Supporting Information

Aqueous Wittig Reaction-mediated Fast Fluorogenic Identification and Single-base Resolution Analysis of 5-Formylcytosine in DNA

Qian Zhou,[1] Kun Li,[1*] Kang-Kang Yu,[1] Na Li,[1] Lei Shi,[1] Hao Chen,[2] Shan-Yong Chen[1] and Xiao-Qi Yu[1*]

[1] *Laboratory of Green Chemistry and Technology (Ministry of Education), College of Chemistry, Sichuan University, Chengdu 610064, P. R. China*

E-mail: kli@scu.edu.cn; xqyu@scu.edu.cn; Fax: +86-28-85415886; Tel: +86-28-85415886

[2]*Department of Pharmaceutical Sciences, College of Pharmacy, University of Tennessee Health Science Centre, Memphis, Tennessee 38163, United States*

Table of Contents

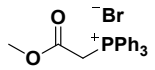
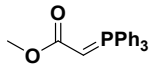
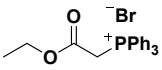
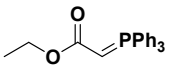
1. Experimental Procedures

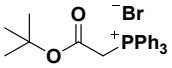
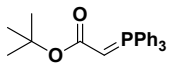
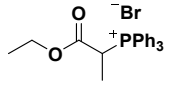
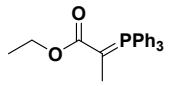
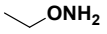
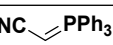
Materials and Instruments

Unless otherwise noted, all chemical reagents were purchased from commercial suppliers and used without further purification. All solvents were dried according to the standard methods prior to use. In the optical spectroscopic studies, all the solvents were either HPLC or spectroscopic grade. Except for 5-methylcytidine was purchased from Aladdin, all modified nucleosides were synthesized [1,2,3]. Unmodified DNA oligonucleotides were purchased from Tsingke Biological Technology and modified DNA oligonucleotides were from Takara.

Thin layer chromatography (TLC) was performed on silica gel plates, and spots were visualized under UV light. Column chromatography was carried out using 200-300 mesh silica gel (Qingdao Ocean Chemicals). NMR spectra were recorded on a Bruker AMX-400 spectrometer at 25 °C (^{1}H NMR: 400 MHz, ^{13}C NMR: 101 MHz) and chemical shifts are expressed in parts per million (ppm) using the internal standard tertramethylsilane or the deuterated solvent ($CDCl_3$, DMSO-d_6, Methanol-d_4, D_2O) as reference. Spin multiplicities in ^{1}H NMR are reported as singlet (s), doublet (d), double doublet (dd), double double doublet (ddd), triplet (t), triplet of triplet (tt), multiplet/overlapping peaks (m) or broad (br). The High-resolution mass spectra (HRMS) were obtained on a Finnigan LCQDECA and a Bruker Daltonics Bio TOF mass spectrometer. pH values were determined by a pH-3c digital pH-meter (Shanghai Lei Ci Device Works, Shanghai, China) with a combined glass-calomel electrode. UV absorption spectra were recorded on a Persee TU-1901 UV-visible spectrophotometer. Fluorescence spectra were measured on a Hitachi F-7000 fluorescence spectrophotometer. HPLC analysis were performed on Waters Associates equipment (Waters 2695 with 2998 Photodiode Array Detector) which equipped with an Ultimate® XB-C18 column (5 μm, 300 Å, 4.6×250 mm, Welch, China).

Compounds Information

	Abbreviation	Structure	Chemical Name	Supplier
1	**MA+**	 CAS: 1779-58-4	(Methoxycarbonylmethyl)triphenylphosphonium bromide	Aladdin
2	**MA**	 CAS: 2605-67-6	Methyl(triphenylphosphoranylidene)acetate	Aladdin
3	**EA+**	 CAS: 1530-45-6	Ethoxycarbonylmethyl(triphenyl)phosphonium bromide	Aladdin
4	**EA**	 CAS: 1099-45-2	Ethyl (triphenylphosphoranylidene)acetate	Innochem

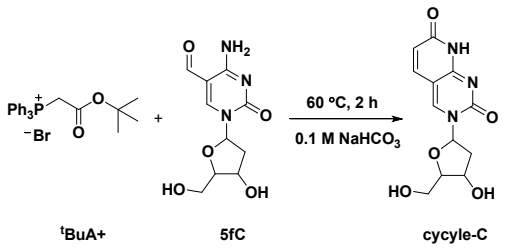
5	**^{t}BuA+**	 CAS: 59159-39-6	(*tert*-Butoxycarbonylmethyl)triphenylphosphonium bromide	Aladdin
6	**^{t}BuA**	 CAS: 35000-38-5	*tert*-Butyl 2-(triphenylphosphoranylidene)acetate	Adamas
7	**MEA+**	 CAS: 30018-16-7	[1-(Ethoxycarbonyl)ethyl]triphenylphosphonium bromide	Adamas
8	**MEA**	 CAS: 5717-37-3	(1-Ethoxycarbonylethylidene)triphenylphosphorane	Adamas
9	**EtONH$_2$**	 CAS: 3332-29-4	Ethoxyamine hydrochloride	Adamas
10	**PPh$_3$=CN**	 CAS: 16640-68-9	(Triphenylphosphoranylidene)acetonitrile	Aladdin

