. M2-NNN MspA proteins (around 1 nM, final concentration) were also added to the *cis* chamber. To promote MspA insertion, a ≥ +200 mV voltage was applied. After a single MspA was inserted into the

planar lipid bilayer, the applied voltage was decreased to 150 mV (or 180 mV) for recording. The current was amplified with an Axopatch 200B integrating patch-clamp amplifier (Axon Instruments, Foster City, CA). Signals were filtered with a Bessel filter at 2 kHz and then acquired by a computer (sampling at 100 μs) after digitization with a Digidata 1440A/D board (Axon Instruments).

DNA immobilized in MspA. After recording a single MspA pore for 5-10 minutes at positive voltages to check its stability, 5'-biotinylated DNA sample (final concentration of 0.25 μM) was added to the *cis* chamber. Streptavidin (0.1 μM), added to solutions in the *cis* chamber, can bind to biotin to prevent the full translocation of the DNA strand through the nanopore. To collect the signal generated from each DNA samples, we applied a sweep protocol. The amplifier applied either 150 mV or 180 mV for 10 s then applied −150 mV to force the DNA out of the pore back into the *cis* compartment. The voltage was then returned to the original value and the sweep protocol repeated for at least 40 times at each voltage.

ONT sequencing protocol. NEB terminal transferase was used for A-tailing the 3' end of the 40-mer control oligos. The reaction mixture was made by 5ul 10X TdT buffer, 5ul 2.5mM CoCl2, 5 pmole DNA, 0.5ul 10mM dATP, 0.5 ul terminal transferase, and 38 ul H_2O. The reaction was Incubated at 37 C for 30 mins, followed by inactivation at 70 C for 10 mins. The DNA was then purified using the Zymo DNA clean up kit (ssDNA Buffer:sample=7:1) and eluted in 10ul warm H_2O. The Oxford Nanopore SQK-RNA002 kit was used for library preparation. The RT adaptor was ligated for 10min at room temperature, then mixed with reverse transcription master mix. 2uL of Superscript IV were added and the mixture was Incubated at 50 C for 50mins, followed by 70 C for 10mins and cooled down to 4 C. Bead clean-up was performed using 40ul samples with 72ul RNAClean XP beads, rotated for 5mins, washed by 70% EtOH and eluted by 20ul H_2O. The RMX adaptor was ligated in 10mins at room temperature, then 40ul RNA Clean XP beads clean-up was used, and the product was washed with 150ul of the wash buffer twice. It was then eluted in 21ul of the elution buffer. The reaction was loaded onto an R9.4.1 flowcell and sequenced on a GridION X5 (Oxford Nanopore) for 24 hs.

Additional results on the MspA readout experiments

Here, we provide a complete report on the experimental results of MspA nanopore detection of all 77 chemically modified tetramers.

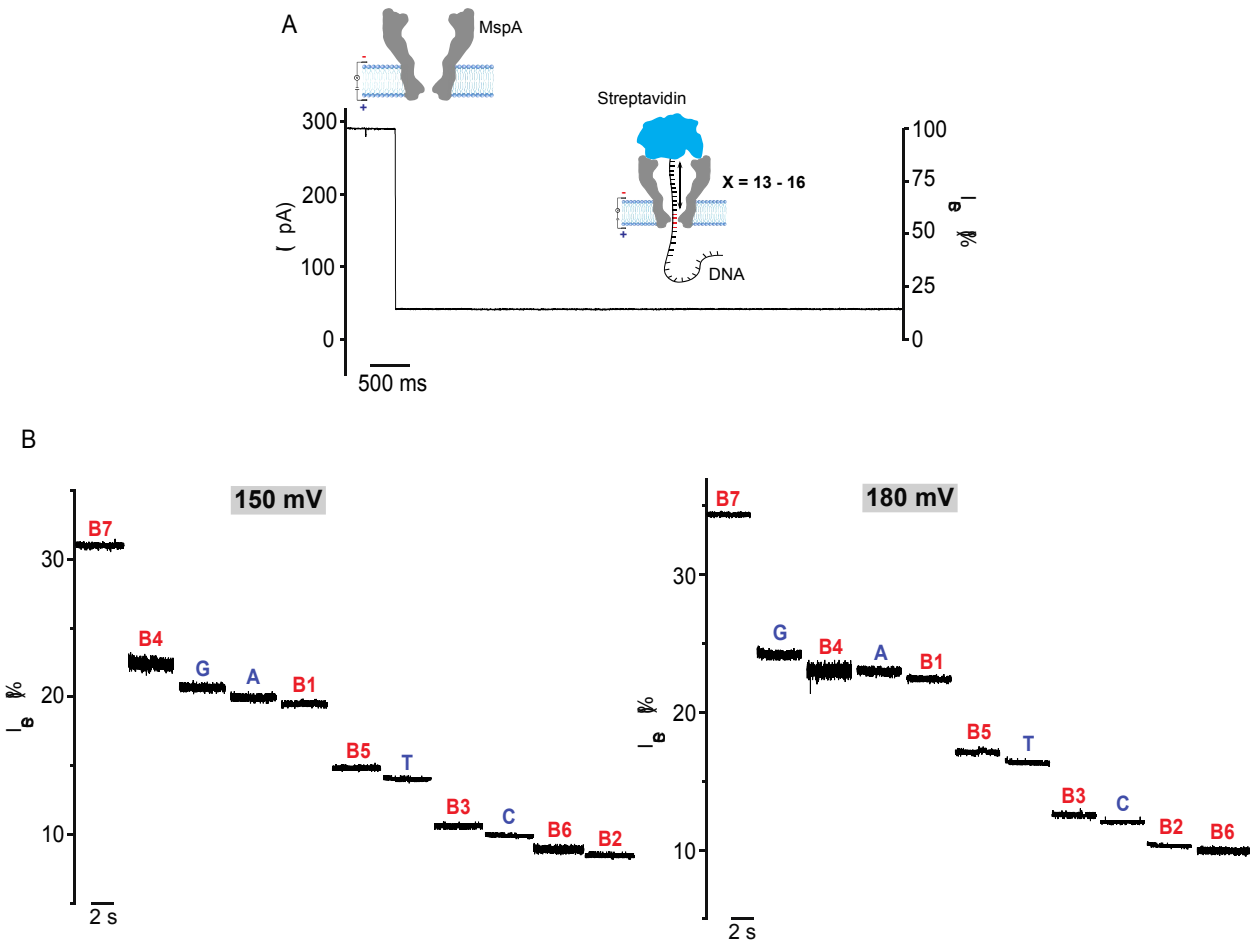
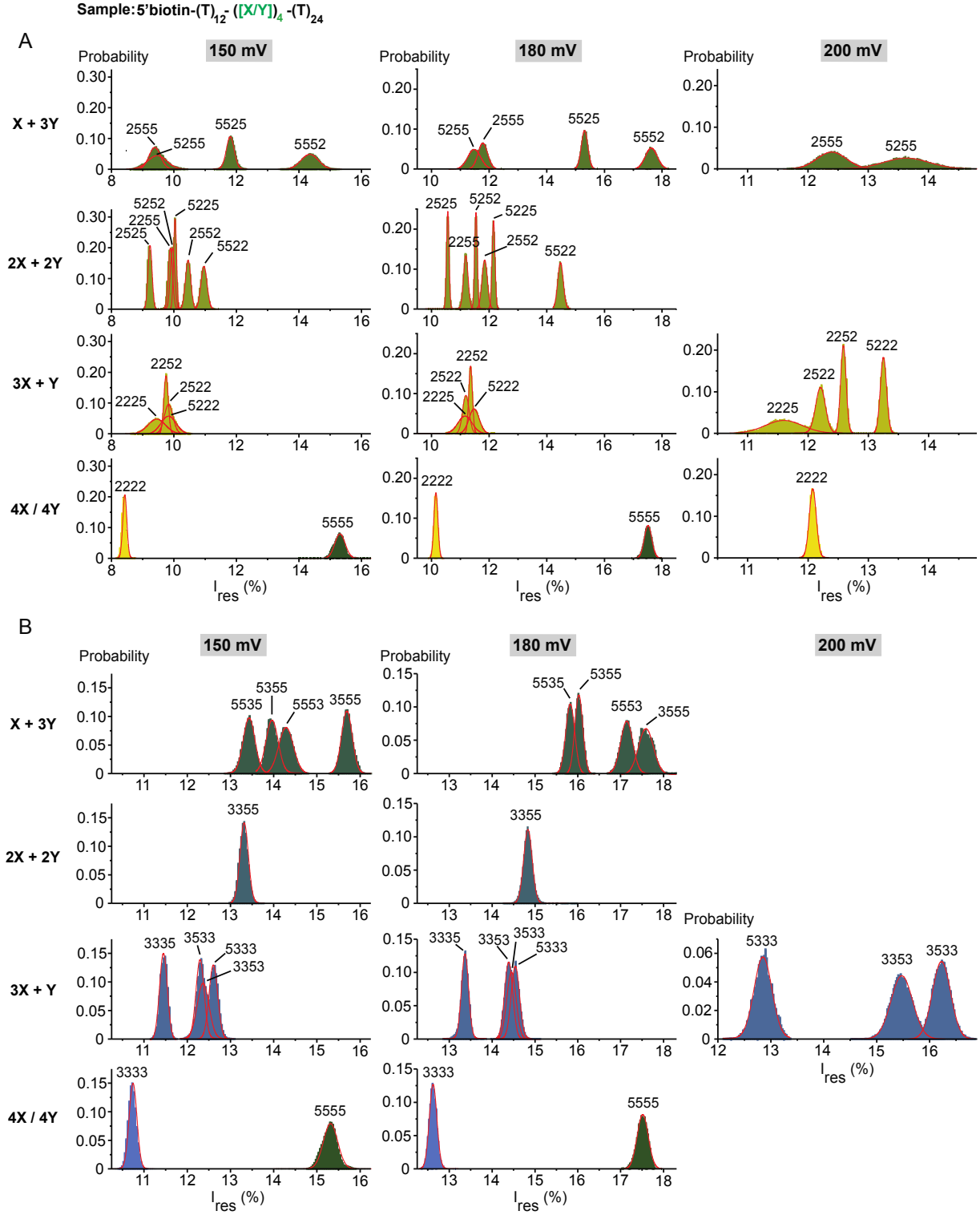


Figure S1. Discrimination of immobilized DNA by MspA nanopore. (A) Schematic diagram of DNA immobilized in the MspA nanopore. Single-stranded DNA (ssDNA) was attached to a streptavidin molecule (cyan) using a biotin linker. Bulky streptavidin prevents ssDNA to translocate through the MspA pore (gray). The residual ion current was recorded as the ssDNA is immobilized within the pore, which is generated by 4 nucleotides in and around the constriction side, at positions 13 – 16 from the biotin-streptavidin end. The open-pore current of MspA is normalized to 100%. **(B)** The representative single-channel recording generated by each tetramer sequence at positions 13-16 from the tethering point to the constriction site (reading head) of the MspA pore. Native nucleotides are highlighted in blue and modified nucleotides in red. Buffer used is 1 M KCl 10 mM HEPES pH 8.0.


Sample: 5'biotin-(T)12- ([X/Y])4 -(T)24
A
150 mV
180 mV
200 mV
Probability
X + 3Y
2X + 2Y
3X + Y
4X / 4Y
2555
5255
5525
5552
5252
5225
2525
2552
5522
2252
2522
5222
2225
2222
5555
I res (%)
B
5355
5535
5553
3555
3355
3335
3533
5333
3353
3333

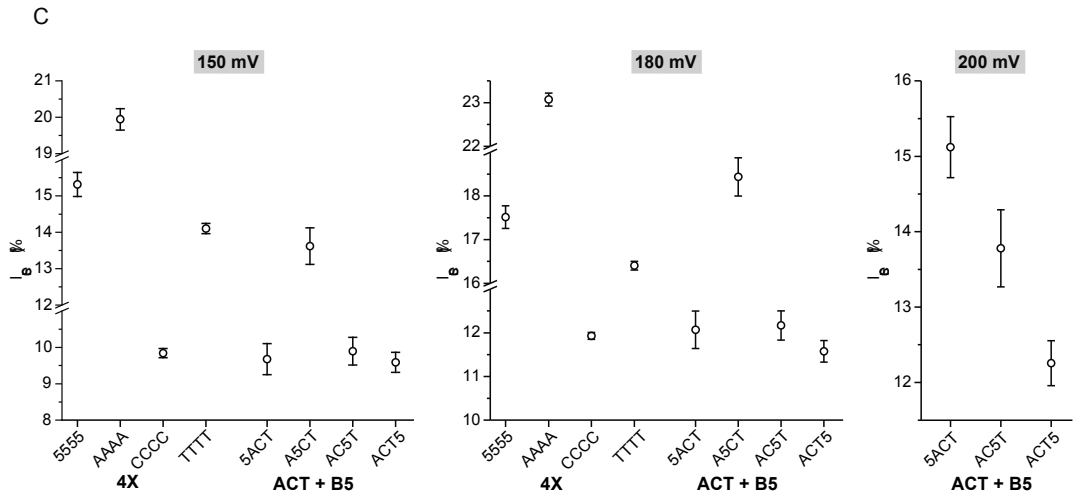


Figure S2. Histograms of the averaged residual ionic currents and the fitted Gaussian curves at various applied voltages for tetramers involving different orderings of B2 and B5 monomers **(A)** and B3 and B5 monomers **(B)** at 150, 180, and 200 mV. All experiments were performed in aqueous buffer (1 M KCl 10 mM HEPES pH 8.0). **(C)** Peak values and full-width half-height values (FWHM), represented as error bars, of the fitted Gaussian distributions around mean residual ionic currents generated by different orderings of B5 with the natural nucleotides (A, C, and T) at 150,180, and 200 mV. All experiments were performed in aqueous buffer (1 M KCl 10 mM HEPES pH 8.0).

Table S1. The mean residual currents (I_{res} (%)) and the full-width half-height (FWHM) values for each oligonucleotide were determined by Gaussian fitting of the residual current histogram from experiments with different combination of natural and modified nucleotides at positions 13 – 16 from the streptavidin anchor at 150 mV.

Combination	X	Y	Sample	I_{res} (%)	FWHM
ACT+X	2		2ACT	8.68	0.60
			A2CT	10.65	0.22
			AC2T	10.14	0.67
			ACT2	9.01	0.32
	3		3ACT	9.60	0.36
			A3CT	10.27	0.70
			AC3T	8.69	0.41
			ACT3	9.52	0.48
	5		5ACT	9.68	0.43
			A5CT	13.62	0.50
			AC5T	9.90	0.38
			ACT5	9.59	0.28
4X	1		B1	19.66	0.39
	2		B2	8.43	0.13
	3		B3	10.75	0.18
	4		B4	22.74	0.51
	5		B5	15.32	0.33
	6		B6	8.49	0.29
	7		B7	31.30	0.12
	A		A	19.94	0.29
	C		C	9.84	0.13
	G		G	20.82	0.50
	T		T	14.10	0.14
3X+Y	2	3	2223	9.13	0.34
			2232	7.36	0.38
			2322	8.34	0.37
			3222	9.45	0.29
		5	2225	9.45	0.55
			2252	9.75	0.14
			2522	9.83	0.27

			5222	9.83	0.48
	3	2	2333	7.91	0.19
			3233	7.48	0.30
			3323	9.44	0.29
			3332	10.45	0.42
		5	3335	11.45	0.18
			3353	12.37	0.27
			3533	12.30	0.19
			5333	12.61	0.20
	5	2	2555	9.39	0.37
			5255	9.46	0.60
			5525	11.80	0.25
			5552	14.35	0.55
		3	3555	15.69	0.25
			5355	13.96	0.28
			5535	13.43	0.27
			5553	14.29	0.34
2X+2Y	2	3	2323	8.53	0.17
			2332	8.07	0.14
			3223	10.02	0.16
			3232	8.18	0.16
			3322	11.34	0.17
			2233	7.59	0.14
		4	2424	12.79	0.20
			2442	13.01	0.59
			4224	12.39	0.12
			4242	12.62	0.19
			4422	12.99	0.21
			2244	10.78	0.18
	2	5	2525	9.23	0.13
			2552	10.45	0.17
			5225	10.03	0.09
			5252	9.95	0.14
			5522	10.96	0.20
			2255	9.89	0.13
	4	5	4545	23.07	0.34

			5454	20.16	0.43
			4554	19.55	0.20
			5445	19.38	0.32
			5544	17.63	0.24
			4455	22.01	0.33
	1	2	1122	11.18	0.27
		3	1133	16.16	0.22
		4	1144	18.09	0.30
		5	1155	17.57	0.21
	3	4	3344	19.07	0.85
		5	3355	13.32	0.19

Table S2. The mean residual currents (I_{res} (%)) and the full-width half-height (FWHM) values for each oligonucleotide, determined by performing Gaussian fitting of the residual current histogram from experiments involving different combination of natural and modified nucleotides at positions 13 – 16 from the streptavidin anchor at 180 mV.

Combination	X	Y	Sample	I_{res} (%)	FWHM
ACT+X	2		2ACT	11.06	0.49
			A2CT	12.93	0.21
			AC2T	12.02	0.59
			ACT2	10.53	0.28
	3		3ACT	12.38	0.38
			A3CT	14.27	0.61
			AC3T	10.74	0.40
			ACT3	11.38	0.42
	5		5ACT	12.07	0.43
			A5CT	18.44	0.44
			AC5T	12.17	0.34
			ACT5	11.58	0.25
4X	1		B1	22.52	0.25
	2		B2	10.15	0.14
	3		B3	12.62	0.18
	4		B4	23.25	0.54
	5		B5	17.51	0.26
	6		B6	9.90	0.27
	7		B7	34.13	0.19
	A		A	23.07	0.30
	C		C	11.93	0.16
	G		G	23.57	0.49
	T		T	16.40	0.20
3X+Y	2	3	2223	11.07	0.22
			2232	8.97	0.33
			2322	9.64	0.26
			3222	11.54	0.26
		5	2225	11.16	0.49
			2252	11.36	0.13
			2522	11.19	0.22

			5222	11.48	0.35
	3	2	2333	9.25	0.16
			3233	9.69	0.26
			3323	12.34	0.24
			3332	12.64	0.39
		5	3335	13.36	0.16
			3353	14.38	0.19
			3533	14.45	0.22
			5333	14.54	0.19
	5	2	2555	11.78	0.33
			5255	11.48	0.45
			5525	15.31	0.22
			5552	17.62	0.42
		3	3555	17.59	0.35
			5355	16.02	0.19
			5535	15.82	0.21
			5553	17.14	0.28
2X+2Y	2	3	2323	9.65	0.15
			2332	9.60	0.17
			3223	12.15	0.17
			3232	10.05	0.17
		3	3322	13.66	0.18
			2233	9.86	0.15
		4	2424	14.14	0.22
			2442	15.94	0.36
			4224	14.57	0.17
			4242	15.22	0.18
			4422	15.80	0.35
			2244	12.58	0.15
	2	5	2525	10.57	0.09
			2552	11.85	0.19
			5225	12.15	0.10
			5252	11.55	0.09
			5522	14.48	0.19
			2255	11.19	0.16
	4		4545	25.65	0.41

		5			
			5454	20.99	0.38
			4554	22.27	0.28
			5445	20.74	0.45
		5	5544	19.56	0.26
			4455	23.70	0.41
	1	2	1122	14.05	0.24
		3	1133	18.93	0.21
		4	1144	21.09	0.23
		5	1155	20.50	0.25
	3	4	3344	20.18	0.87
		5	3355	14.83	0.20

Table S3. The mean residual currents (I_{res} (%)) and the full width half height values (FWHM) for each oligonucleotide were determined by performing Gaussian fits to the residual current histogram from experiments with different combination of natural and modified nucleotides at position x = 13 – 16 from the streptavidin anchor at 200 mV.

Combination	X	Y	Sample	I_{res} (%)	FWHM
ACT+X	2		2ACT	12.08	0.33
			ACT2	11.57	0.44
	5		5ACT	15.12	0.40
			AC5T	13.78	0.51
			ACT5	12.26	0.30
4X	2		B2	12.08	0.11
3X+Y	2	3	2322	9.93	0.11
		5	2225	11.60	0.61
			2252	12.59	0.09
			2522	12.21	0.16
			5222	13.25	0.10
	3	2	2333	10.00	0.16
		5	3353	15.47	0.41
			3533	16.22	0.34
			5333	12.85	0.34
	5	2	2555	12.39	0.49
			5255	13.63	0.75
2X+2Y	2	3	2323	10.57	0.16
			2332	10.35	0.17
			3232	10.86	0.26
		4	2442	17.19	0.25
			4422	17.98	0.55

Two-step event identification scheme for ONT readouts with NN processing

The main challenges faced when analyzing nanopore current signals are illustrated in **Figure S3.** The figure shows the extreme variations in the current levels, which can either stay close to the mean (as illustrated on the example CCCC) or deviate more than 15% from the mean (as illustrated on the example 2233). Therefore, to automatically extract the regions from the ONT current readouts that correspond to modified nucleotides without resorting to basecalling, we developed a two-step identification scheme depicted in **Figure S3.** The first step is to estimate the current level for the polyA region, which is subsequently used for calibration purposes. We used kernel density estimation of the signal level distribution (14), followed by identification of the levels that have the two largest probabilities in the estimated distribution. This approach is justified by the observation that in our oligo structure, the polyA regions constitute the longest signal component. As polyT current levels are expected to be lower than polyA levels, we subsequently filtered out readout regions that are trailed by nearly flat regions with a mean level value lower than that observed for the polyA tails, using a finite state machine (15).These regions are expected to bear the signal from the chemically modified nucleotides.

