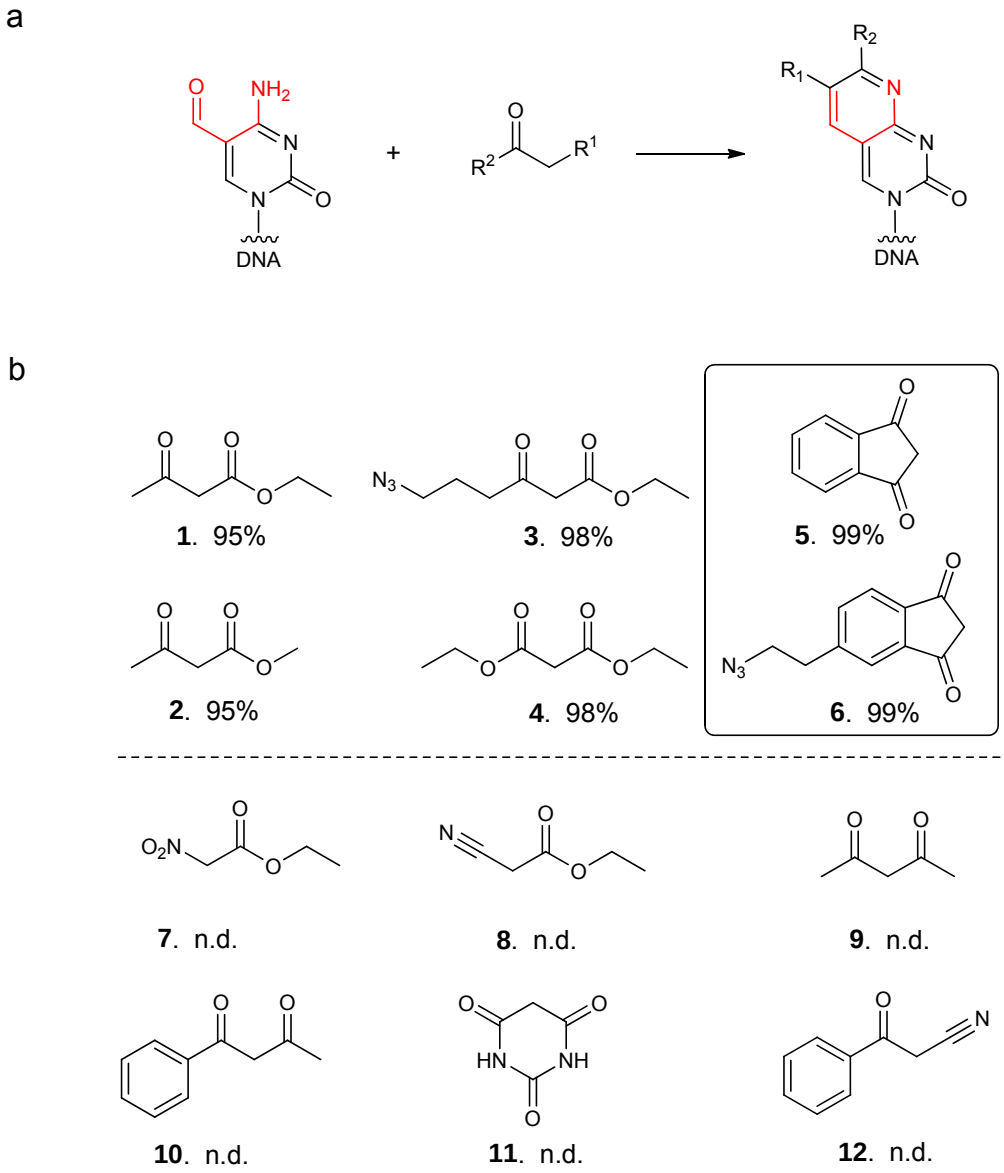
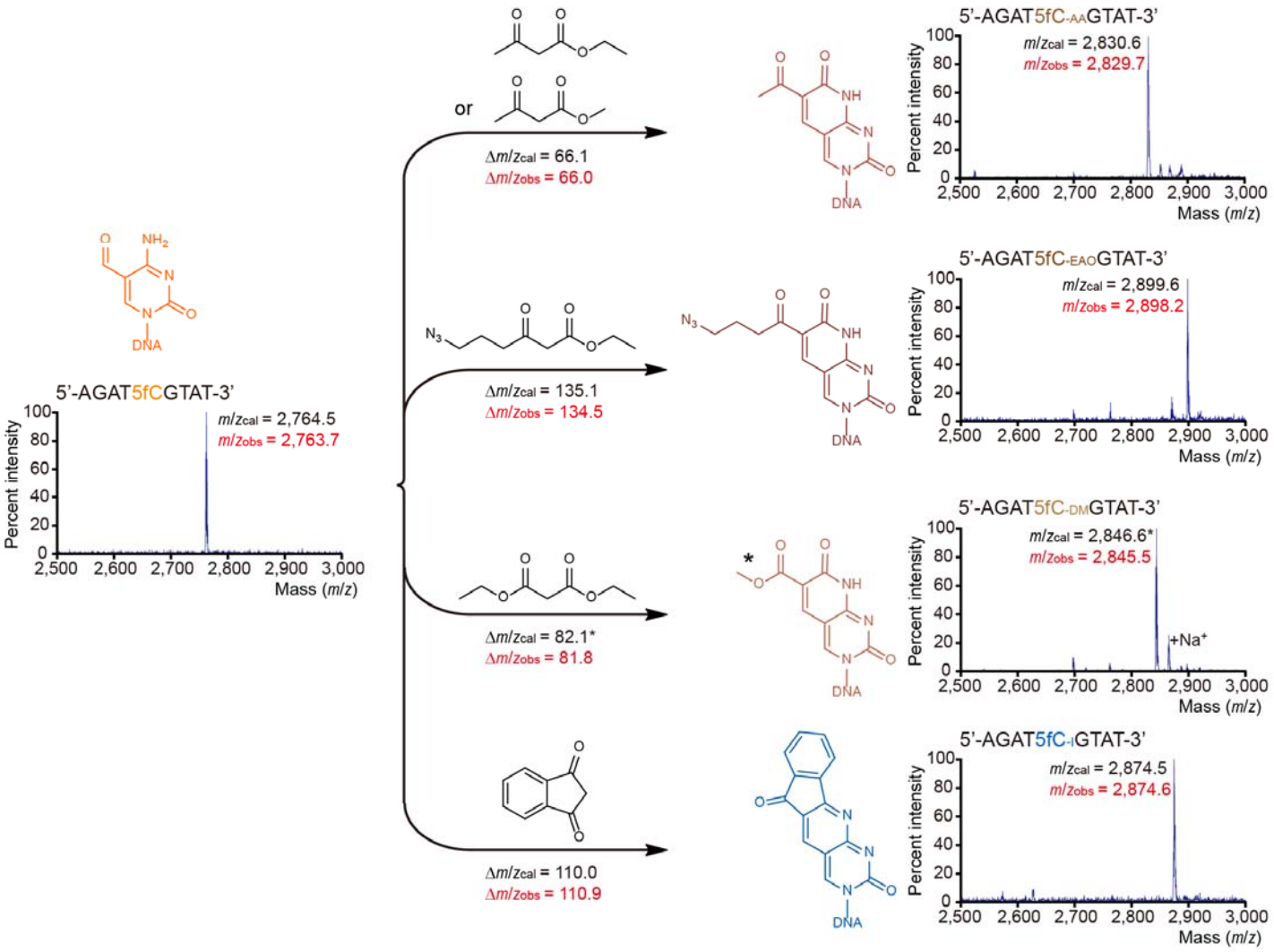