HPLC Analysis

The HPLC analysis of the derivatization reaction mixtures were performed on Waters Associates equipment (Waters 2695 with 2998 Photodiode Array Detector) which equipped with an Ultimate® XB-C18 column (5 μm, 300 Å, 4.6×250 mm, Welch, China) with mobile phase A (H_2O) and B (CH_3OH) with flow rate of 1 mL/min at 35°C (B Conc.: 20% / 0-10 min).

Chemical Synthesis

3-((2R,4S,5R)-4-hydroxy-5-(hydroxymethyl)tetrahydrofuran-2-yl)pyrido[2,3-d]pyrimidine-2,7(3H,8H)-dione (cycle-C)



Scheme S1. The synthetic route of **cycle-C**.

To a solution of 5fC (43 mg, 0.1685 mmol) in 0.1 M $NaHCO_3$, phosphonium salt **^{t}BuA+** (193 mg, 0.4220 mmol) was added. After stirring at 60 °C for 2 h, the solvent was evaporated in vacuo and the crude product was

purified by silica gel column chromatography (CH_2Cl_2: Methanol = 10:1, v/v) to afford **cycle-C** as a white solid. ^{1}H NMR (400 MHz, D_2O) δ 8.95 (s, 1H), 7.77 (d, J = 9.6 Hz, 1H), 6.37 (d, J = 9.6 Hz, 1H), 6.20 (t, J = 6.0 Hz, 1H), 4.39 - 4.36 (m, 1H), 4.16 - 4.13 (m, 1H), 3.73-3.91 (m, 2H), 2.66 - 2.29(m, 2H); ^{13}C NMR (101 MHz, D_2O) δ 146.5, 139.2, 118.0, 103.6, 88.3, 87.5, 69.6, 60.5, 40.1; HRMS (ESI) m/z calcd for $[M+Na]^+$: 302.0753, found: 302.0752.

2. Supplementary Figures

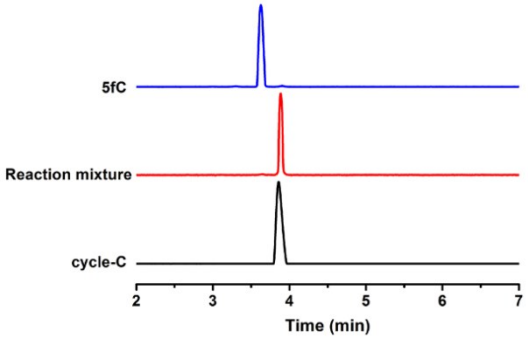


Fig. S1 HPLC chromatograms (UV = 260 nm) of 5fC, **cycle-C**, and the reaction mixture of **^{t}BuA+** and 5fC after incubation in 0.1 M $NaHCO_3$ at 37 °C for 1.5 h.