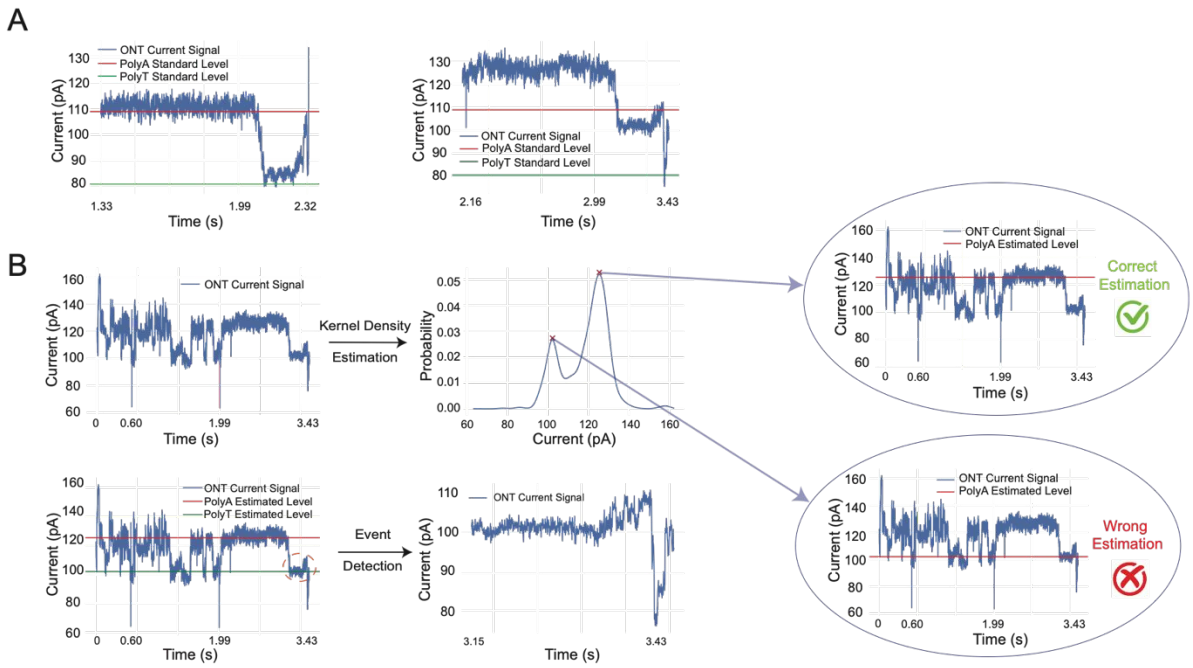


Figure S3. (**A**) (Left) Raw current readout of a control oligo bearing the content CCCC. (Right) A raw current readout bearing the content 2233. The red and green lines represent the expected standard levels for polyA and polyT regions, respectively. (**B**) Analysis of nanopore sequencing results for chemically modified nucleotides. (Top Left) Raw current readout for a control oligo containing the sequence 2233. (Top Right) Visualization of the kernel density estimation method: Two peaks correspond to two possible polyA region levels. (Bottom) The procedure for determining which level to use for calibration, based on the mean value of the “nearly-flat” region following the predicted polyA region. An example of the current level corresponding to the highest peak, which was used to correctly estimate the location of the polyA region. Building upon this step, the results show that one can also isolate the signal region which corresponds to the chemically modified nucleotides.

Summary of results from model-based classification procedure

We trained ResNet models on 12 permutation classes in which the composition is fixed, but the orderings of the modified nucleotides are different. What we refer to as a “superclass” combines different choices and orderings of the modified nucleotides (the superclass contains 66 out of 77 tetramers, as for 11 tetramers an insufficient number of training samples was available). The number of valid sequenced reads (i.e., reads

containing modified nucleotides) for each class is shown in **Table S4**. To perform unbiased training, we balanced out the sizes of the classes by setting a lower bound for subsampling of reads in different classes. We also set an upper bound on the number of training samples used for each class, in order to prohibit one/several classes to dominant the training set. For finer classification involving permutations of nucleobases within a class, we set the lower bound to 1000, and the upper bound to 5000. For the classification task on all 66 classes, we set the lower bound to 2000, and the upper bound to 3500. These choices are necessitated by two conflicting requirements: To balance out the class sizes and retain a training set as large as possible. The classification results are shown in **Figure S4**. From the confusion matrices we observe that almost all combinations can be easily distinguished from each other with very high accuracies (i.e., the diagonal values are significantly larger than the off-diagonal values). However, there are some tetramer instances that are hard to classify, such as 3223 (when compared to a tetramer in {2233, 3322, 2332, 3223, 2323, 3232}). The average classification accuracies for each model trained are listed in the caption of **Figure S4.**

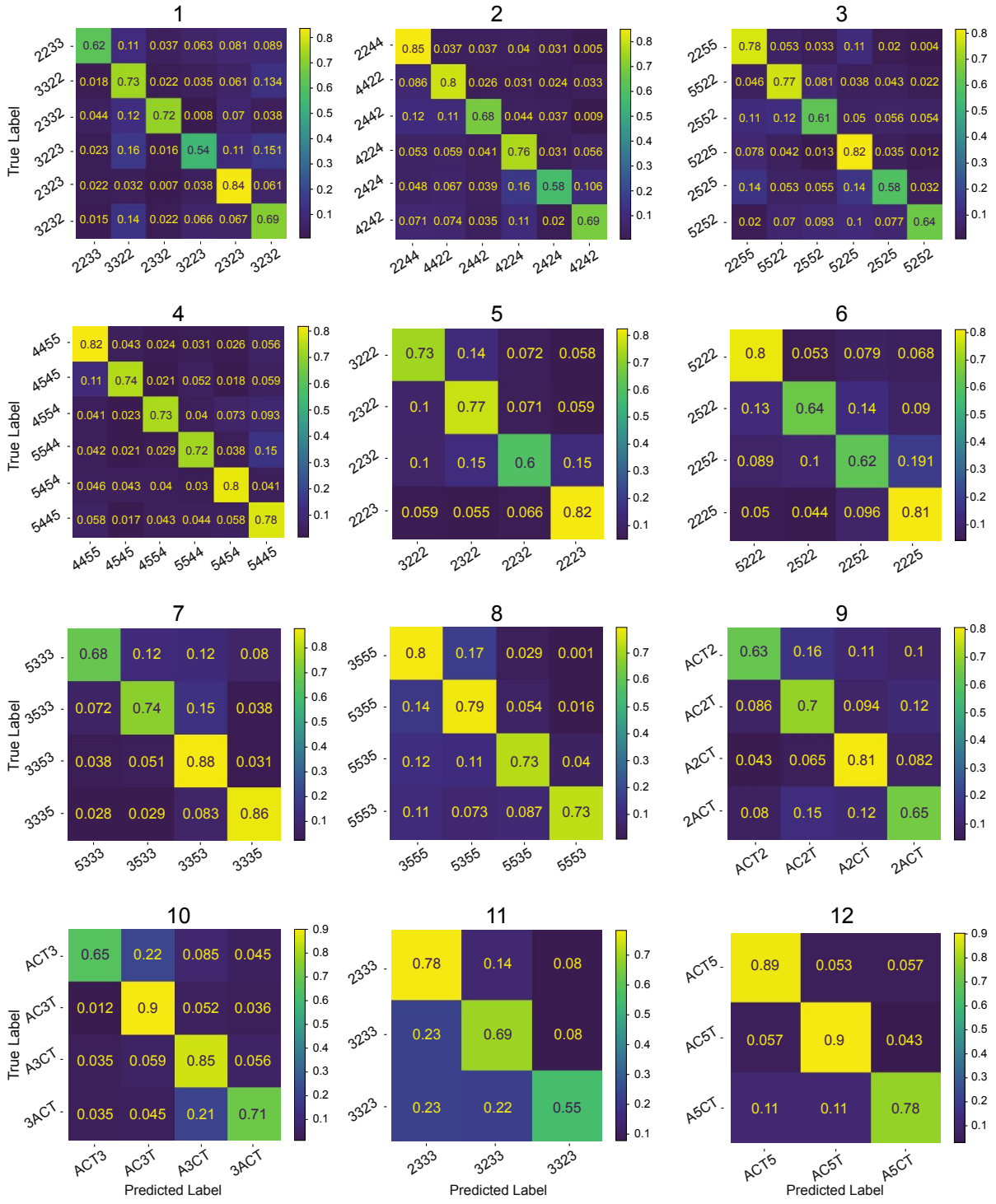


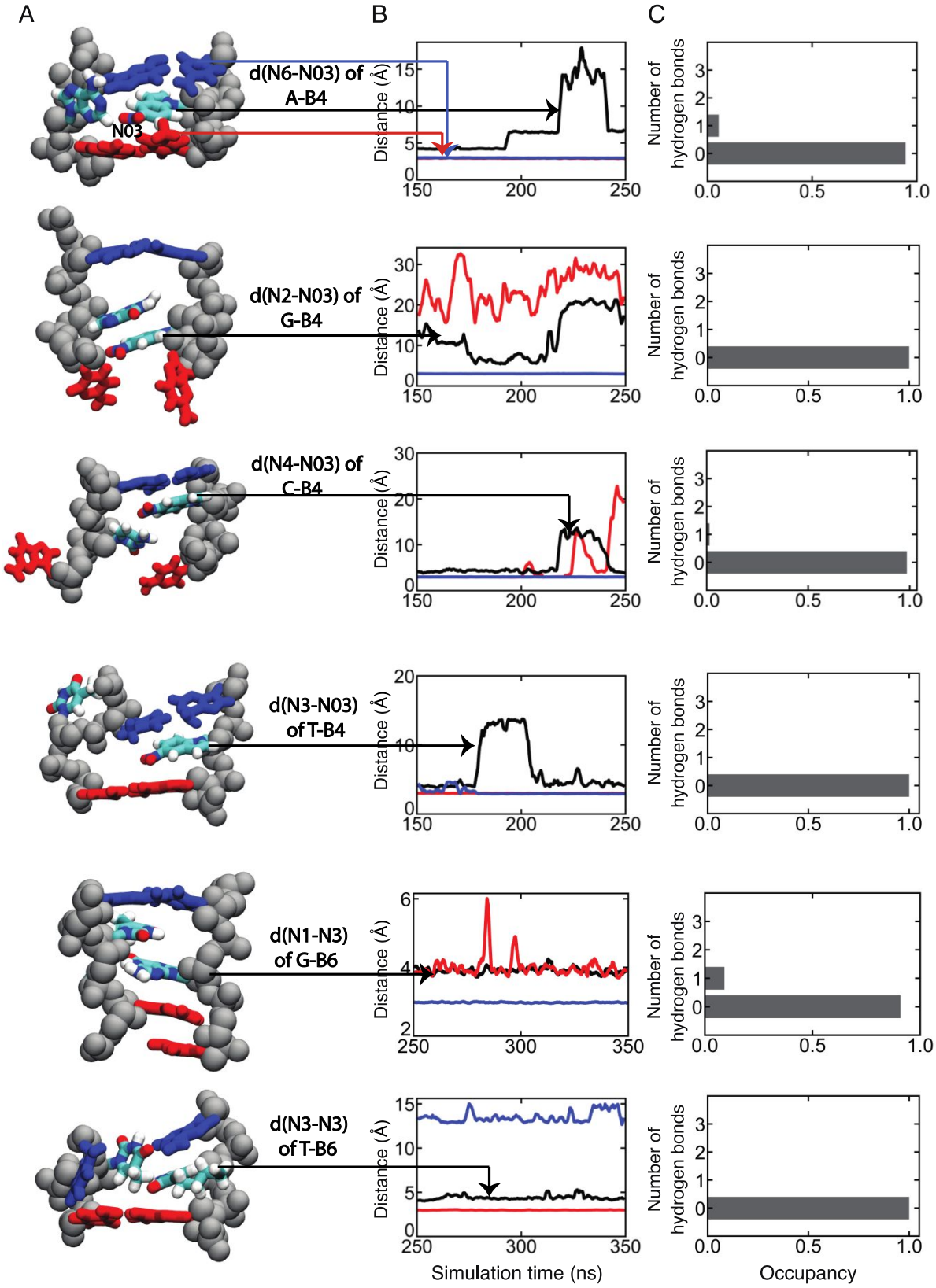
Figure S4. Classification performance of 12 different classes of tetramers. The names of the classes are listed in the subfigures, along with their average classification accuracies: (1) 69.39 ± 0.93%, (2) 72.25% ± 1.46%, (3) 68.87% ± 0.90%, (4) 77.84% ± 0.96%, (5) 72.18% ± 1.79%, (6)

71.97% ± 0.54%, (7) 81.27% ± 0.93%, (8) 79.17% ± 1.87%, (9) 69.66% ± 0.48%, (10) 80.04% ± 0.69%, (11) 70.81% ± 1.15%, (12) 88.00% ± 1.31%.

Class Name	Number of valid reads	Class Name	Number of valid reads	Class Name	Number of valid reads	Class Name	Number of valid reads	Class Name	Number of valid reads
3332	39	**5255**	74	**2555**	204	**5ACT**	315	**7777**	712
5525	750	**TTTT**	1390	**ACT3**	1717	**3323**	1808	**3555**	1885
5552	1944	**A5CT**	2133	**5535**	2315	**3233**	2344	**5333**	2430
5553	2460	**4444**	2553	**GGGG**	2607	**6666**	2632	**2424**	2706
1144	2723	**4422**	2740	**1133**	3134	**3353**	3167	**4242**	3310
4224	3377	**3223**	3732	**ACT2**	3837	**3322**	3865	**2442**	3967
2255	4039	**4545**	4072	**4455**	4500	**3333**	4506	**5555**	4630
5225	4657	**4554**	4827	**2ACT**	4827	**1122**	4844	**5355**	4925
A2CT	5197	**CCCC**	5198	**5522**	5236	**3232**	5324	**3ACT**	5403
5544	5485	**AC2T**	5505	**2333**	5612	**5222**	5905	**2222**	5958
5454	6090	**5445**	6163	**3222**	6395	**2244**	6484	**2252**	6509
3533	6526	**AC5T**	6532	**3355**	6556	**2522**	6799	**2233**	7047
2525	7403	**A3CT**	7448	**2225**	7563	**1155**	7591	**2223**	7700
3344	7716	**AAAA**	7927	**3335**	7952	**2552**	9525	**2232**	9955

ACT5	11768	**1111**	13502	**2322**	13915	**2323**	15927	**5252**	16104
2332	17890	**AC3T**	22040						

Table S4. The number of valid reads for each tetramer class (77 classes in total), arranged in ascending order.


A
B
C
N03
d(N6-N03) of A-B4
d(N2-N03) of G-B4
d(N4-N03) of C-B4
d(N3-N03) of T-B4
d(N1-N3) of G-B6
d(N3-N3) of T-B6
Distance (Å)
Number of hydrogen bonds
Simulation time (ns)
Occupancy

D

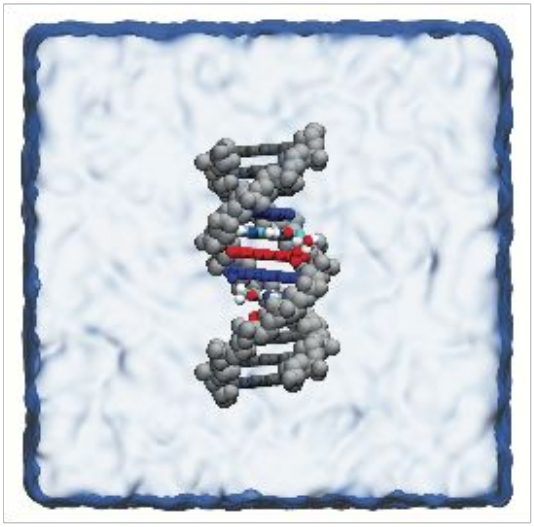


Figure S5. Interactions between modified and natural bases that do not involve stable hydrogen bonds. **(A)** Microscopic configurations of modified base pairs (from top to bottom: B2—G, A—B4, G—B4, C—B4, T—B4, G—B6, and T—B6). The backbone of the dodecamer is shown using silver spheres whereas the bases are drawn as molecular bonds. Unnatural bases and the natural bases that pair with them are colored according to the atom type (cyan for carbon, blue for nitrogen and red for oxygen). Base pairs immediately adjacent to the modified base pair are colored in red or blue.

(B) Distance between the key atoms of the modified base pair during the last 100 ns of the 350 ns MD simulation. The red curve and blue curve show the N1—N3 distance for the two adjacent base pairs, whose pairing patterns can either remain intact or be disrupted. The arrows starting from panel A to panel B indicate the correspondence between the basepairs and the curves. The label specifies the atoms used to compute the distance. The curves show a running average of the 10 ps-sampled data with a 2 ns averaging window. (**C**) Probability of observing the specified number of hydrogen bonds within a modified base pair. The H-bonding probabilities were computed using the final 100 ns of a 350 ns all-atom MD simulation of a DNA dodecamer. (**D**) Initial state of a simulation system where a DNA dodecamer containing chemically modified nucleotides is immersed in electrolyte solution. The backbone of the dodecamer is shown using silver spheres whereas the bases are drawn as molecular bonds. Chemically modified bases and the natural bases that pair with them are colored according to the atom type (cyan for carbon, blue for nitrogen and red for oxygen). Base pairs immediately adjacent to the modified base pair are colored in red or blue.

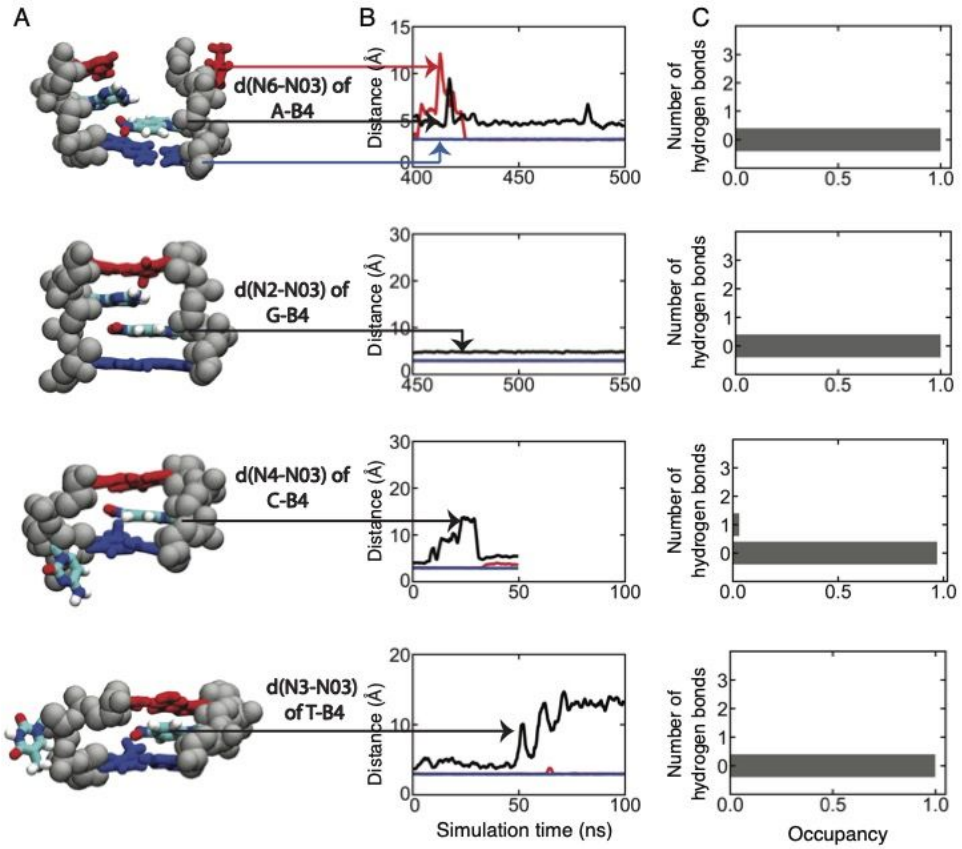


Figure S6. Interactions between B4 and natural bases in long DNA strands **(A)** Microscopic configurations of modified base pairs (from top to bottom: A—B4, G—B4, C—B4 and T—B4). The backbone of the dodecamer is shown using silver spheres whereas the bases are drawn as molecular bonds. B4 bases and the natural bases that pair with them are colored according to the atom type (cyan for carbon, blue for nitrogen and red for oxygen). Base pairs immediately adjacent to the modified base pair are colored in red or blue. In contrast to simulations reported in Figure S4, here each DNA dodecamer contains only one B4 base. Extra bonds between donor(N1) and acceptor(N3) (The equilibrium length was set as 2.9 Å. The spring constant was set as $1\mathrm{kcal/mol/Å^2}$.) are applied the terminal base pairs, preventing DNA from fraying and thereby mimicking an environment of a longer DNA strand. **(B)** Distance between the key atoms of the modified base pair during the last 50/100 ns of the MD simulation. The red curve and blue curve show the N1—N3 distance for the two adjacent base pairs, whose pairing patterns can either remain intact or be disrupted. The arrows starting from panel A to panel B indicate the correspondence between the basepairs and the curves. The label specifies the atoms used to compute the distance. The curves show a running average of the 10 ps-sampled data with a 2 ns averaging window. **(C)** Probability

of observing the specified number of hydrogen bonds within a modified base pair. The H-bonding probabilities were computed using the final 50/100ns of the all-atom MD simulations of a DNA dodecamer.