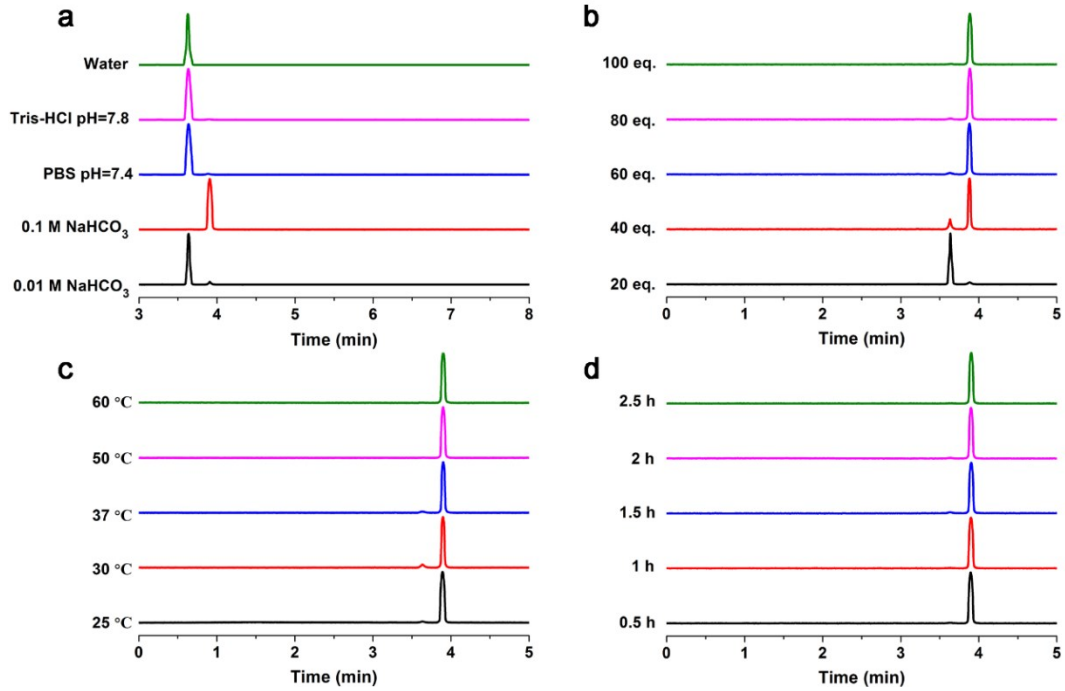


Fig. S2 Optimization of derivatization conditions for 5fC by **^{t}BuA+**, including the pH, molar ratio of **^{t}BuA+**/5fC**,** temperature and reaction time.

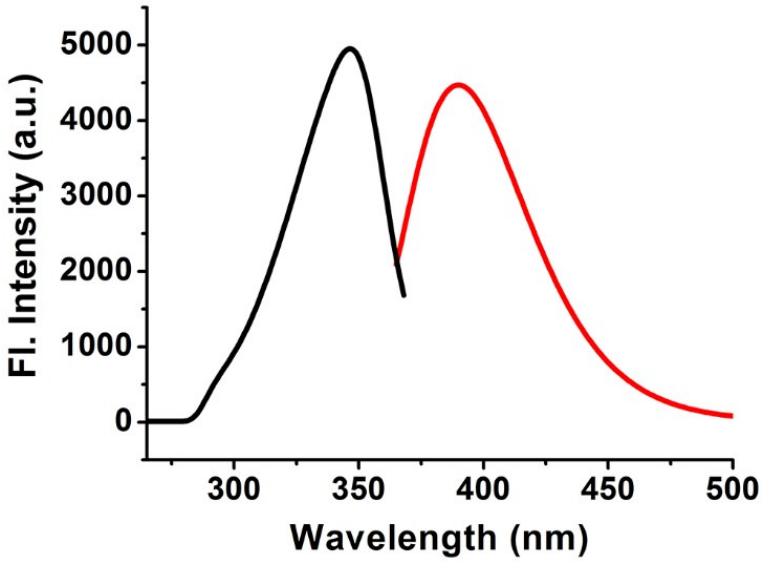


Fig. S3 Fluorescence excitation and emission spectrum of **cycle-C** in PBS.

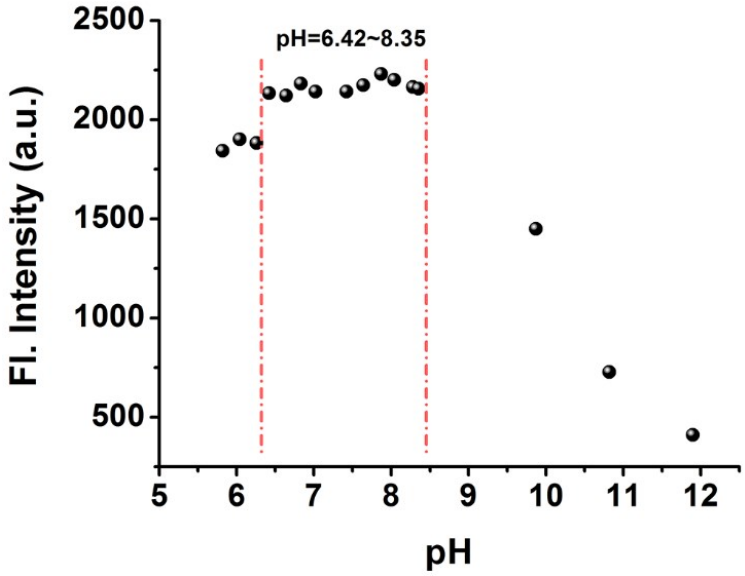


Fig. S4 Fluorescence intensity of **cycle-C** in different pH buffers.

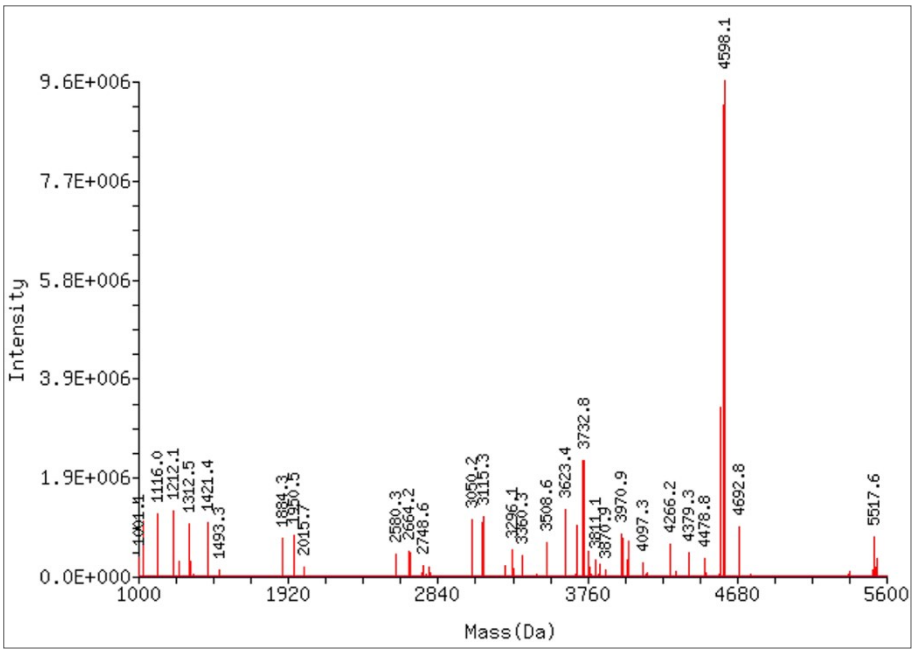


Fig. S5 ESI-MS Spectrum of ODN-5fC after labelling with **^{t}BuA+**. 5'-GACTCAA**cycle-C**AGCCGTA-3', caculated 4598.0, found 4598.1.

Antibody-based immunoprecipitation and chemical-assisted biotin pulldown with aldehyde reactive probes (ARPs), such as *o*-(biotinylcarbazoylmethyl) hydroxylamine, have been used to isolate and enrich 5fC-containing DNA fragments in the genome followed by sequencing, but such affinity-based approaches have limited resolution (100-400 bases).[4] In addition, the poor specificity of antibodies and ARPs for targeted modifications always leads to false positive signals.[5] Currently, the gold standard established for detecting C modifications at single-base resolution is bisulfite sequencing (BS-seq), where unmodified C, 5fC and 5caC, are efficiently converted to uracil (U) by sodium bisulfite-mediated deamination and read as thymine (T) in subsequent sequencing, while 5mC and 5hmC are resistant to this chemical conversion and therefore still be read as C. That is, this method does not distinguish 5fC from C and 5caC or 5mC from 5hmC. Additional chemical derivatization steps before sodium bisulfite treatment can be used to overcome these limitations. The most recent breakthrough in this field came with several new methods (Table S1), allowing selectively positional measurement of the four C oxidations, respectively.

Sequencing method	Readout of residue in sequencing method				
	C	5mC	5hmC	5fC	5caC
Bisulfite sequencing (BS-seq)	T	C	C	T	T
Oxidative BS-seq (oxBS-seq)	T	C	T	T	T
Reductive BS-seq (redBS-seq)	T	C	C	C	T
TET-assisted BS-seq (TAB-seq)	T	T	C	T	T
5fC CAB-seq (fCAB-seq)	T	C	C	C	T
5caC CAB-seq (caCAB-seq)	T	C	C	T	C
Methylase-assisted BS-seq (MAB-seq)	C	C	C	T	T
5caC MAB-seq (caMAB-seq)	C	C	C	C	T

Table S1. Bisulfite sequencing and its modifications.[6]