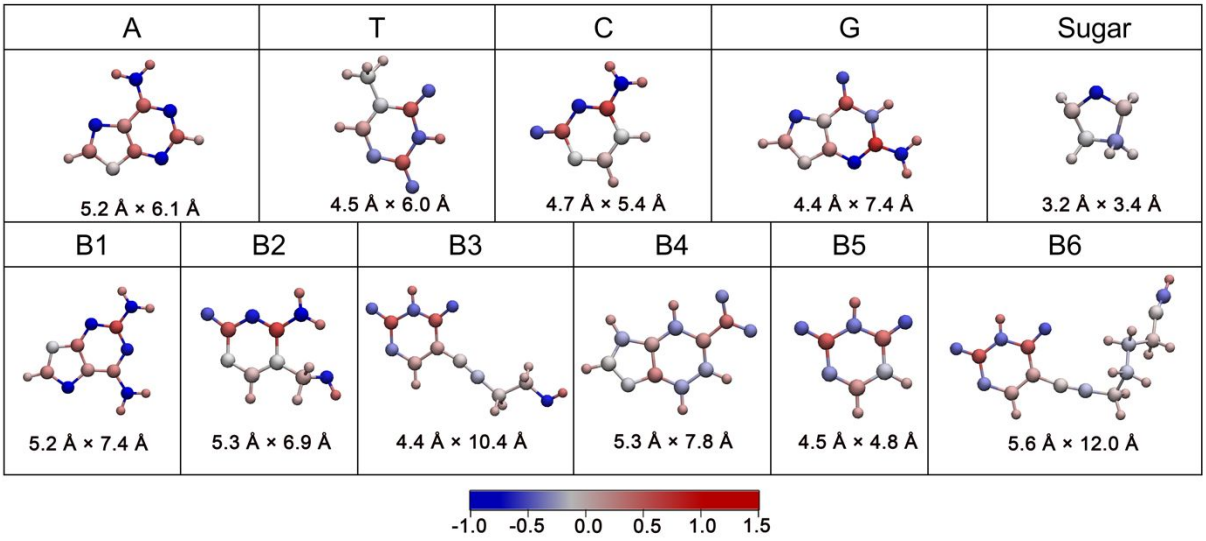


Table S5. Charge distribution and dimensions of the natural and modified bases and deoxyribose moieties in simulation. The chemical moieties are shown using a ball-and-stick representation, with the atoms colored by their charge according to the color bar. The dimensions of each base, specified as short axis length × long axis length, were averaged over the last 100 ns of a 350 ns/250ns all-atom MD trajectory of a DNA dodecamer.

Although it may appear that issues observed with some nucleobases may prohibit the use of them in storage applications, or implementations that include the other recommended chemically modified nucleotides, further experiments are required to confirm which combinations may cause disruptions. Once such combinations are identified, well-known methods from coding theory – such as constrained coding (16) – may be used to eliminate the offending patterns with minimal loss in the information rate.

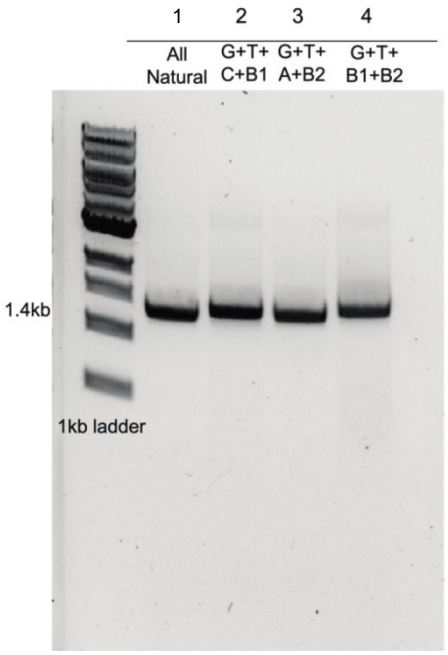
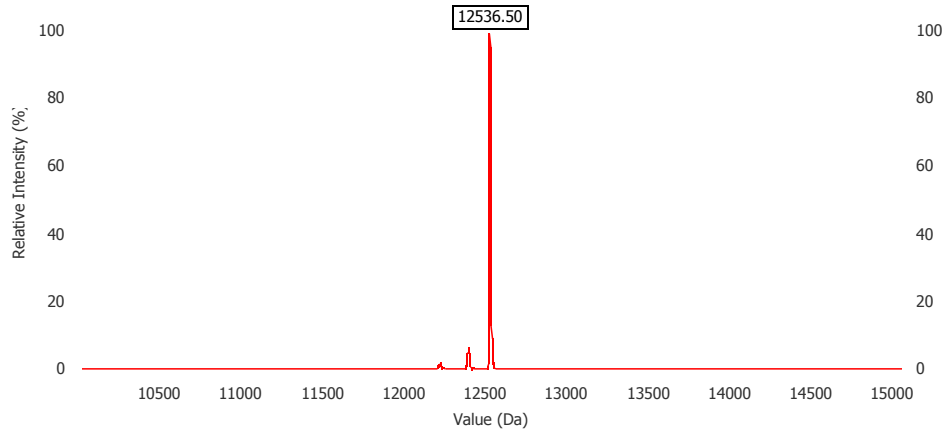


Figure S7. As a starting point to experimentally evaluate the effect of chemically modified nucleotides on DNA structure, we performed a simple PCR reaction on a 1.4kb double stranded DNA from a commonly used vector, pUC19 plasmid, using Q5 polymerase. The reaction was either supplied by all four natural nucleotides or B1 and B2 as substitutes for A and C. The final PCR products were run on 1% agarose gel. The results indicate successful incorporation of B1 and B2 into DNA duplex structure when only one of them (lanes 2 and 3) or two of them (lane 4) were used instead of the natural nucleotides.

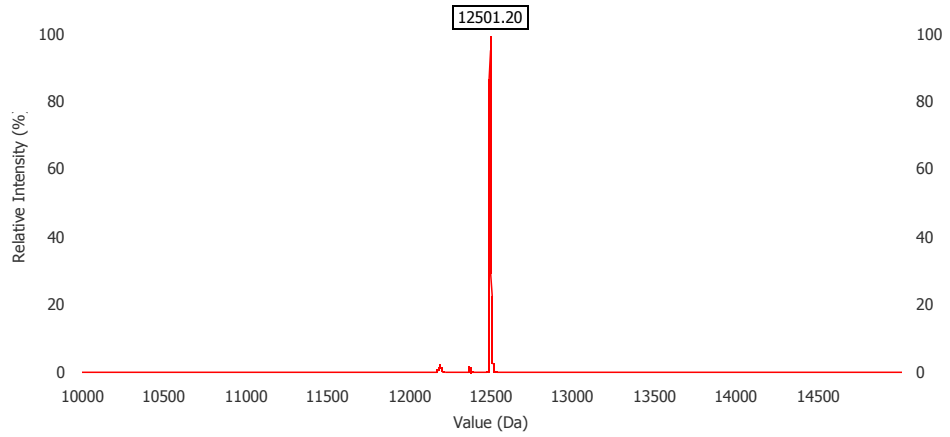
Synthetic DNA oligos QC data

Below (**Figures S8-S84**) we provide the Electrospray Ionization Mass Spectrometry (ESI-MS) plots for all synthetic DNA oligos containing natural/ chemically modified nucleotides. The data was generated by Integrated DNA Technologies (IDT).



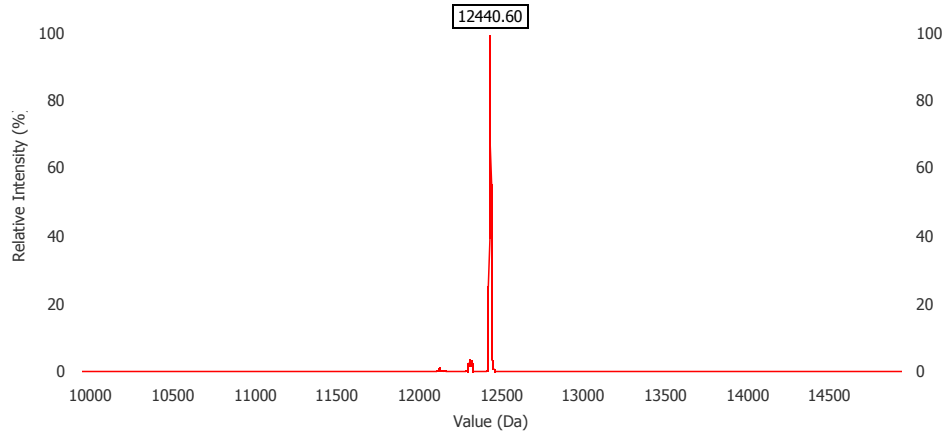
Sequence Name:	C1
Sequence:	5'-/5Biosg/TTT TTT TTT TTT AAA ATT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12535.3
Measured Molecular Weight:	12536.50

Figure S8. ESI-MS plot of Control A.



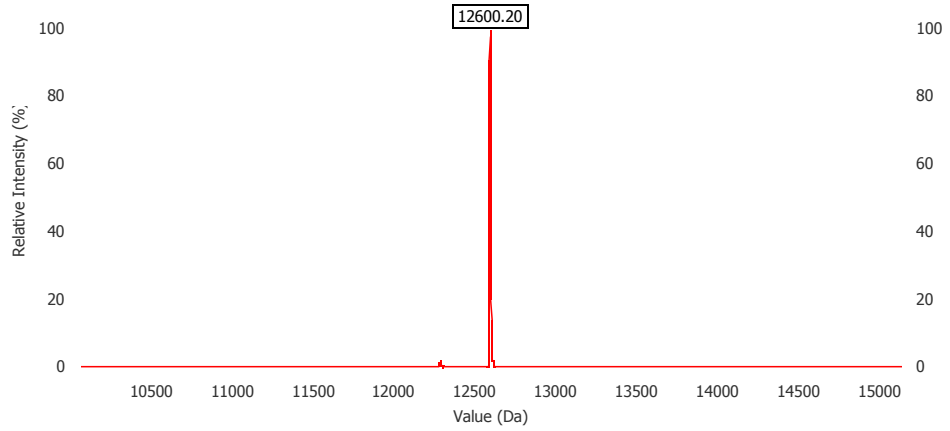
Sequence Name:	C2
Sequence:	5'-/5Biosg/TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12499.3
Measured Molecular Weight:	12501.20

Figure S9. ESI-MS plot of Control T.



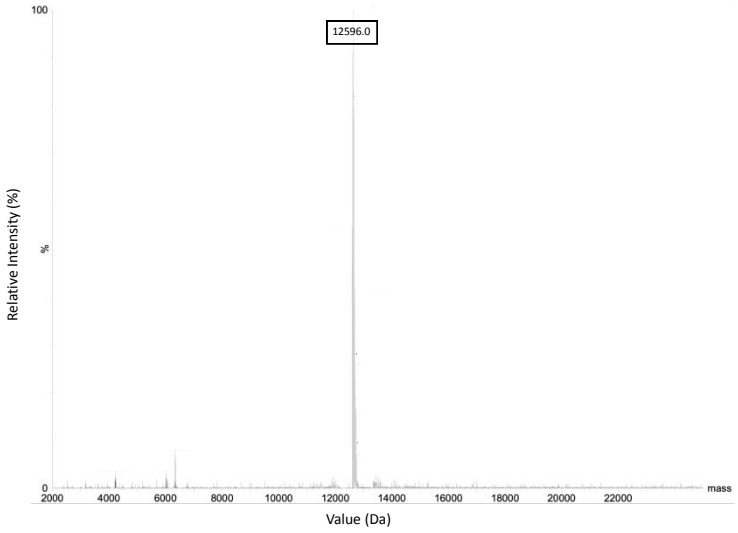
Sequence Name:	C3
Sequence:	5'-/5Biosg/TTT TTT TTT TTT CCC CTT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12439.2
Measured Molecular Weight:	12440.60

Figure S10. ESI-MS plot of Control C.



Sequence Name: C4
Sequence: 5'-/5Biosg/TTT TTT TTT TTT GGG GTT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight: 12599.3
Measured Molecular Weight: 12600.20

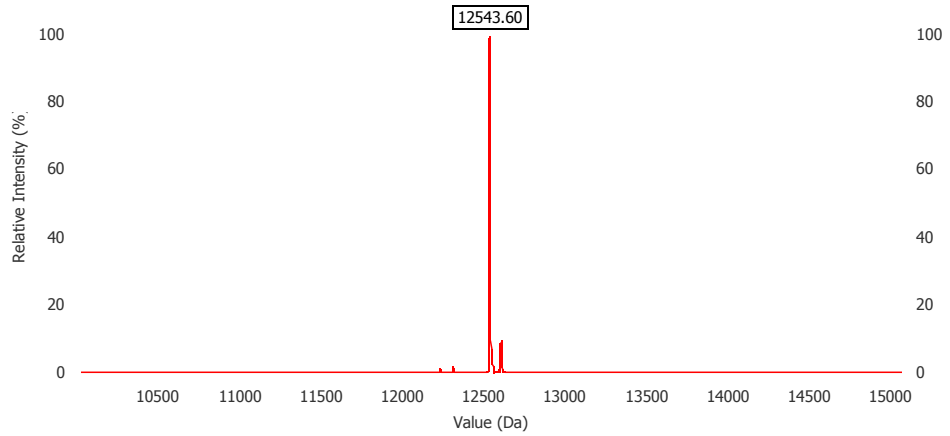
Figure S11. ESI-MS plot of Control G.



Sequence Name: C5
Sequence: 5'-/5Biosg/TTT TTT TTT TTT /i6diPr//i6diPr//i6diPr//i6diPr/
TTT TTT TTT TTT TTT TTT TTT TTT- 3'
Calculated Molecular Weight: 12595.4

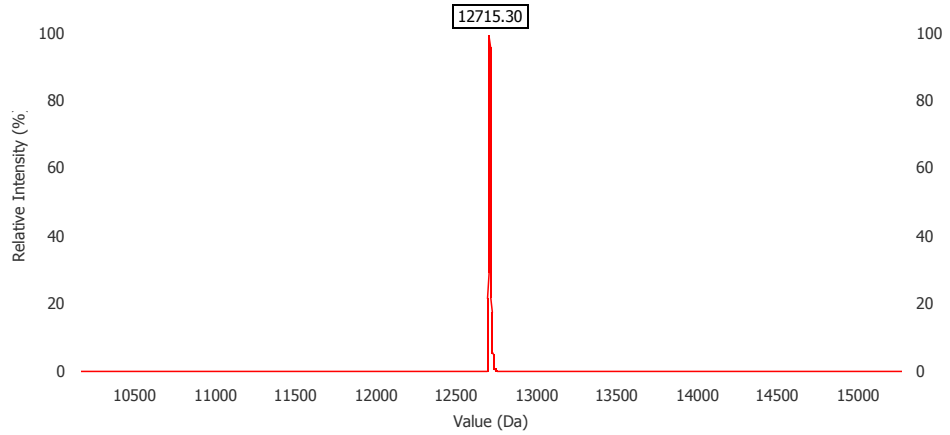
Measured Molecular Weight: 12596.0

Figure S12. ESI-MS plot of B1.



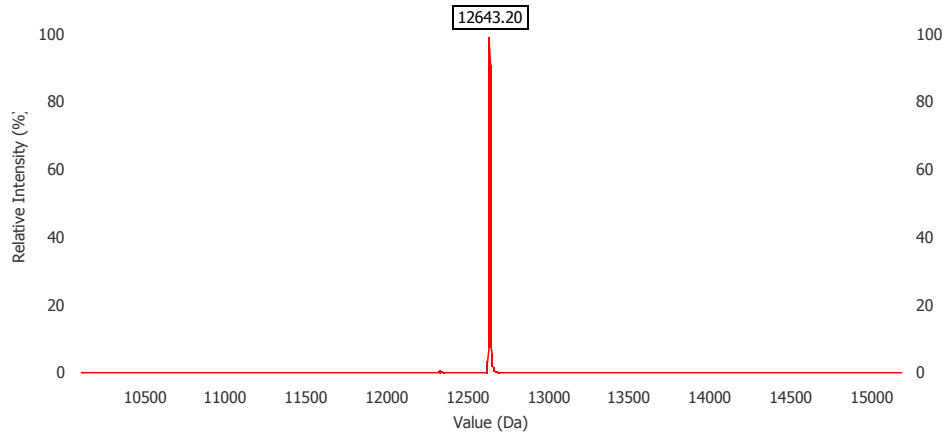
Sequence Name:	C6
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//i5HydMe-dC/ TTT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12544.3
Measured Molecular Weight:	12543.60

Figure S13. ESI-MS plot of B2.



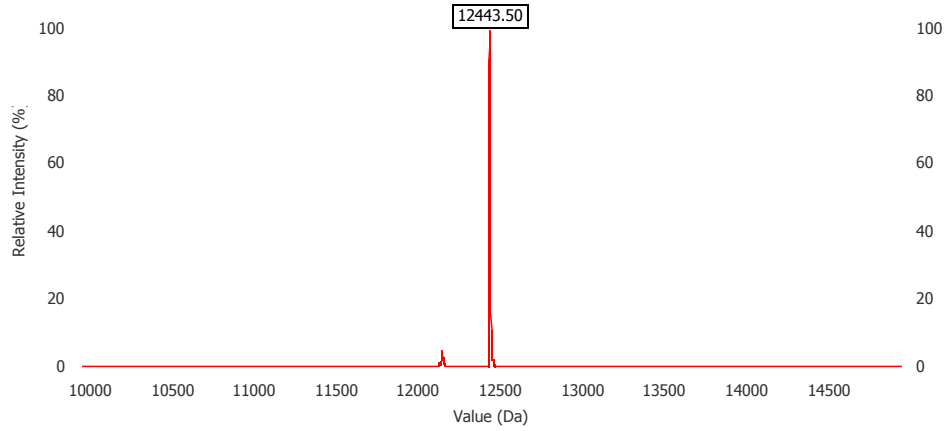
Sequence Name:	C7
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12715.5
Measured Molecular Weight:	12715.30

Figure S14. ESI-MS plot of B3.



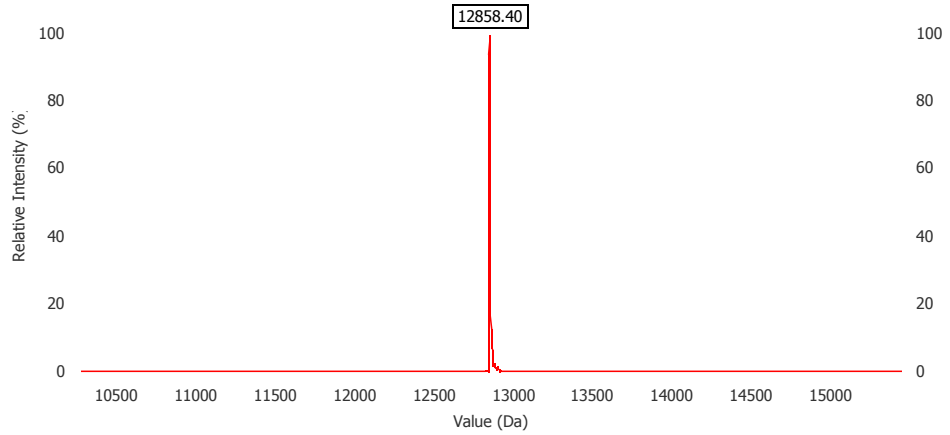
Sequence Name:	C8
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//i5NitInd//i5NitInd//i5NitInd/TTT TTT TTT TTT TTT TTT TTT TTT -3'
Calculated Molecular Weight:	12643.4
Measured Molecular Weight:	12643.20

Figure S15. ESI-MS plot of B4.



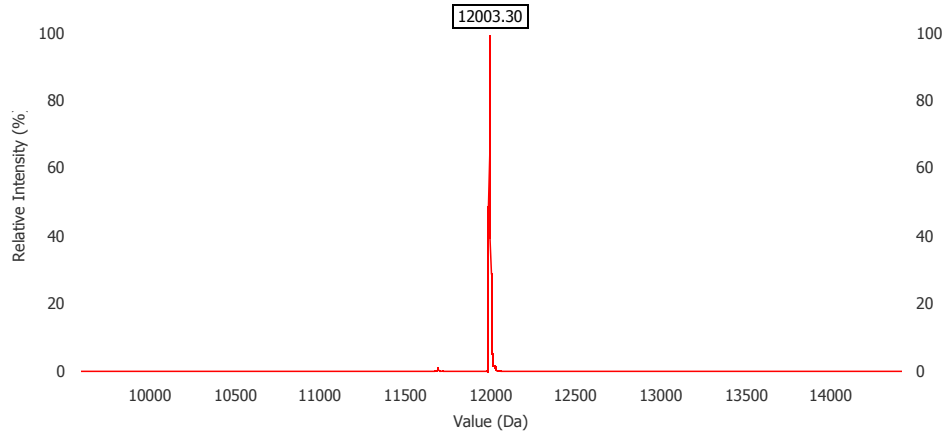
Sequence Name:	C9
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12443.2
Measured Molecular Weight:	12443.50

Figure S16. ESI-MS plot of B5.



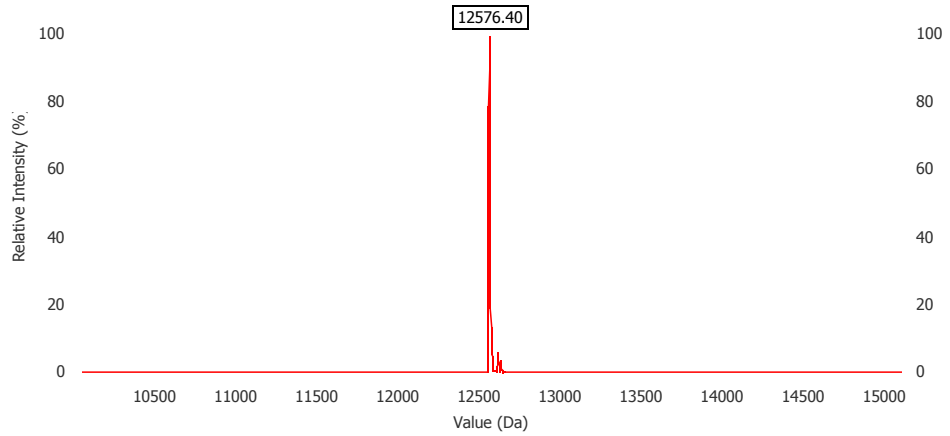
Sequence Name:	KT_Click Control
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5OctdU//i5OctdU//i5OctdU/ /i5OctdU/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12859.8
Measured Molecular Weight:	12858.40

Figure S17. ESI-MS plot of B6.



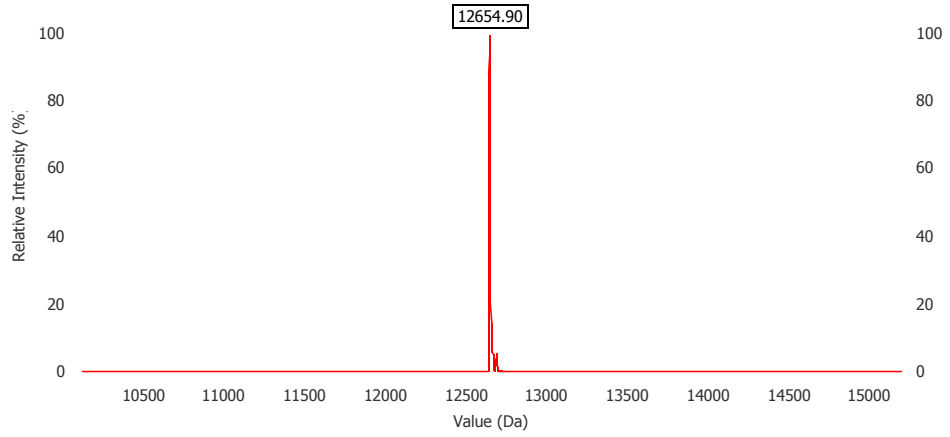
Sequence Name:	Abasic control
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /idSp//idSp//idSp//idSp/TTT TTT TTT TTT TTT TTT TTT TTT -3'
Calculated Molecular Weight:	12002.9
Measured Molecular Weight:	12003.30

Figure S18. ESI-MS plot of B7.



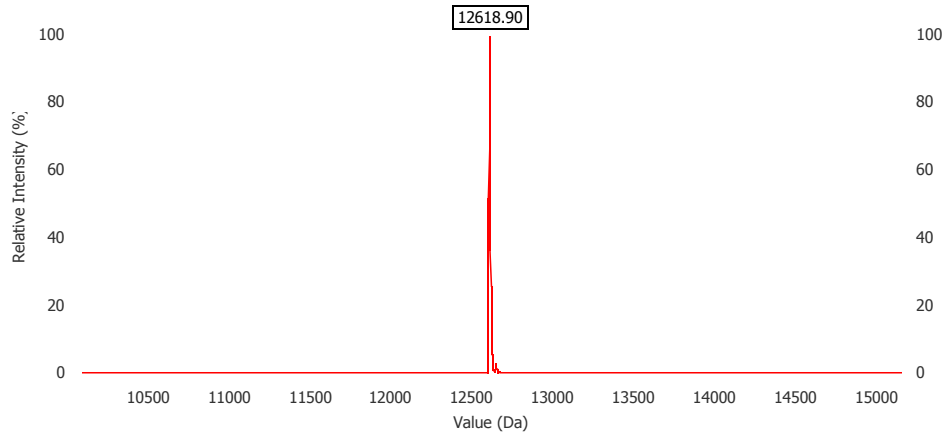
Sequence Name:	C10
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i6diPr//i6diPr//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12577.4
Measured Molecular Weight:	12576.40

Figure S19. ESI-MS plot of 1122.



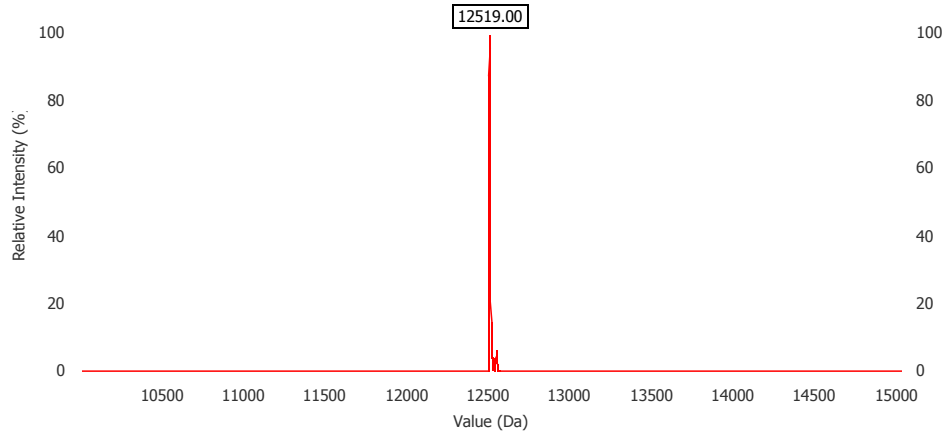
Sequence Name:	C11
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i6diPr//i6diPr//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12655.4
Measured Molecular Weight:	12654.90

Figure S20. ESI-MS plot of 1133.



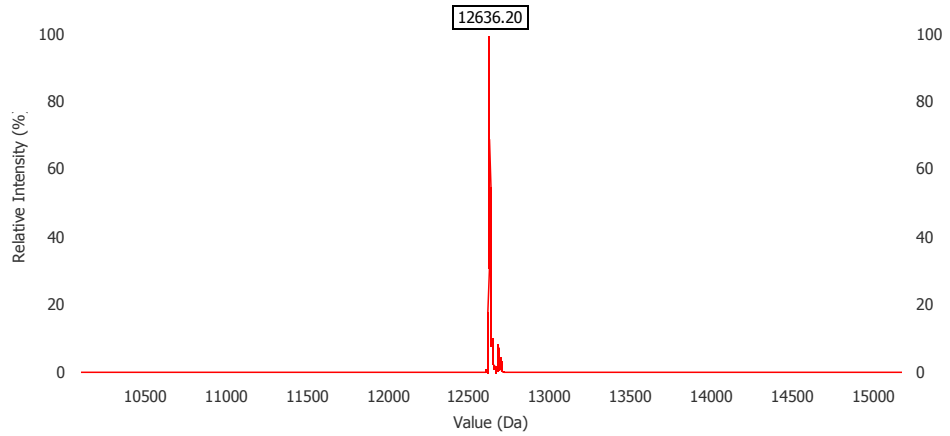
Sequence Name:	C12
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i6diPr//i6diPr//i5NitInd//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12619.4
Measured Molecular Weight:	12618.90

Figure S21. ESI-MS plot of 1144.



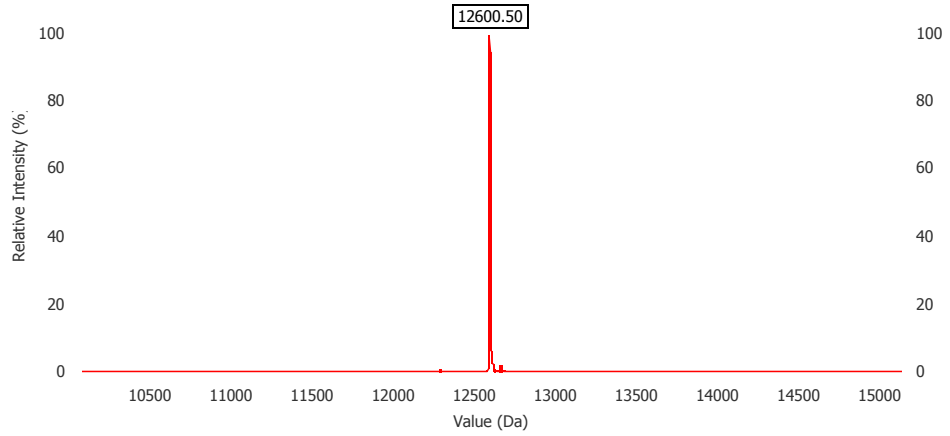
Sequence Name:	C13
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i6diPr//i6diPr//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12519.3
Measured Molecular Weight:	12519.00

Figure S22. ESI-MS plot of 1155.



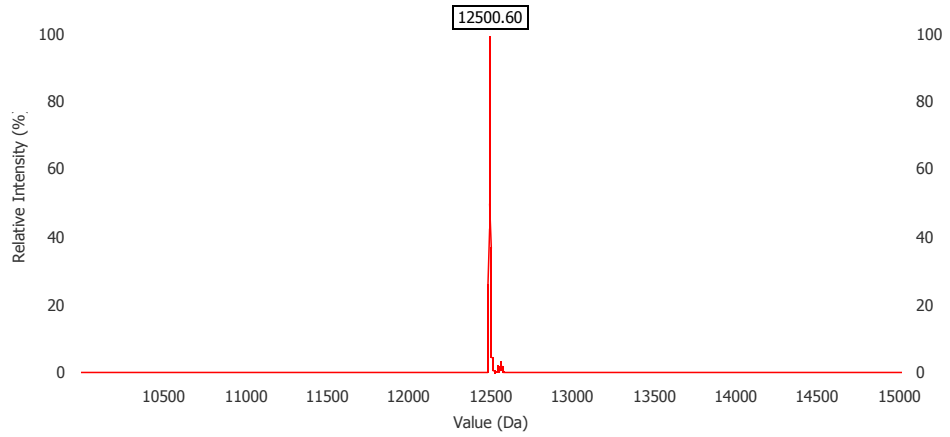
Sequence Name:	C14
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12637.4
Measured Molecular Weight:	12636.20

Figure S23. ESI-MS plot of 2233.



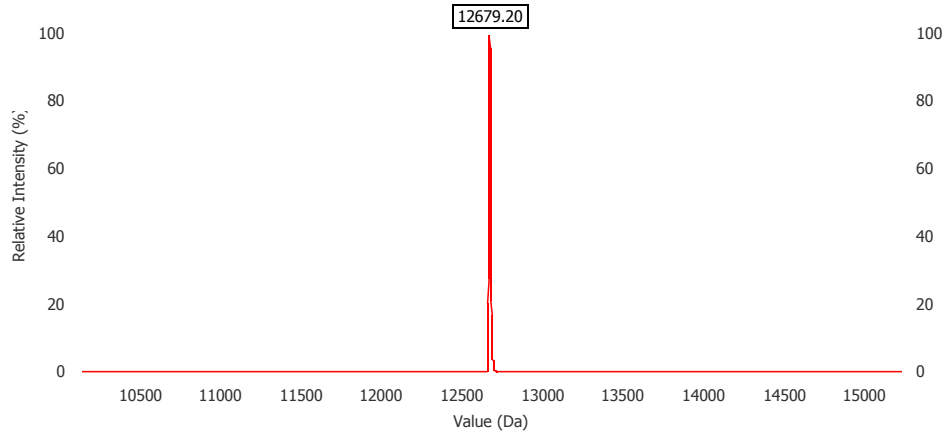
Sequence Name:	C15
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//i5NitInd//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12601.4
Measured Molecular Weight:	12600.50

Figure S24. ESI-MS plot of 2244.



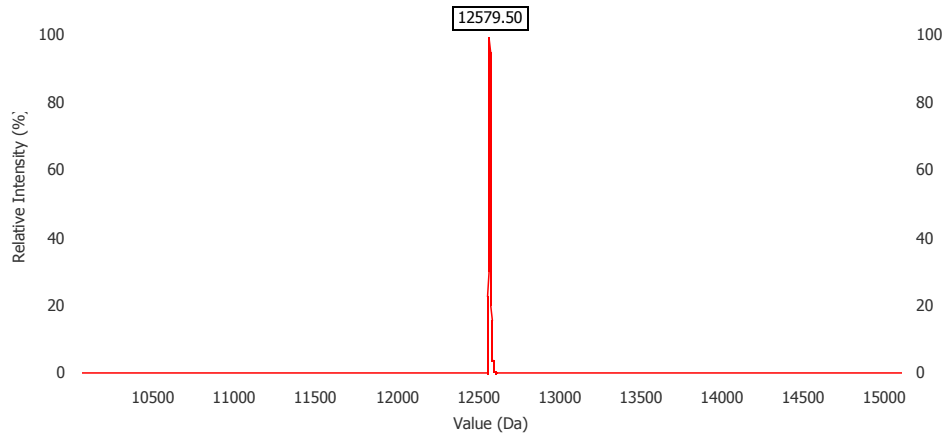
Sequence Name:	C16
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12501.2
Measured Molecular Weight:	12500.60

Figure S25. ESI-MS plot of 2255.



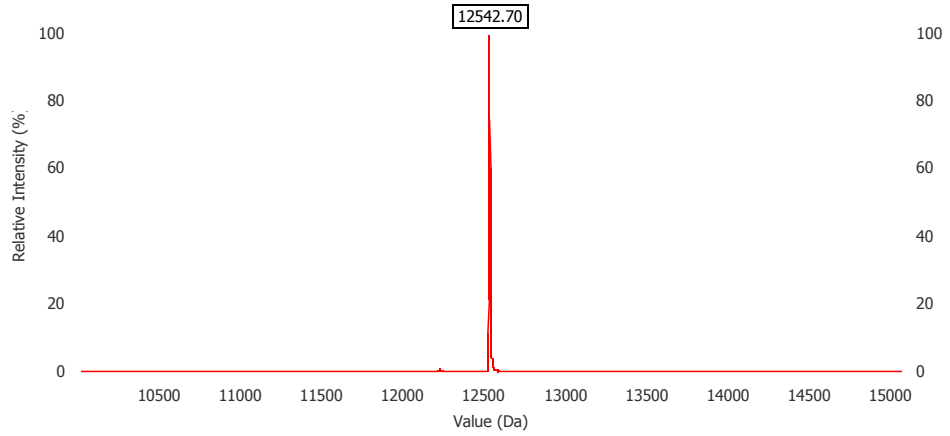
Sequence Name:	C17
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//i5NitInd//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT-3'
Calculated Molecular Weight:	12679.4
Measured Molecular Weight:	12679.20

Figure S26. ESI-MS plot of 3344.



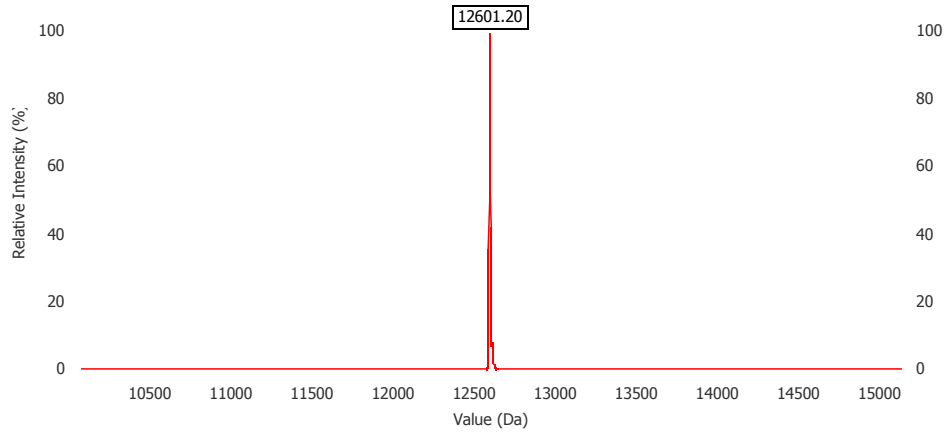
Sequence Name:	C18
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12579.3
Measured Molecular Weight:	12579.50

Figure S27. ESI-MS plot of 3355.



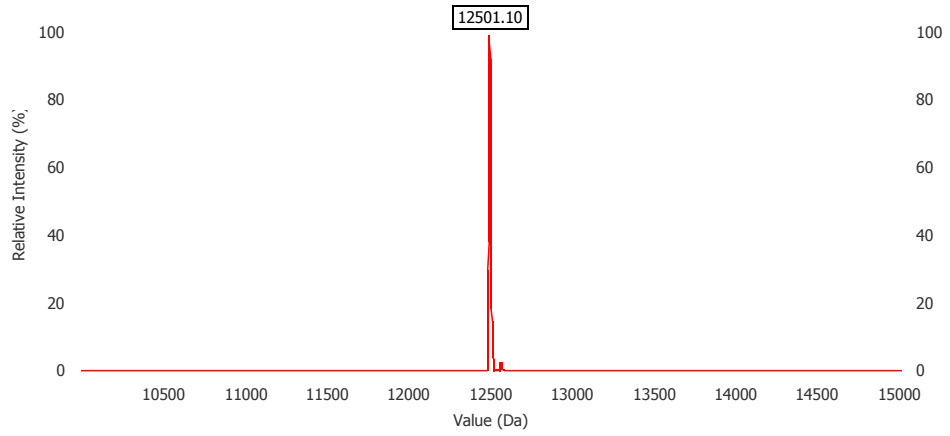
Sequence Name:	C19
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//i5NitInd//ideoxyU//ideoxyU/T TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12543.3
Measured Molecular Weight:	12542.70

Figure S28. ESI-MS plot of 4455.



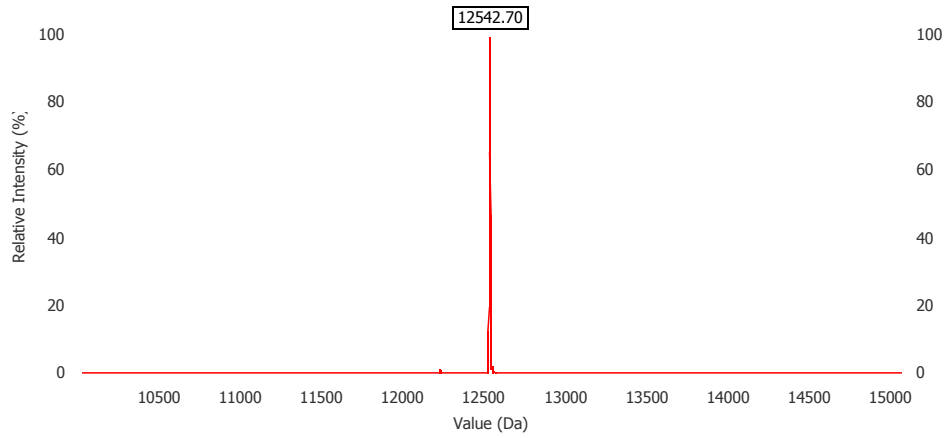
Sequence Name:	C8866
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//i5NitInd//i5HydMe-dC//i5HydMe-dC/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12601.4
Measured Molecular Weight:	12601.20

Figure S29. ESI-MS plot of 4422.