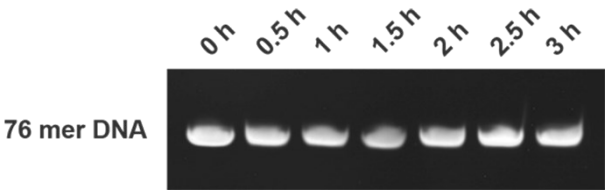


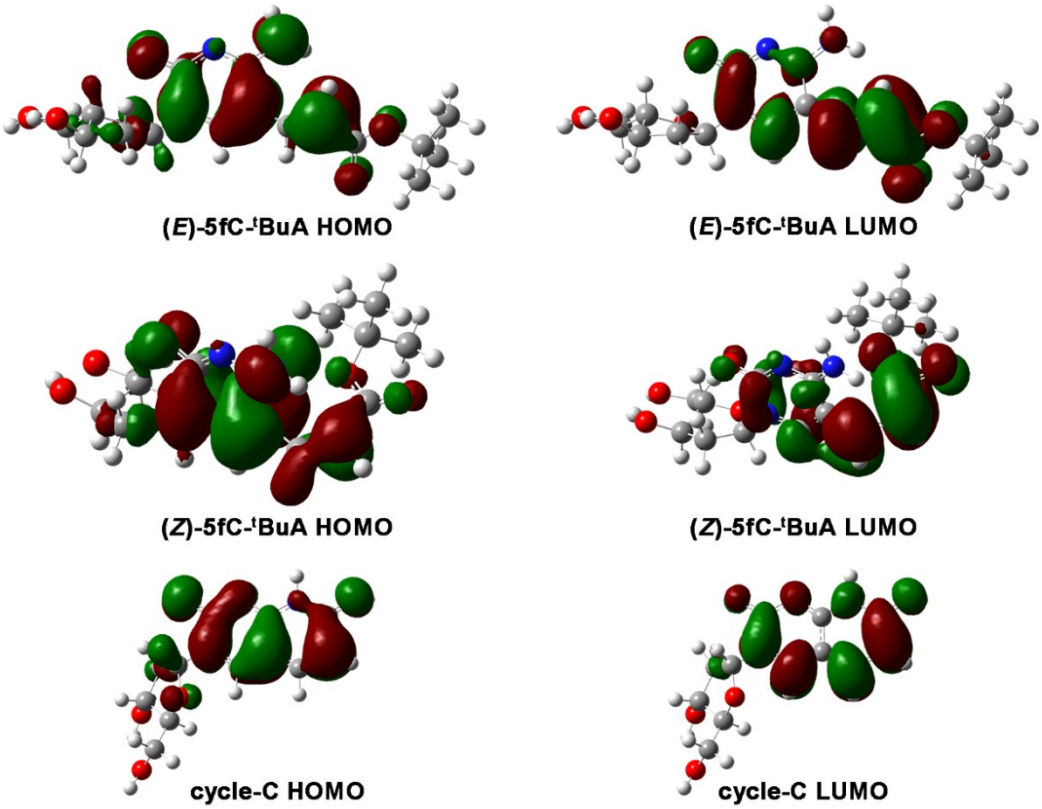
Fig. S6 Denaturing polyacrylamide gel electrophoresis (PAGE) shows no noticeable degradation after reaction with **^{t}BuA+** at different time point.

3. Table of Oligonucleotides Sequences

Oligomer	**Sequence (from 5' to 3')**
ODN-C	GACTCAA**C**AGCCGTA
ODN-AP	GACTCAA**AP**AGCCGTA
ODN-5fU	GACTCAA**5fU**AGCCGTA
ODN-5fC	GACTCAA**5fC**AGCCGTA
76-mer dsDNA-5fC	a) CCTCACCATCTCAACCAATATTATATTATGTGTATATT**5fC**GATATTTTG TGTTATAATATTGAGGGAGAAGTGGTGA b) TCACCACTTCTCCCTCAATATTATAACACAAAATATCGAATATACACAT AATATAATATTGGTTGAGATGGTGAGG
Forward primer	CCCTTTTATTATTTTAATTAATATTATATT
Reverse primer	CTCCGACATTATCACTACCATCAACCACCCATCCTACCTGGACTACATTCT TATTCAGTATTCACCACTTCTCCCTCAAT
Sequencing primer	CTCCGACATTATCACTACCA

4. DFT Calculations

The Gaussian 09 program6 was used. [7]



The HOMO and LUMO plots of all relevant structures.

Below we report the Cartesian coordinates of all relevant structures.

Structure of (*E*)-5fC-^{t}BuA

C	0.54056100	0.47390800	-0.18464000
C	-0.57068300	-0.33644700	-0.11506200
N	-1.82621600	0.13397800	0.06287000
C	-2.05155500	1.54154700	0.17713400
N	-0.98075700	2.36413400	0.02702000
C	0.24783800	1.89540800	-0.16449100
C	1.84189600	-0.16695700	-0.27924800
N	1.22278400	2.81333400	-0.34604100
C	3.07501100	0.33285600	-0.04097500
C	4.26479500	-0.53455400	-0.18003500
O	5.37477900	0.14664900	0.16531200
O	4.24411600	-1.70215900	-0.55422500
C	6.73441700	-0.45076400	0.13323400
C	7.61574000	0.70767700	0.60587300
C	6.80963200	-1.62537000	1.11321600
C	7.09102200	-0.85423500	-1.30056100
O	-3.19776300	1.95307700	0.38491000
C	-2.94925800	-0.83336400	0.11481500
O	-3.84811500	-0.61264100	-0.96742700
C	-5.16708100	-0.34103100	-0.46173800