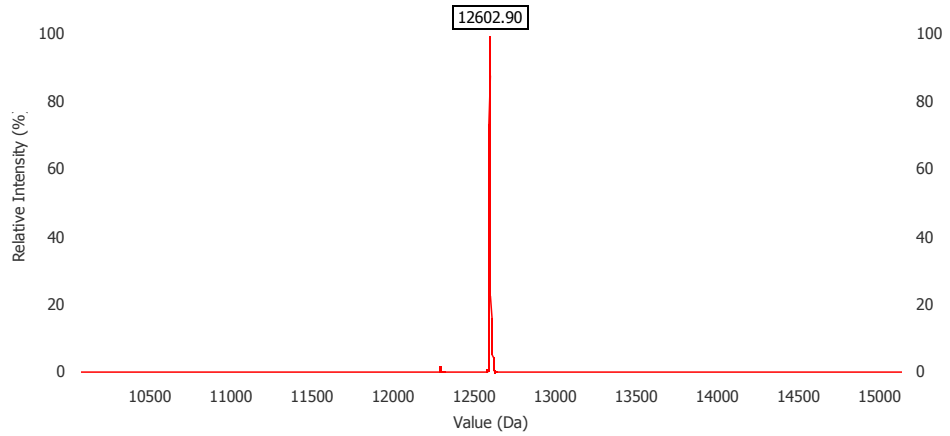
Sequence Name:	C9966
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12501.2
Measured Molecular Weight:	12501.10

Figure S30. ESI-MS plot of 5522.



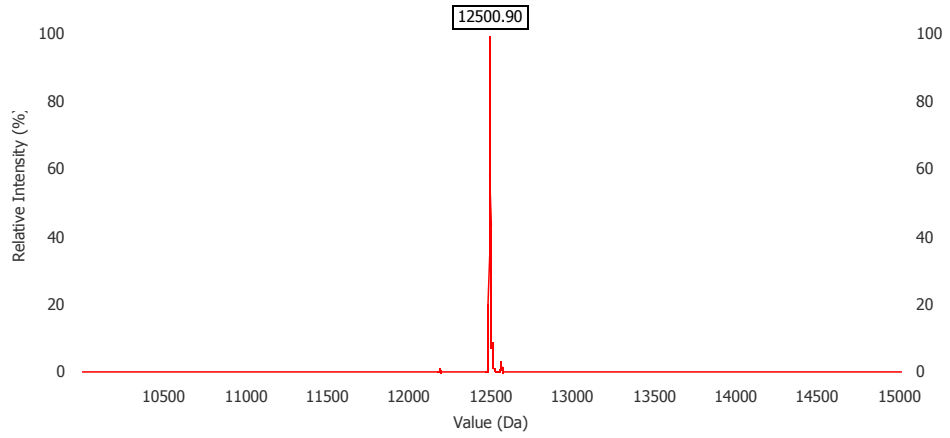
Sequence Name:	C9988
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//i5NitInd//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12543.3
Measured Molecular Weight:	12542.70

Figure S31. ESI-MS plot of 5544.



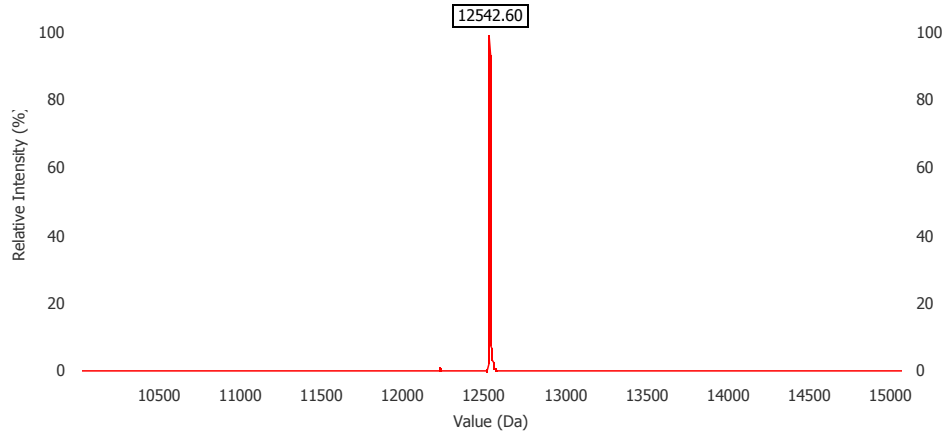
Sequence Name:	C6886
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5NitInd//i5NitInd//i5HydMe-dC/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12601.4
Measured Molecular Weight:	12602.90

Figure S32. ESI-MS plot of 2442.



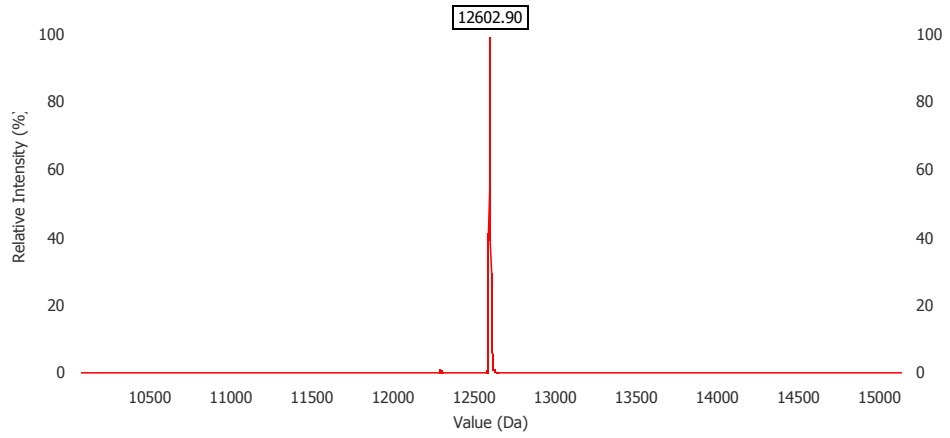
Sequence Name:	C6996
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//ideoxyU//ideoxyU/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12501.2
Measured Molecular Weight:	12500.90

Figure S33. ESI-MS plot of 2552.



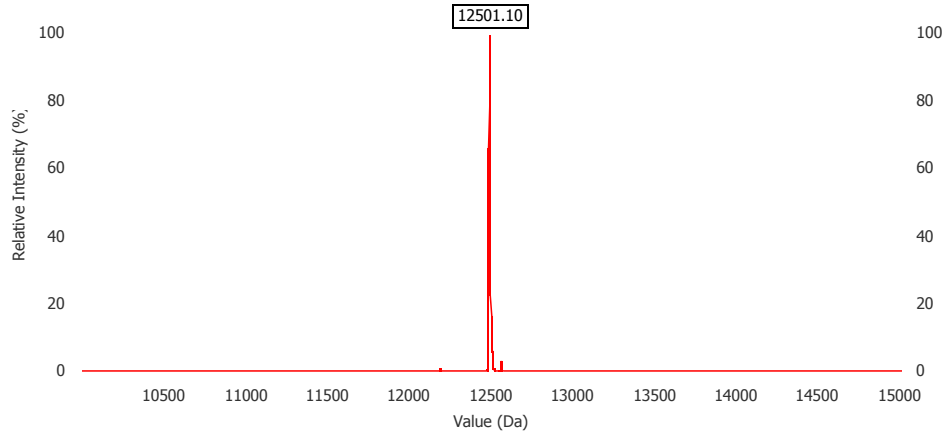
Sequence Name:	C8998
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//ideoxyU//ideoxyU//i5NitInd/T TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12543.3
Measured Molecular Weight:	12542.60

Figure S34. ESI-MS plot of 4554.



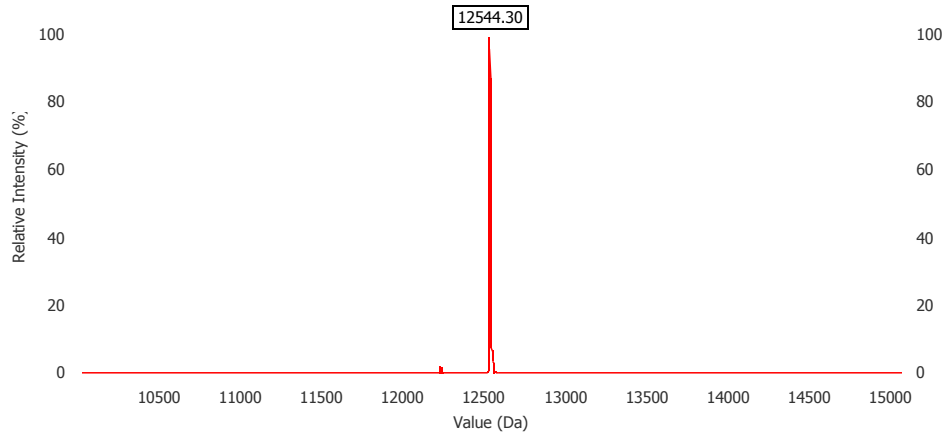
Sequence Name:	C8668
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//i5HydMe-dC//i5HydMe-dC//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12601.4
Measured Molecular Weight:	12602.90

Figure S35. ESI-MS plot of 4224.



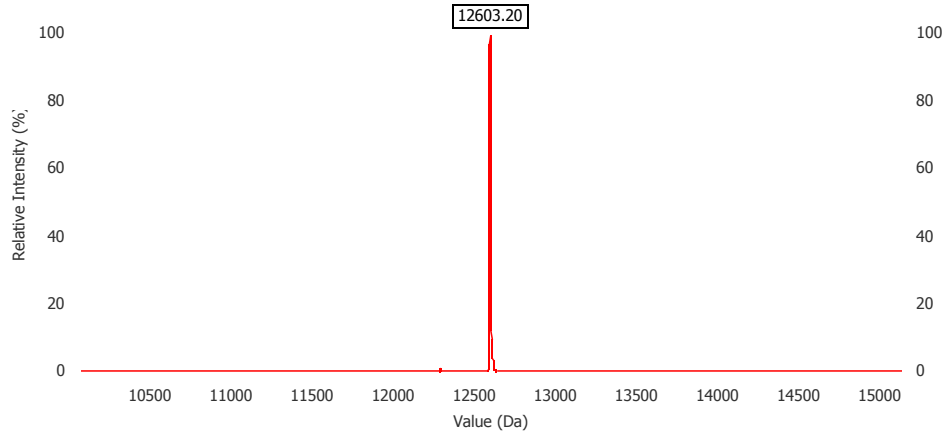
Sequence Name:	C9669
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//i5HydMe-dC//i5HydMe-dC/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12501.2
Measured Molecular Weight:	12501.10

Figure S36. ESI-MS plot of 5225.



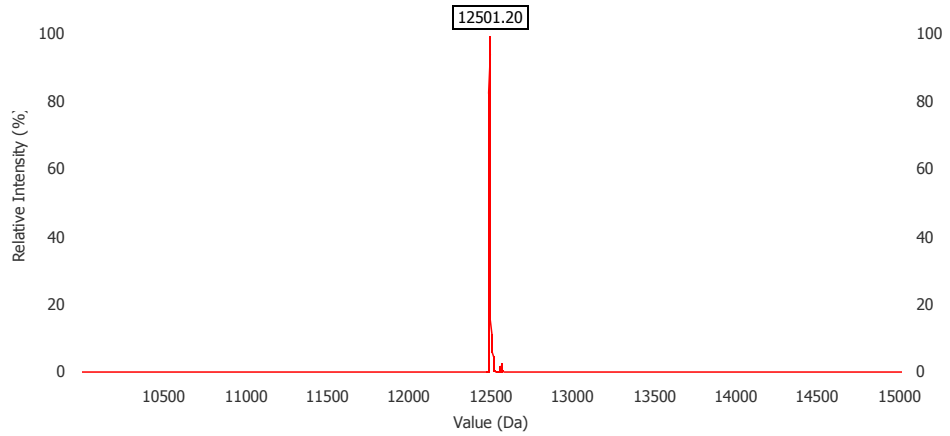
Sequence Name:	C9889
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//i5NitInd//i5NitInd//ideoxyU/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12543.3
Measured Molecular Weight:	12544.30

Figure S37. ESI-MS plot of 5445.



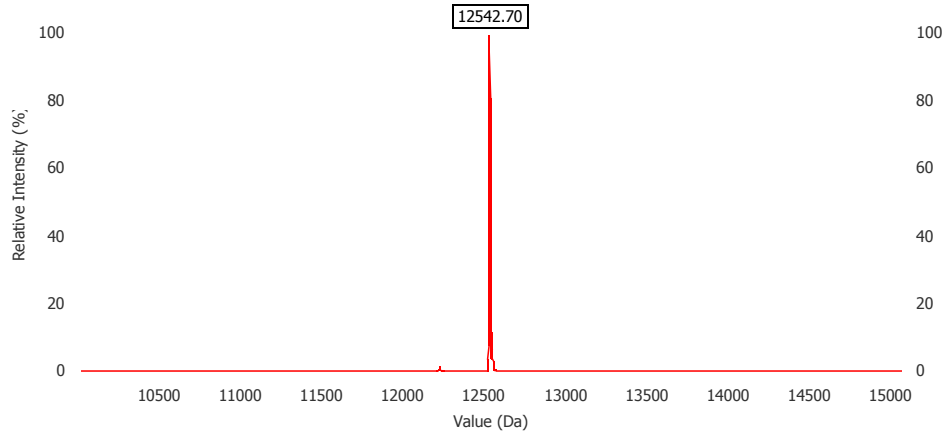
Sequence Name:	C6868
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5NitInd//i5HydMe-dC//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12601.4
Measured Molecular Weight:	12603.20

Figure S38. ESI-MS plot of 2424.



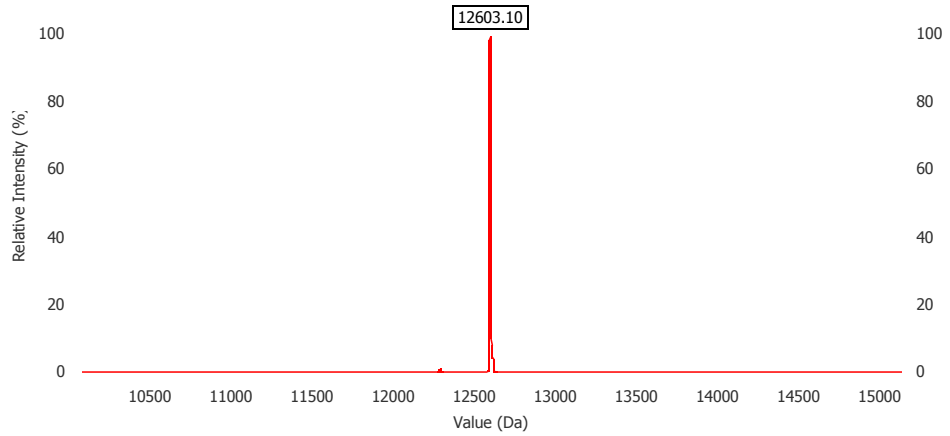
Sequence Name:	C6969
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//ideoxyU//i5HydMe-dC/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12501.2
Measured Molecular Weight:	12501.20

Figure S39. ESI-MS plot of 2525.



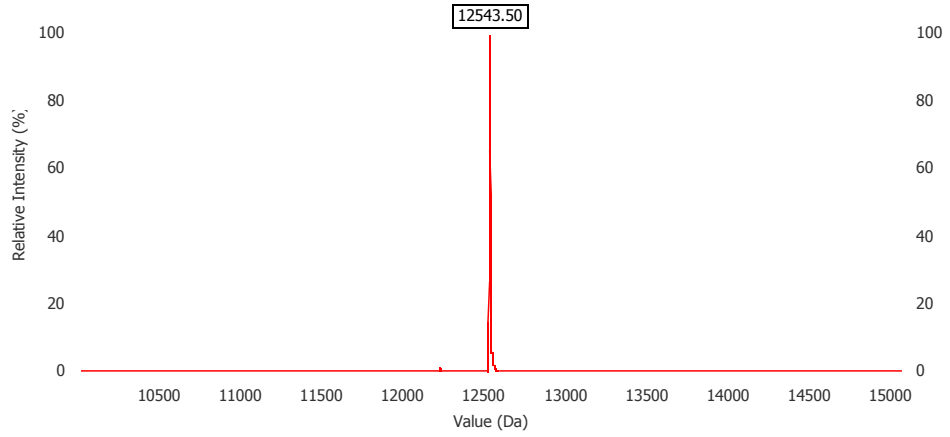
Sequence Name:	C8989
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//ideoxyU//i5NitInd//ideoxyU/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12543.3
Measured Molecular Weight:	12542.70

Figure S40. ESI-MS plot of 4545.



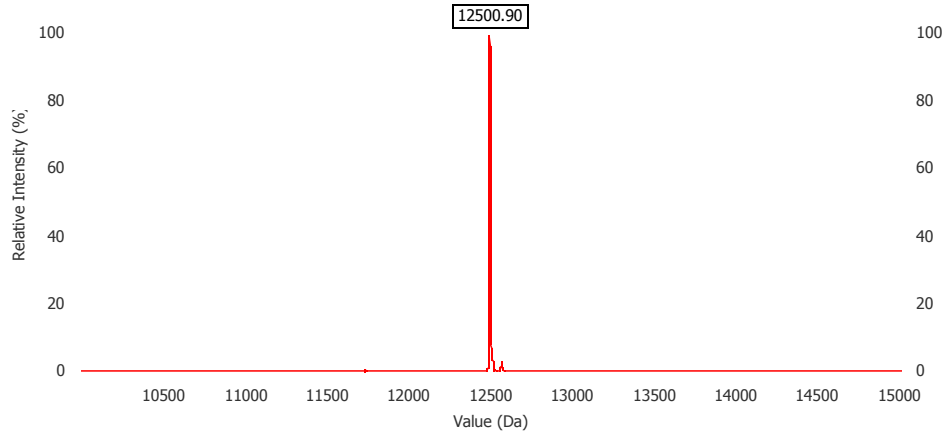
Sequence Name:	C8686
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5NitInd//i5HydMe-dC//i5NitInd//i5HydMe-dC/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12601.4
Measured Molecular Weight:	12603.10

Figure S41. ESI-MS plot of 4242.



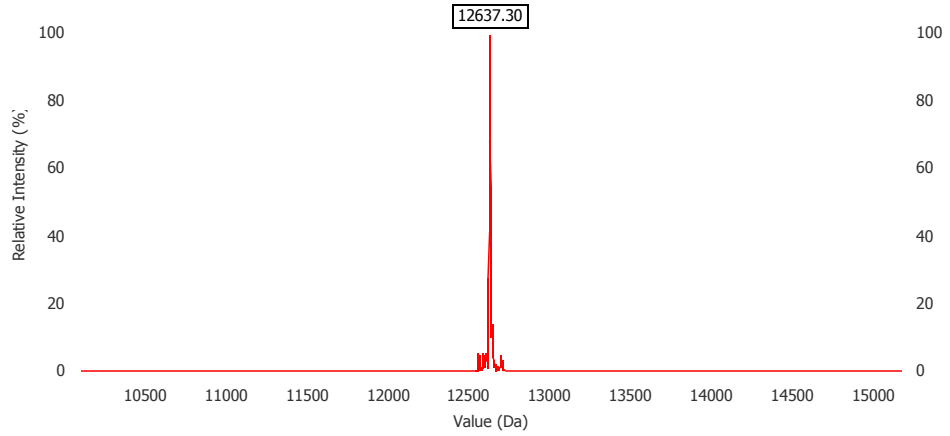
Sequence Name:	C9898
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//i5NitInd//ideoxyU//i5NitInd/T TTT TTT TTT TTT TTT TTT TTT TT -3'
Calculated Molecular Weight:	12543.3
Measured Molecular Weight:	12543.50

Figure S42. ESI-MS plot of 5454.



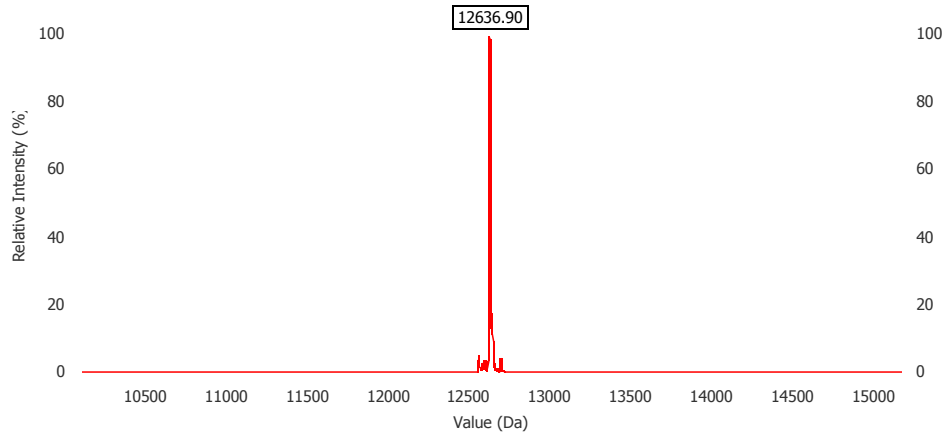
Sequence Name:	C9696
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//i5HydMe-dC//ideoxyU/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12501.2
Measured Molecular Weight:	12500.90

Figure S43. ESI-MS plot of 5252.



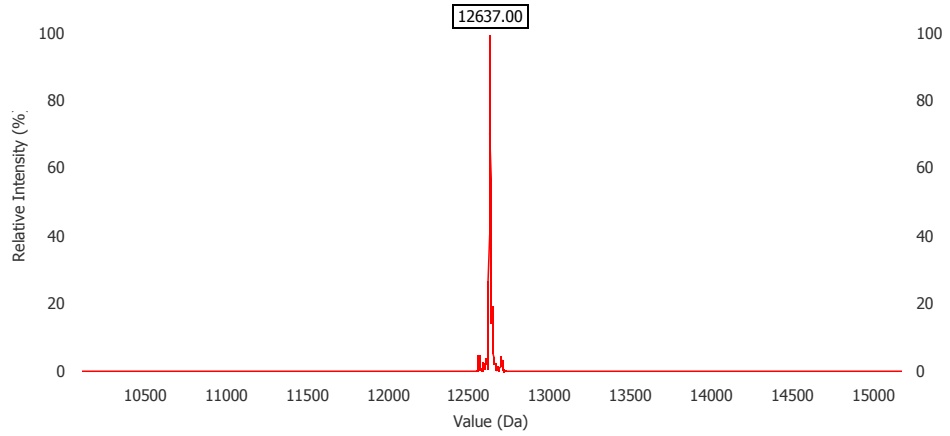
Sequence Name:	C7766
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12637.4
Measured Molecular Weight:	12637.30

Figure S44. ESI-MS plot of 3322.



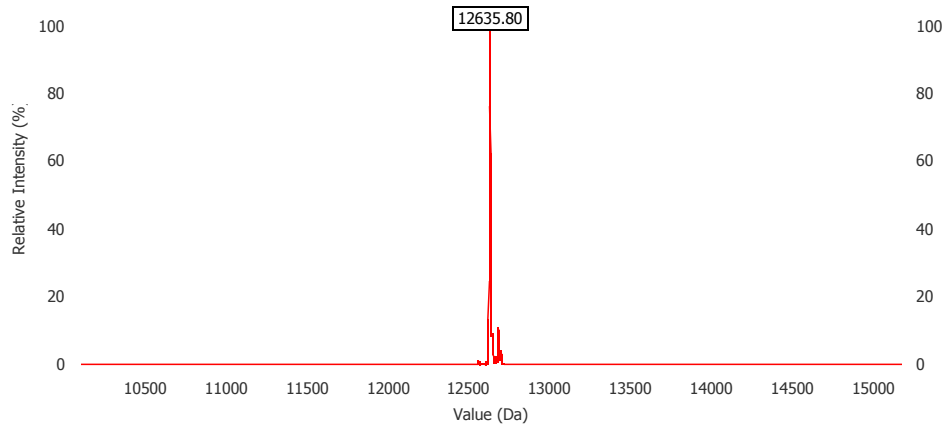
Sequence Name:	C6776
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//iSuper-dT//iSuper-dT/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12637.4
Measured Molecular Weight:	12636.90

Figure S45. ESI-MS plot of 2332.



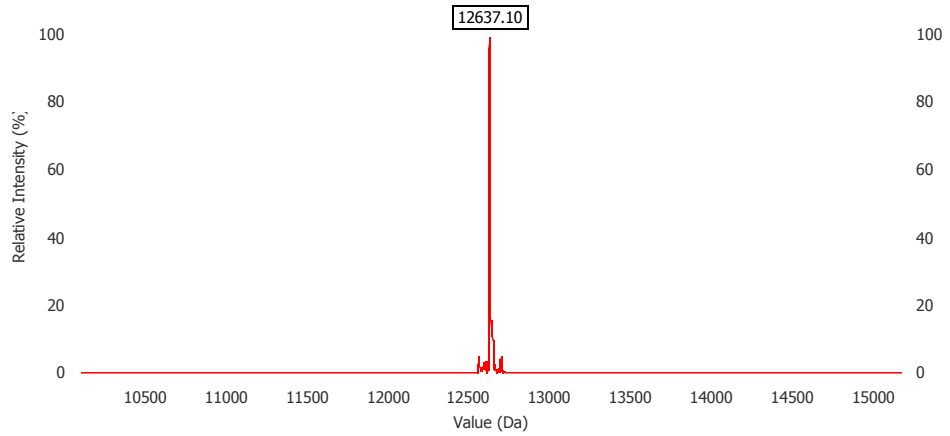
Sequence Name:	C7667
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//i5HydMe-dC//i5HydMe-dC/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12637.4
Measured Molecular Weight:	12637.00

Figure S46. ESI-MS plot of 3223.



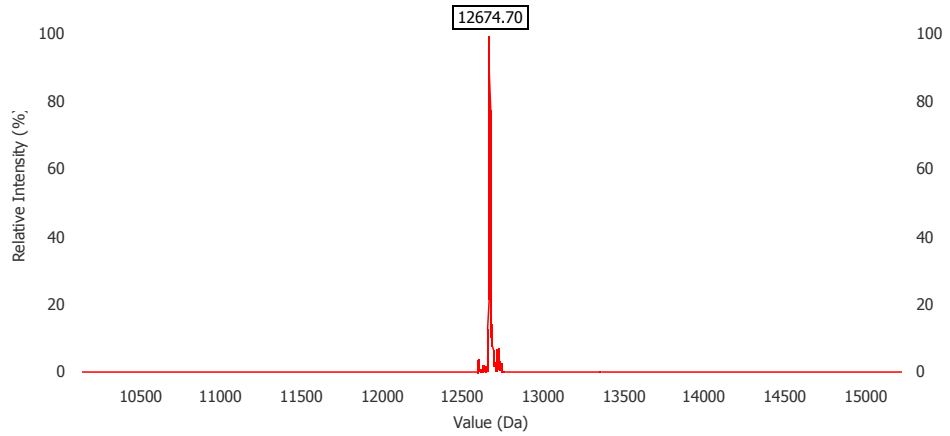
Sequence Name:	C6767
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//iSuper-dT//i5HydMe-dC/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12637.4
Measured Molecular Weight:	12635.80

Figure S47. ESI-MS plot of 2323.



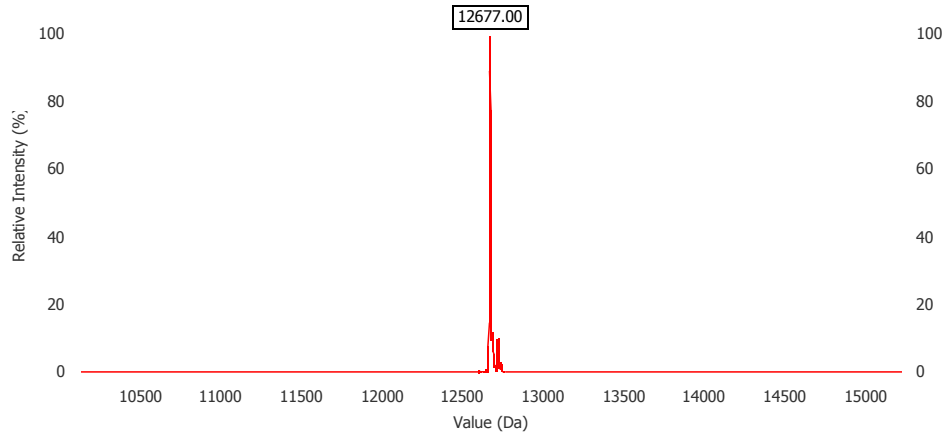
Sequence Name:	C7676
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//i5HydMe-dC//iSuper-dT/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12637.4
Measured Molecular Weight:	12637.10

Figure S48. ESI-MS plot of 3232.



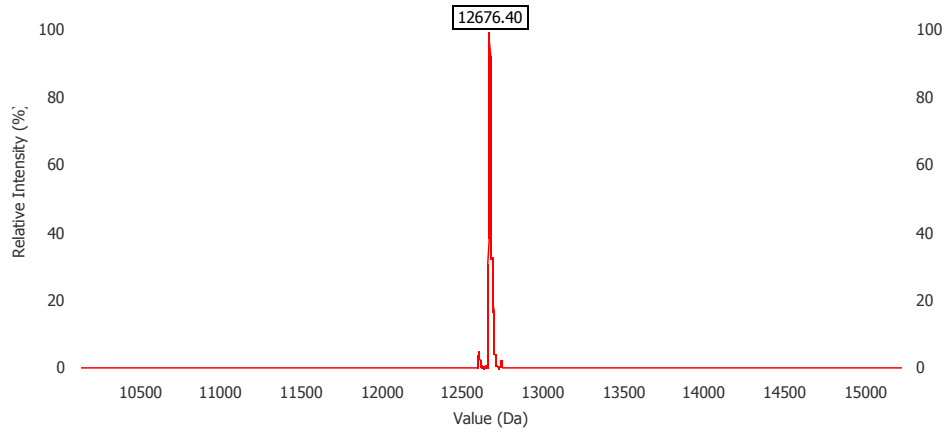
Sequence Name:	C6777
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//iSuper-dT//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12676.4
Measured Molecular Weight:	12674.70

Figure S49. ESI-MS plot of 2333.



Sequence Name:	C7677
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//i5HydMe-dC//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12676.4
Measured Molecular Weight:	12677.00

Figure S50. ESI-MS plot of 3233.



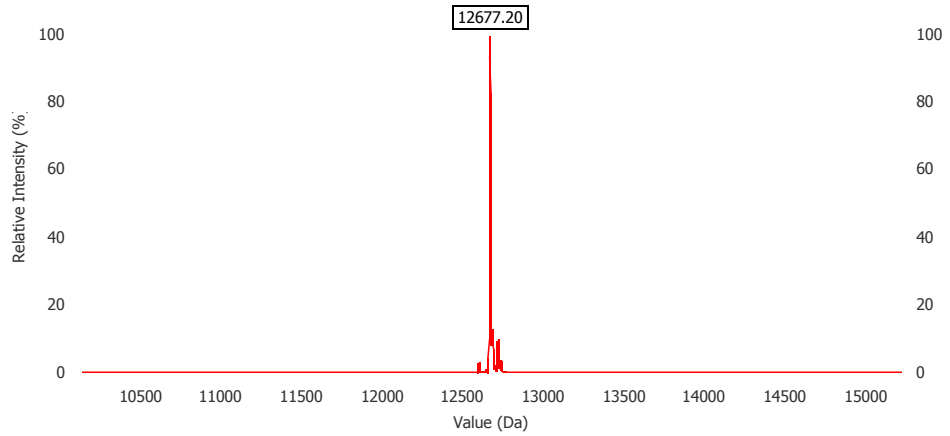
Sequence Name: C7767

Sequence: 5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//i5HydMe-dC/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'

Calculated Molecular Weight: 12676.4

Measured Molecular Weight: 12676.40

Figure S51. ESI-MS plot of 3323.



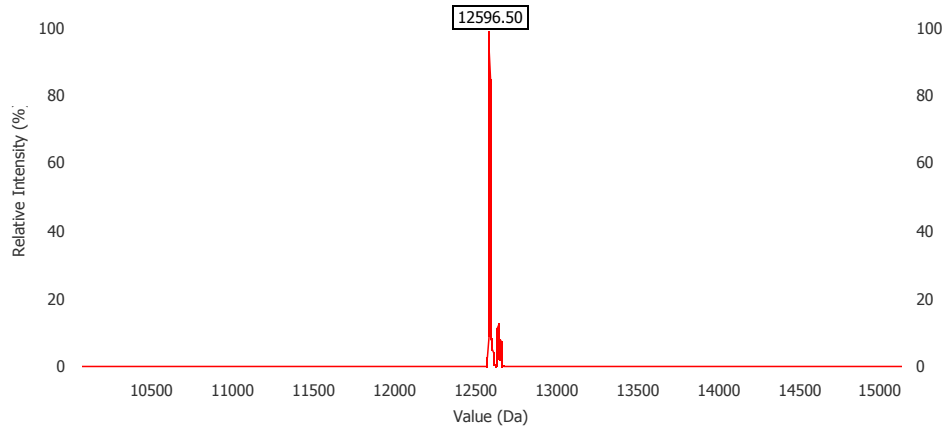
Sequence Name: C7776

Sequence: 5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//iSuper-dT/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T -3'

Calculated Molecular Weight: 12676.4

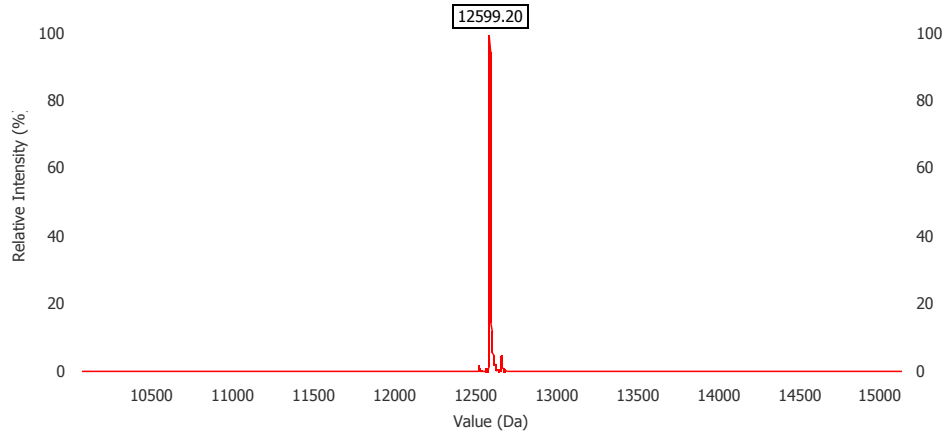
Measured Molecular Weight: 12677.20

Figure S52. ESI-MS plot of 3332.



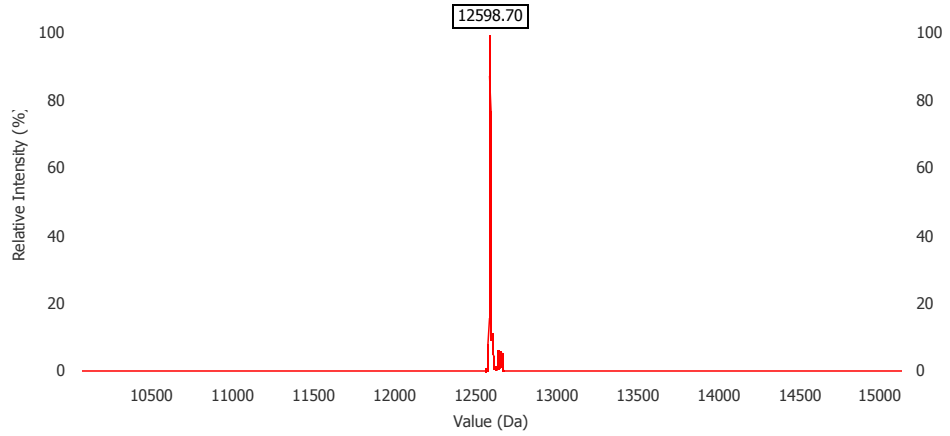
Sequence Name:	C7666
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//i5HydMe-dC//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12598.4
Measured Molecular Weight:	12596.50

Figure S53. ESI-MS plot of 3222.



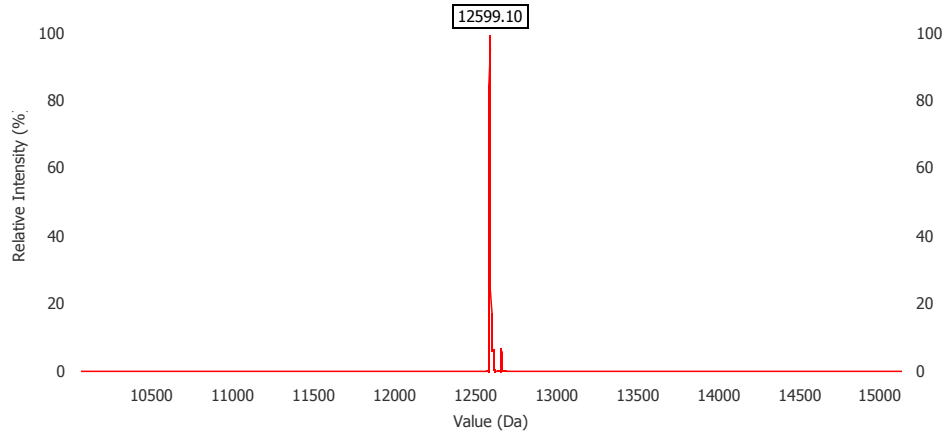
Sequence Name:	C6766
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//iSuper-dT//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12598.4
Measured Molecular Weight:	12599.20

Figure S54. ESI-MS plot of 2322.



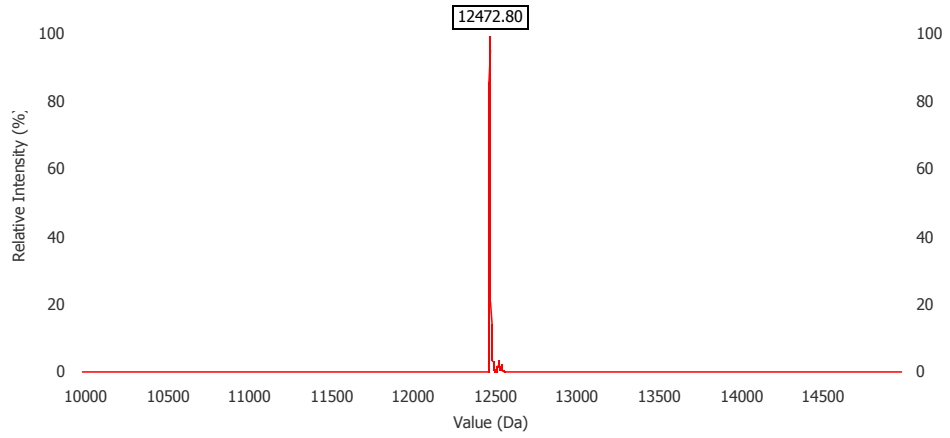
Sequence Name:	C6676
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//iSuper-dT/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12598.4
Measured Molecular Weight:	12598.70

Figure S55. ESI-MS plot of 2232.



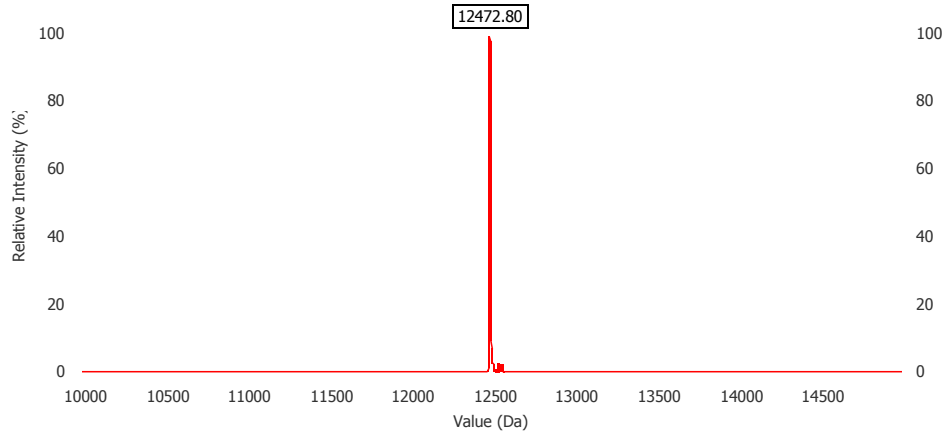
Sequence Name:	C6667
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//i5HydMe-dC/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12598.4
Measured Molecular Weight:	12599.10

Figure S56. ESI-MS plot of 2223.



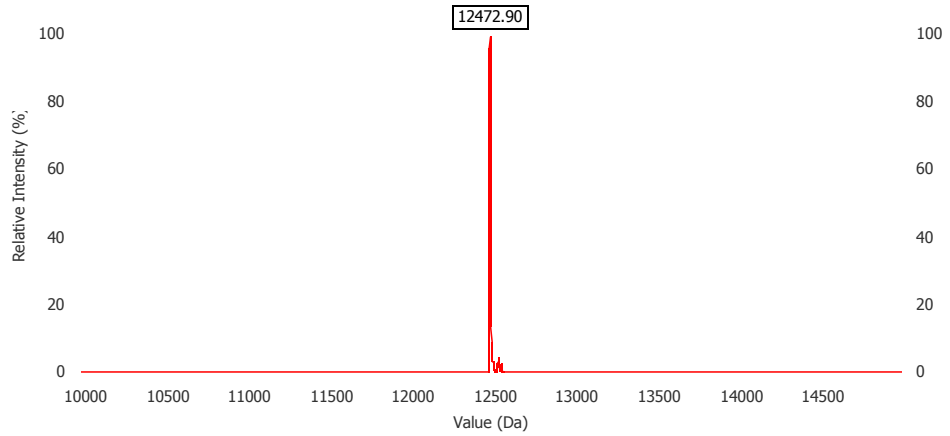
Sequence Name:	C6999
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//ideoxyU//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12472.2
Measured Molecular Weight:	12472.80

Figure S57. ESI-MS plot of 2555.