C	-5.20022100	-1.03039900	0.91232000
C	-3.77151300	-0.81545600	1.41604500
O	-6.15387200	-0.47092400	1.80817800
C	-6.19306900	-0.87571400	-1.44860200
O	-7.47283500	-0.64543000	-0.83798900
H	-0.47148900	-1.41461000	-0.18147300
H	1.80538400	-1.22205400	-0.54833600
H	0.93105300	3.77482000	-0.46113000
H	2.10684100	2.55202200	-0.75496100
H	3.26262100	1.34104800	0.30681400
H	8.66278400	0.39135600	0.62557600
H	7.52557200	1.56389800	-0.06946500
H	7.33116000	1.02530400	1.61355900
H	7.84762800	-1.96598000	1.18388300
H	6.19096200	-2.46182600	0.78537100
H	6.48655300	-1.31204500	2.11112600
H	6.96616300	-0.00447900	-1.97930000
H	8.14048800	-1.16403300	-1.33407100
H	6.47303400	-1.68236500	-1.65000600
H	-2.47841400	-1.80669400	-0.04097000
H	-5.29550700	0.73916900	-0.32374900
H	-5.40189400	-2.10288400	0.77420700
H	-3.43410700	-1.59208900	2.10543800
H	-3.70855500	0.15677200	1.90620200
H	-7.01724800	-0.53861200	1.36930000
H	-6.11833600	-0.35219700	-2.40943800
H	-6.02716000	-1.94792500	-1.61473800
H	-8.16386400	-1.02793000	-1.39457000

Structure of (*Z*)-5fC-^{t}BuA

C	1.19341100	-1.60603900	-0.44320500
C	-0.06191300	-1.40303900	-0.94476000
N	-1.11113300	-0.99457700	-0.17782200
C	-0.91102000	-0.71616300	1.20574100
N	0.29645900	-1.03117200	1.74470100
C	1.30368600	-1.46976000	0.99438300
C	2.27749100	-2.04723000	-1.33970100
N	2.43634600	-1.82178800	1.63260000
C	3.53348400	-1.57352200	-1.47348600
O	-1.83869300	-0.22975200	1.86572700
C	-2.41563500	-0.76248900	-0.83219100
O	-2.73195700	0.62992300	-0.85856900
C	-3.97928200	0.86082500	-0.18286800
C	-4.73851100	-0.46967700	-0.31504900

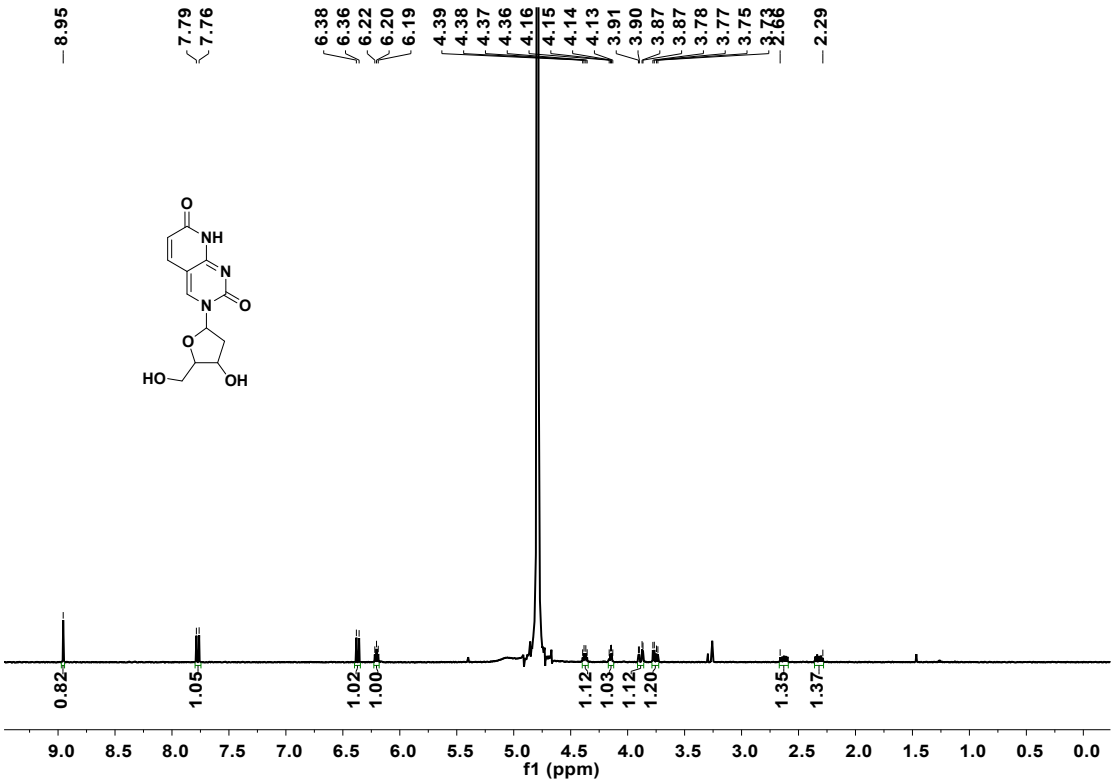
C	-3.61091100	-1.49690800	-0.19810000
O	-5.72518900	-0.66598900	0.69180200
C	-4.68168700	2.04253900	-0.83319400
O	-5.94921500	2.15602300	-0.16606500
C	4.17645300	-0.38158800	-0.85888800
O	5.39548900	-0.24551900	-0.86748100
O	3.30346600	0.50639300	-0.37072600
C	3.72239800	1.79998600	0.24137300
C	2.38355900	2.42906900	0.63164600
C	4.58308500	1.53743500	1.48004600
C	4.43919100	2.65513500	-0.80724200
H	-0.27269700	-1.54635300	-1.99904000
H	2.02268100	-2.88826700	-1.98400900
H	2.45181500	-1.80257600	2.64190200
H	3.20057000	-2.26489600	1.14681900
H	4.20426400	-2.09566200	-2.14920900
H	-2.26499300	-1.08151600	-1.86611100
H	-3.79424200	1.06287200	0.87917600
H	-5.20514400	-0.52210300	-1.30989300
H	-3.82120700	-2.43132600	-0.72256200
H	-3.42854900	-1.70631500	0.85659000
H	-6.33399100	0.08790100	0.63085100
H	-4.09296400	2.95998400	-0.71196200
H	-4.82089700	1.84928700	-1.90464600
H	-6.47418100	2.84363800	-0.59643000
H	2.55792900	3.40265500	1.09907000
H	1.75270900	2.57698500	-0.24999800
H	1.84796000	1.79392400	1.34283500
H	4.76671700	2.48746600	1.99209600
H	5.54441200	1.09484000	1.21561400
H	4.06148500	0.87154200	2.17397600
H	3.81872500	2.76245200	-1.70263500
H	4.61308000	3.65365500	-0.39352000
H	5.40079800	2.22451100	-1.08979100

Structure of cycle-C

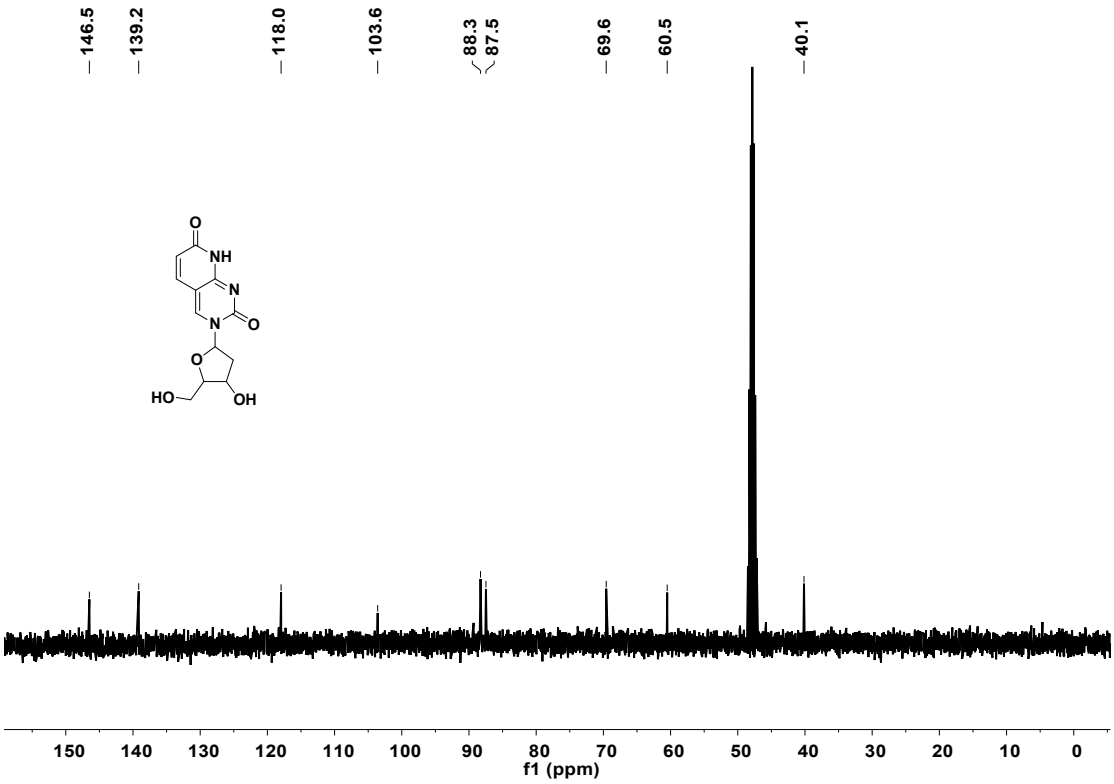
C	-1.91178400	-0.67693300	-0.09201800
C	-0.55334000	-0.51497800	-0.28567400
N	-0.01837700	0.71059800	-0.39887700
C	-0.82088000	1.89484800	-0.33234100
N	-2.16315400	1.74594400	-0.15725900
C	-2.67499500	0.53164900	-0.03910400
C	-2.58162000	-1.94013600	0.03977900
N	-4.02707100	0.41683400	0.14137700