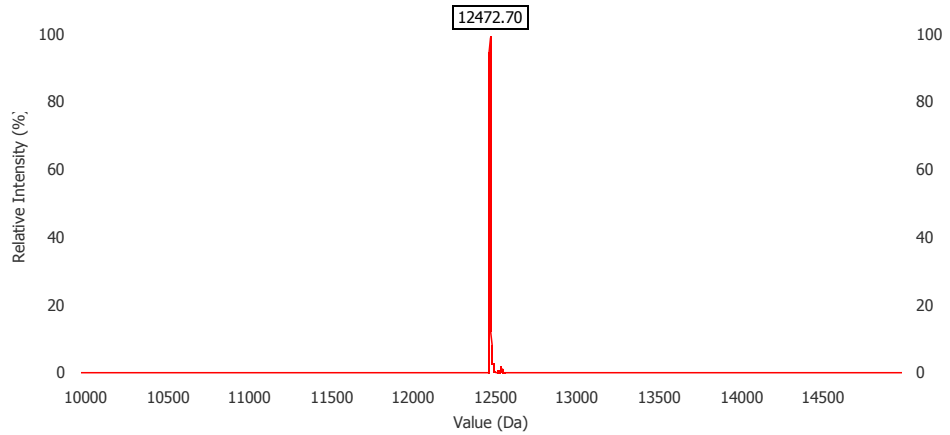
Sequence Name:	C9699
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//i5HydMe-dC//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12472.2
Measured Molecular Weight:	12472.80

Figure S58. ESI-MS plot of 5255.



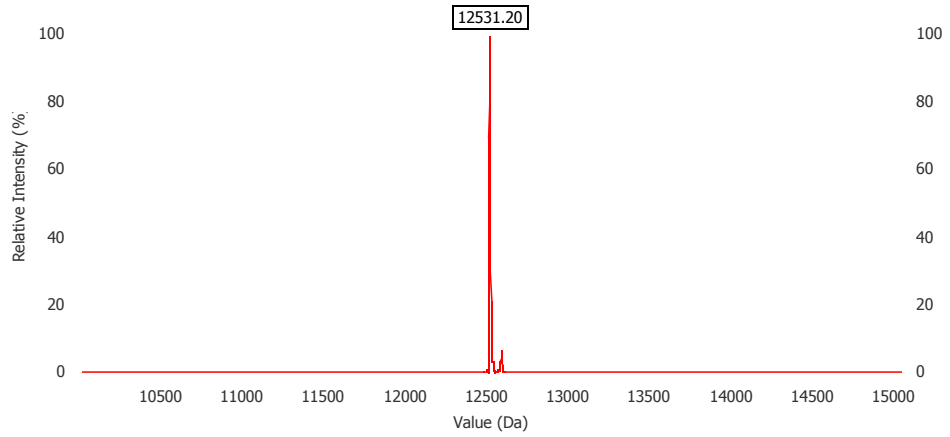
Sequence Name:	C9969
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//i5HydMe-dC/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12472.2
Measured Molecular Weight:	12472.90

Figure S59. ESI-MS plot of 5525.



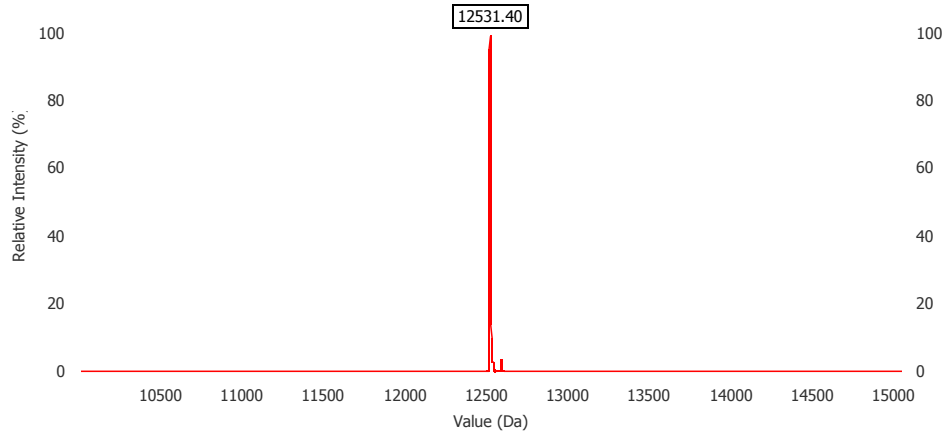
Sequence Name:	C9996
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//ideoxyU/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12472.2
Measured Molecular Weight:	12472.70

Figure S60. ESI-MS plot of 5552.



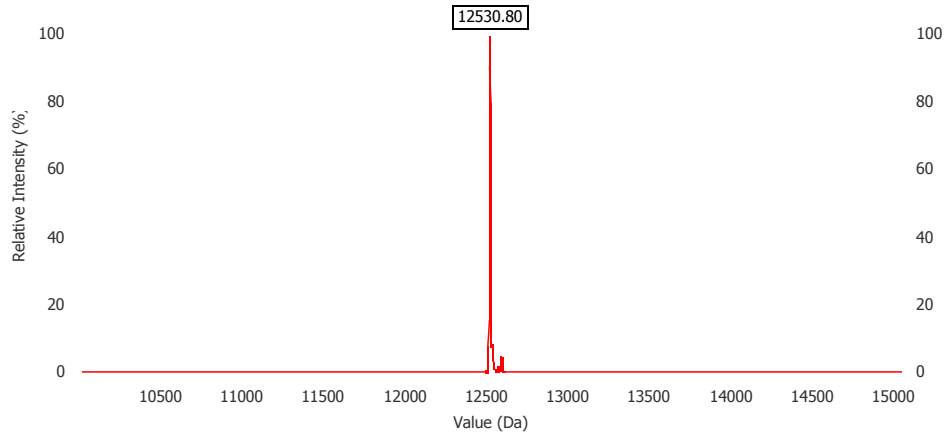
Sequence Name:	C9666
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//i5HydMe-dC//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12530.3
Measured Molecular Weight:	12531.20

Figure S61. ESI-MS plot of 5222.



Sequence Name:	C6966
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//ideoxyU//i5HydMe-dC/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12530.3
Measured Molecular Weight:	12531.40

Figure S62. ESI-MS plot of 2522.



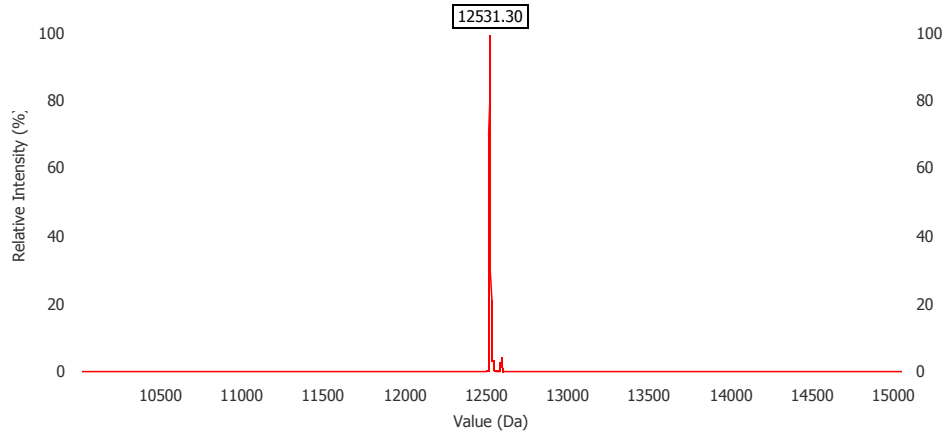
Sequence Name: C6696

Sequence: 5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//ideoxyU/ /i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT TTT T -3'

Calculated Molecular Weight: 12530.3

Measured Molecular Weight: 12530.80

Figure S63. ESI-MS plot of 2252.



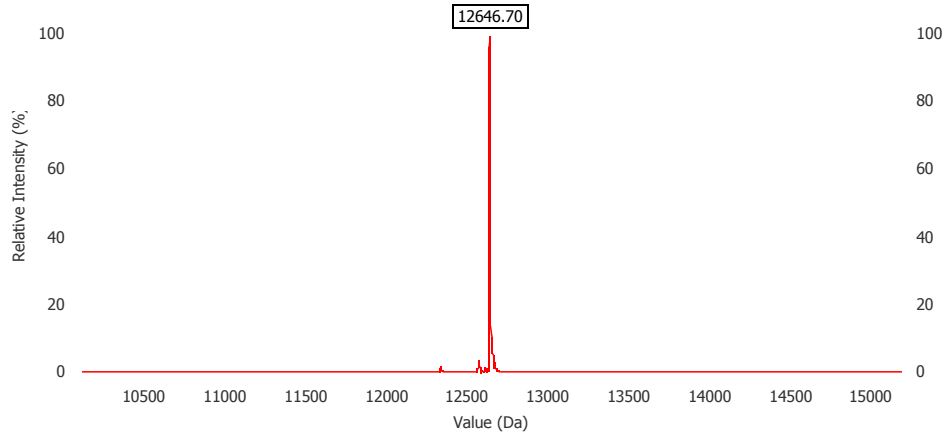
Sequence Name: C6669

Sequence: 5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC//i5HydMe-dC//i5HydMe-dC/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'

Calculated Molecular Weight: 12530.3

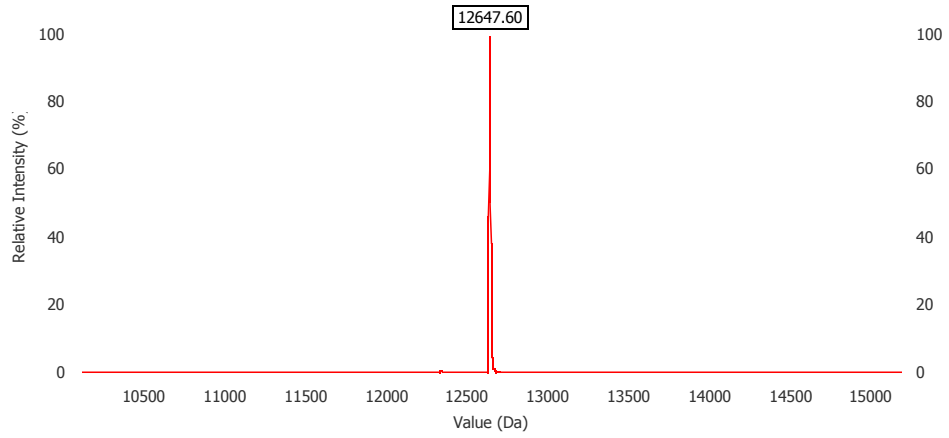
Measured Molecular Weight: 12531.30

Figure S64. ESI-MS plot of 2225.



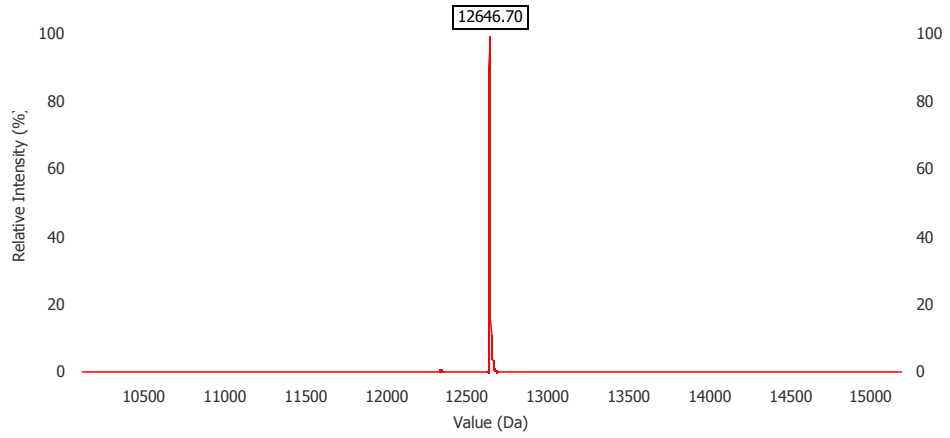
Sequence Name:	C9777
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//iSuper-dT//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12647.4
Measured Molecular Weight:	12646.70

Figure S65. ESI-MS plot of 5333.



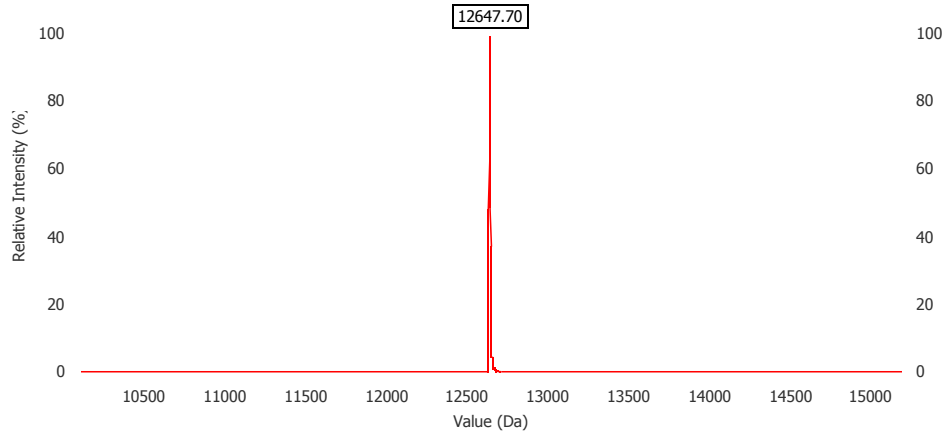
Sequence Name:	C7977
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//ideoxyU//iSuper-dT/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12647.4
Measured Molecular Weight:	12647.60

Figure S66. ESI-MS plot of 3533.



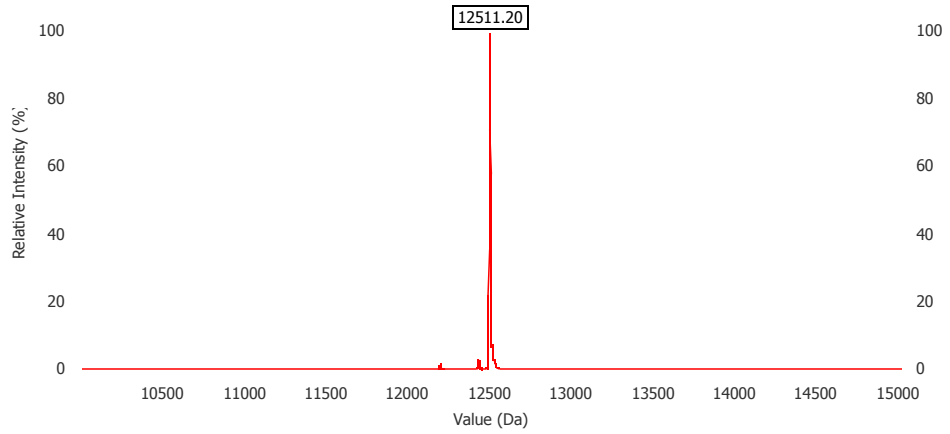
Sequence Name:	C7797
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//ideoxyU/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12647.4
Measured Molecular Weight:	12646.70

Figure S67. ESI-MS plot of 3353.



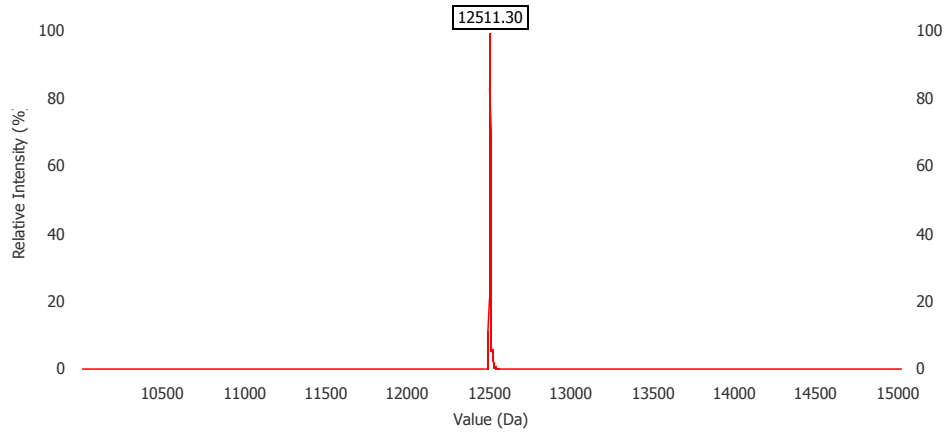
Sequence Name:	C7779
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//iSuper-dT//iSuper-dT/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12647.4
Measured Molecular Weight:	12647.70

Figure S68. ESI-MS plot of 3335.



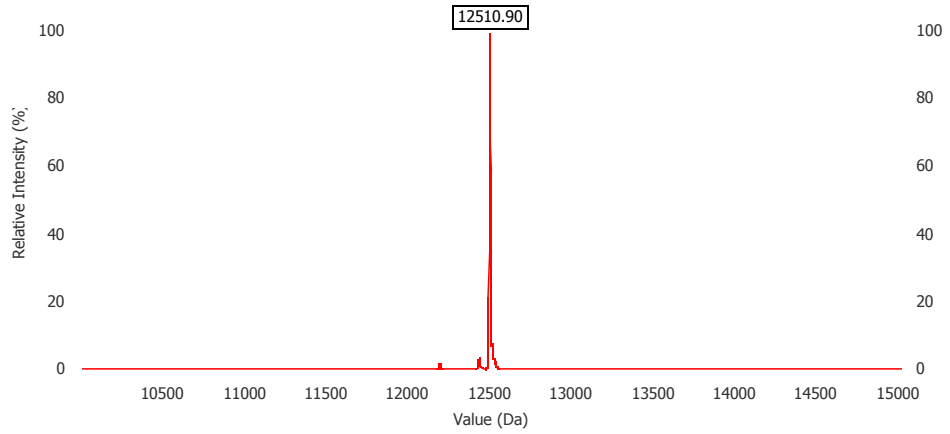
Sequence Name:	C7999
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT//ideoxyU//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12511.2
Measured Molecular Weight:	12511.20

Figure S69. ESI-MS plot of 3555.



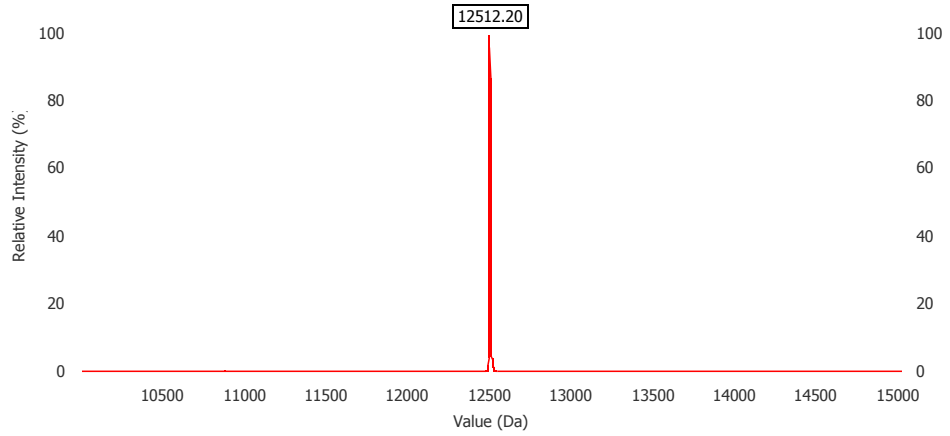
Sequence Name:	C9799
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//iSuper-dT//ideoxyU/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12511.2
Measured Molecular Weight:	12511.30

Figure S70. ESI-MS plot of 5355.



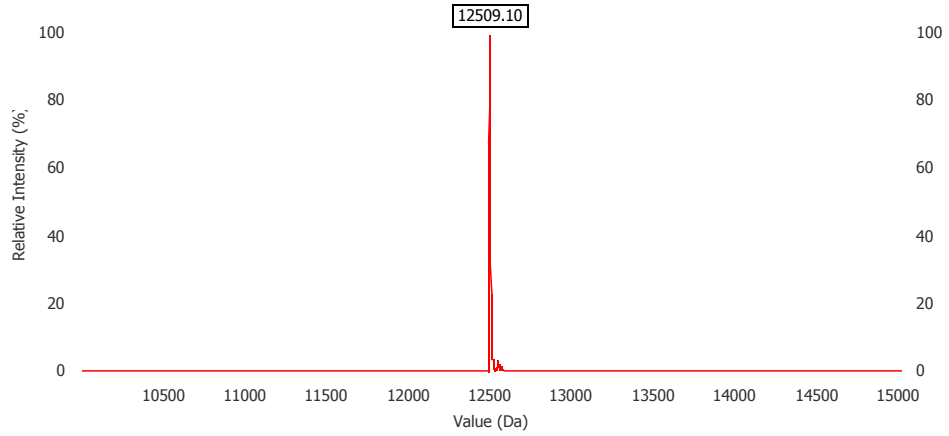
Sequence Name:	C9979
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//iSuper-dT/ /ideoxyU/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12511.2
Measured Molecular Weight:	12510.90

Figure S71. ESI-MS plot of 5535.



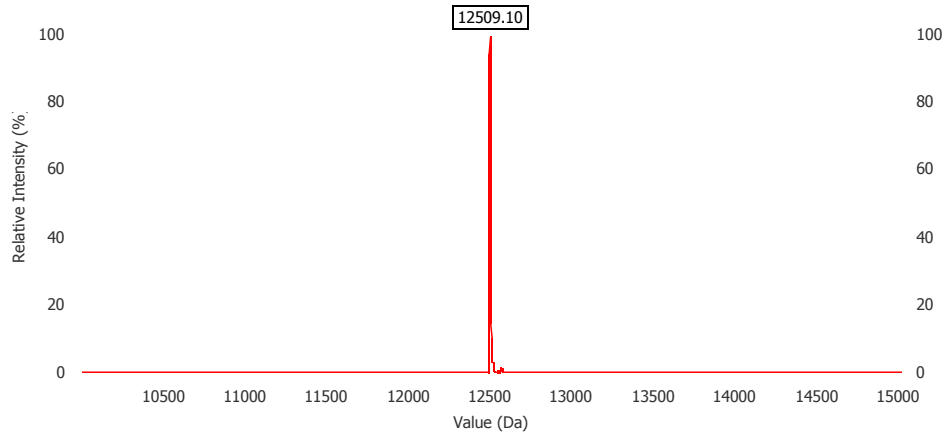
Sequence Name:	C9997
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU//ideoxyU//ideoxyU/ /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12511.2
Measured Molecular Weight:	12512.20

Figure S72. ESI-MS plot of 5553.



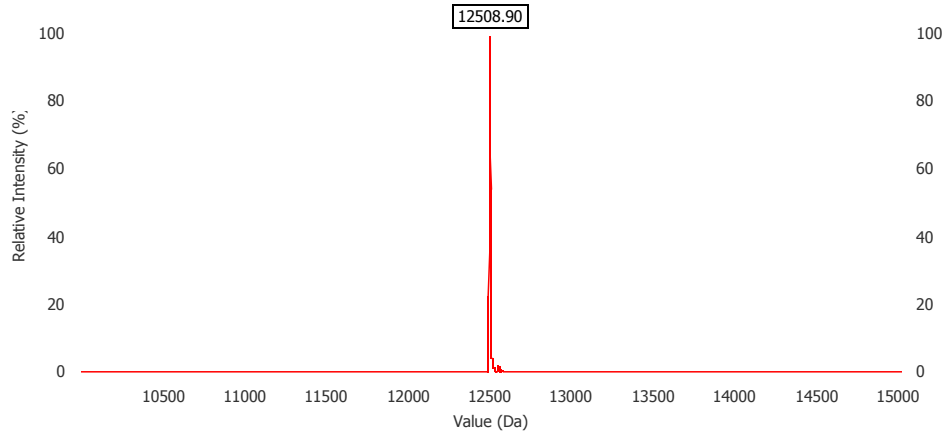
Sequence Name:	ACTC6
Sequence:	5'-/5Biosg/TTT TTT TTT TTT ACT/i5HydMe-dC/TT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12508.3
Measured Molecular Weight:	12509.10

Figure S73. ESI-MS plot of ACT2.



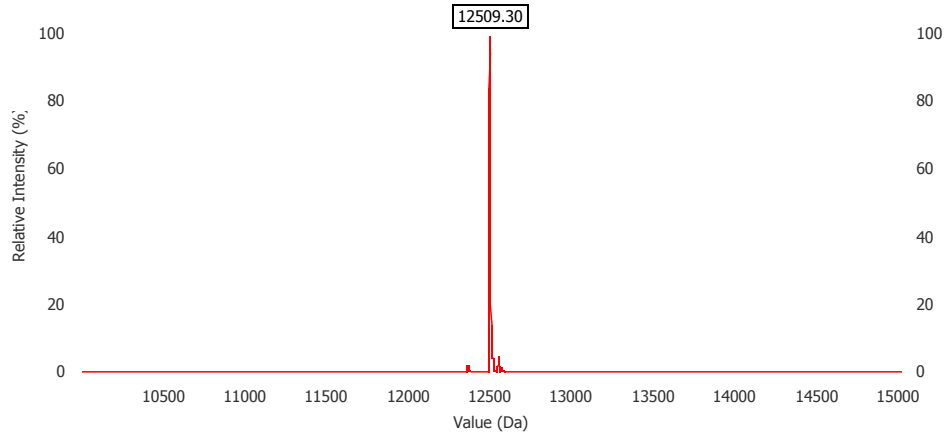
Sequence Name:	ACC6T
Sequence:	5'-/5Biosg/TTT TTT TTT TTT AC/i5HydMe-dC/ TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12508.3
Measured Molecular Weight:	12509.10

Figure S74. ESI-MS plot of AC2T.



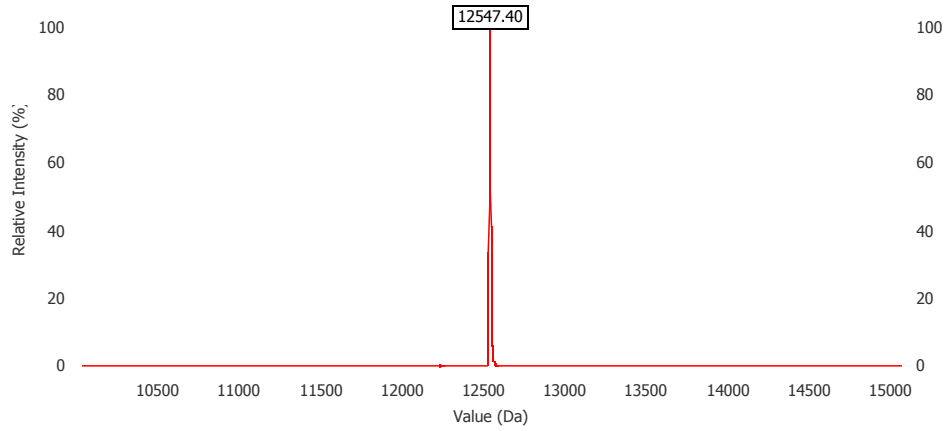
Sequence Name:	AC6CT
Sequence:	5'-/5Biosg/TTT TTT TTT TTT A/i5HydMe-dC/C TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12508.3
Measured Molecular Weight:	12508.90

Figure S75. ESI-MS plot of A2CT.



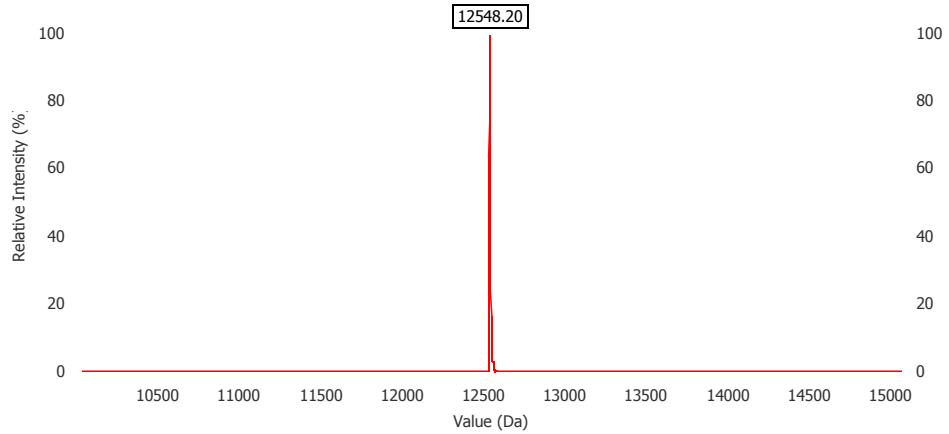
Sequence Name:	C6ACT
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /i5HydMe-dC/AC TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12508.3
Measured Molecular Weight:	12509.30

Figure S76. ESI-MS plot of 2ACT.



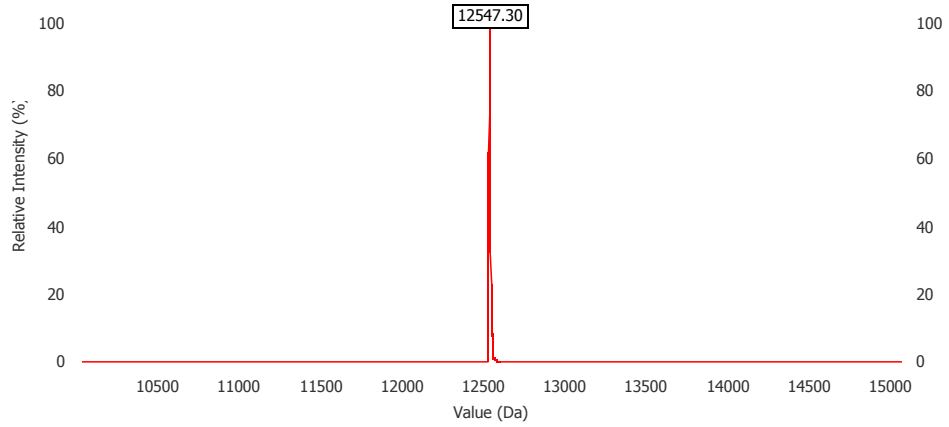
Sequence Name:	ACTC7
Sequence:	5'-/5Biosg/TTT TTT TTT TTT ACT /iSuper-dT/TT TTT TTT TTT TTT TTT TTT TTT -3'
Calculated Molecular Weight:	12547.3
Measured Molecular Weight:	12547.40

Figure S77. ESI-MS plot of ACT3.



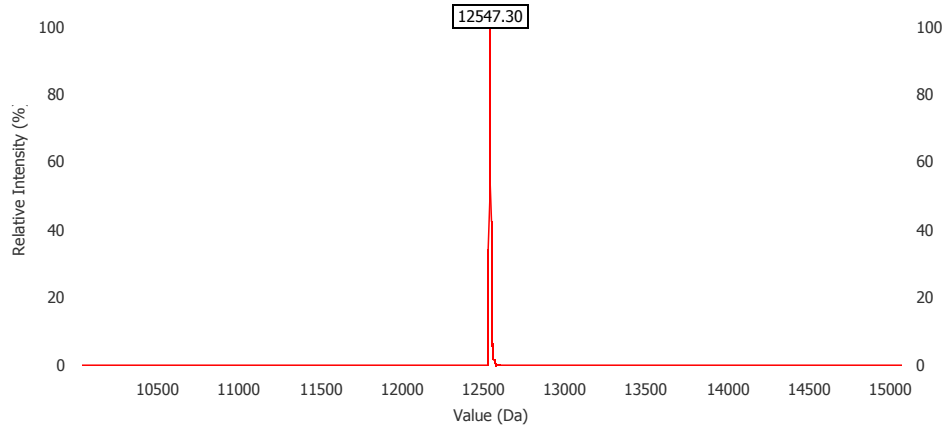
Sequence Name:	ACC7T
Sequence:	5'-/5Biosg/TTT TTT TTT TTT AC/iSuper-dT/ TTT TTT TTT TTT TTT TTT TTT TTT -3'
Calculated Molecular Weight:	12547.3
Measured Molecular Weight:	12548.20

Figure S78. ESI-MS plot of AC3T.



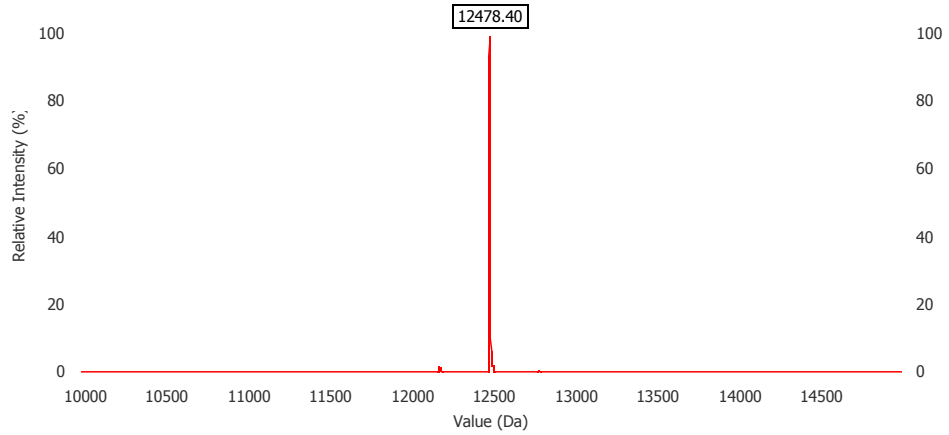
Sequence Name:	AC7CT
Sequence:	5'-/5Biosg/TTT TTT TTT TTT A/iSuper-dT/C TTT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12547.3
Measured Molecular Weight:	12547.30

Figure S79. ESI-MS plot of A3CT.



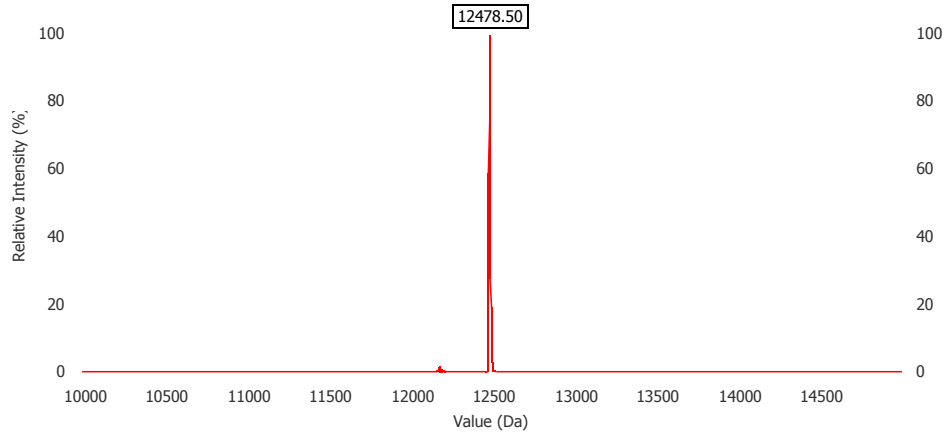
Sequence Name:	C7ACT
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /iSuper-dT/AC TTT TTT TTT TTT TTT TTT TTT TTT T-3'
Calculated Molecular Weight:	12547.3
Measured Molecular Weight:	12547.30

Figure S80. ESI-MS plot of 3ACT.



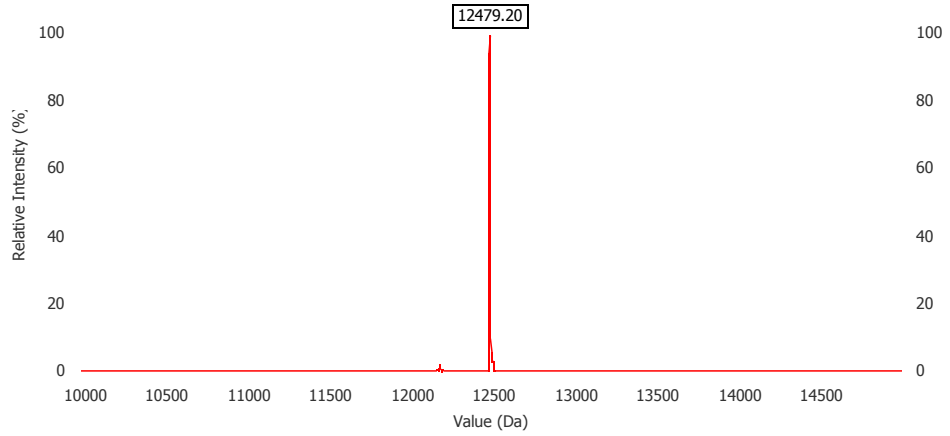
Sequence Name:	ACTC9
Sequence:	5'-/5Biosg/TTT TTT TTT TTT ACT /ideoxyU/TT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12479.2
Measured Molecular Weight:	12478.40

Figure S81. ESI-MS plot of ACT5.



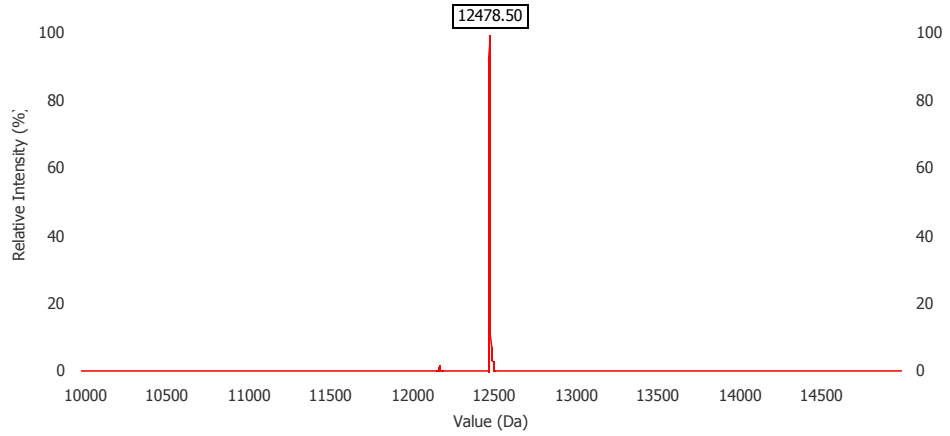
Sequence Name:	ACC9T
Sequence:	5'-/5Biosg/TTT TTT TTT TTT AC/ideoxyU/ TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12479.2
Measured Molecular Weight:	12478.50

Figure S82. ESI-MS plot of AC5T.



Sequence Name:	AC9CT
Sequence:	5'-/5Biosg/TTT TTT TTT TTT A/ideoxyU/C TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12479.2
Measured Molecular Weight:	12479.20

Figure S83. ESI-MS plot of A5CT.



Sequence Name:	C9ACT
Sequence:	5'-/5Biosg/TTT TTT TTT TTT /ideoxyU/AC TTT TTT TTT TTT TTT TTT TTT T -3'
Calculated Molecular Weight:	12479.2
Measured Molecular Weight:	12478.50

Figure S84. ESI-MS plot of 5ACT.

References

1. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2009;NA-NA.

2. Frauer C, Hoffmann T, Bultmann S, Casa V, Cardoso MC, Antes I, et al. Recognition of 5-Hydroxymethylcytosine by the Uhrf1 SRA Domain. Xu S, editor. PLoS ONE. 2011 Jun 22;6(6):e21306.

3. Drew HR, Wing RM, Takano T, Broka C, Tanaka S, Itakura K, et al. Structure of a B-DNA dodecamer: conformation and dynamics. Proceedings of the National Academy of Sciences. 1981 Apr 1;78(4):2179–83.

4. Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J Chem Phys. 2020 Jul 28;153(4):044130.

5. Darden T, York D, Pedersen L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. The Journal of Chemical Physics. 1993 Jun 15;98(12):10089–92.

6. Andersen HC. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. Journal of Computational Physics. 1983 Oct;52(1):24–34.

7. Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. J Comput Chem. 1992 Oct;13(8):952–62.

8. Martyna GJ, Tobias DJ, Klein ML. Constant pressure molecular dynamics algorithms. The Journal of Chemical Physics. 1994 Sep;101(5):4177–89.

9. Hart K, Foloppe N, Baker CM, Denning EJ, Nilsson L, MacKerell AD. Optimization of the CHARMM Additive Force Field for DNA: Improved Treatment of the BI/BII Conformational Equilibrium. J Chem Theory Comput. 2012 Jan 10;8(1):348–62.

10. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics. 1983 Jul 15;79(2):926–35.

11. Yoo J, Aksimentiev A. New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. Phys Chem Chem Phys. 2018;20(13):8432–49.

12. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics. 1996 Feb;14(1):33–8.

13. Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH. Single-molecule DNA detection with an engineered MspA protein nanopore. Proceedings of the National Academy of Sciences of the United States of America. 2008;

14. Scott DW. Multivariate Density Estimation: Theory, Practice, and Visualization [Internet]. 1st ed. Wiley; 2015 [cited 2021 Jun 22]. (Wiley Series in Probability and Statistics). Available from: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118575574

15. Gill A. Introduction to the theory of finite-state machines. New York: McGraw-Hill; 1962.

16. Schouhamer Immink KA. Coding techniques for digital recorders. New York: Prentice Hall; 1991. 297 p.