C	-3.92648300	-1.99690700	0.21905900
O	-0.25515000	2.98811300	-0.43677500
C	1.44862800	0.90744300	-0.62145000
O	2.02334500	-0.31545000	-1.05110200
C	3.04327800	-0.74188000	-0.12350100
C	3.50433400	0.55510300	0.55815500
C	2.19779500	1.34472000	0.65050500
O	4.07293500	0.35080400	1.84381000
C	4.13337200	-1.47017500	-0.89384800
O	5.14733300	-1.77339800	0.07499200
C	-4.73251600	-0.78338800	0.27969200
O	-5.95782100	-0.75545300	0.43926400
H	0.13031700	-1.35106600	-0.36830600
H	-1.99140400	-2.85073600	-0.00619500
H	-4.57068800	1.27396200	0.17874100
H	-4.45538600	-2.93642200	0.32216600
H	1.52165500	1.64073000	-1.42570600
H	2.61061800	-1.41149000	0.63360400
H	4.22091100	1.07289100	-0.09544300
H	2.33802500	2.42493500	0.68716800
H	1.65191000	1.02978200	1.54572200
H	4.83425900	-0.23898200	1.72105900
H	3.74088700	-2.38558600	-1.35250700
H	4.52617400	-0.81668800	-1.68304600
H	5.90305700	-2.17807700	-0.37098500

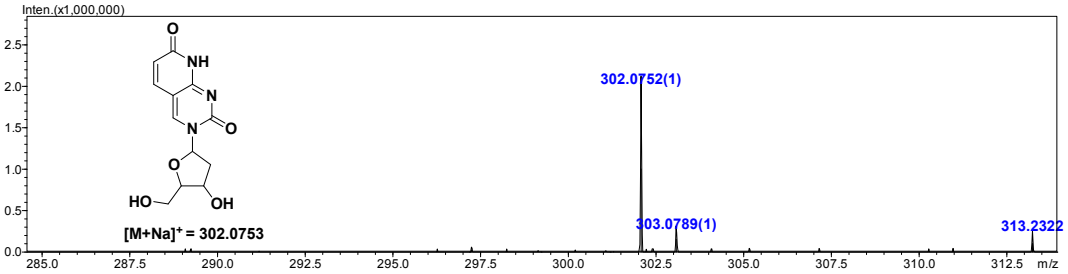
5. ^{1}H NMR, ^{13}C NMR and ESI-MS Spectra



The ^{1}H NMR spectrum of **cycle-C** in D_2O.



The ^{13}C NMR spectrum of **cycle-C** in D_2O.



The ESI-MS spectrum of **cycle-C**.

6. References

[1] Wang, C. J.; Xie, F.; Zhang, W. B.; *Chinese J. Org. Chem.*, **2008**, *28*, 503-505.

[2] Safadi, Y. E.; Paillart, J. C.; Laumond, G.; Aubertin, A. M.; Burger, A.; Marquet, R.; Boudou, V. V.; *J. Med. Chem.* **2010**, *53*, 1534-1545.

[3] Guo, P.; Yan, S.; Hu, J.; Xing, X.; Wang, C.; Xu, X.; Qiu, X.; Ma, W.; Lu, C.; Weng, X.; Zhou, X.; *Org. Lett.* **2013**, *15*, 3266-3269.

[4] Shen, L.; Wu, H.; Diep, D.; Yamaguchi, S.; D'Alessio, A. C.; Fung, H. L.; Zhang, K.; Zhang, Y.; *Cell* **2013**, *153*, 692-706.

[5] Lentini, A.; Lagerwall, C.; Vikingsson, S.; Mjoseng, H. K.; Douvlataniotis, K.; Vogt, H.; Green, H.; Meehan, R. R.; Benson, M.; Nestor, C. E.; *Nat. Methods* **2018**, *15*, 499-504.

[6] Berney, M.; McGouran, J. F.; *Nat. Rev. Chem.* **2018**, *2*, 332-348.

[7] Gaussian 09, Revision D.01, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; Gaussian, Inc., Wallingford CT, **2013**.