Supplementary Figure 1

Chemical labeling of 5fC via Friedländer reaction

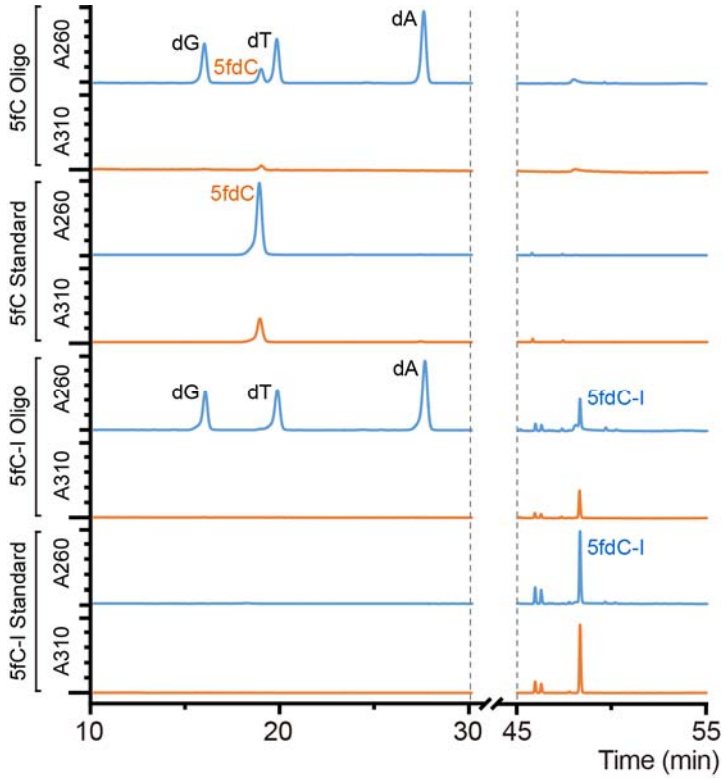
(a) Proposed chemical labeling of 5fC based on the principle of Friedländer reaction[30]. **(b)** Chemicals screened for 5fC labeling, with reaction yield indicated. **1**, ethyl acetoacetate (EA). **2**, methyl acetoacetate (MA). **3**, ethyl 6-azido-3-oxohexanoate (EAO). **4**, diethyl malonate (DM). **5**, 1,3-Indandione (I). **6**, 5-(2-azidoethyl)-1, 3-indandione (AI). **7**, ethyl nitroacetate. **8**, ethyl cyanacetate. **9**, acetylacetone. **10**, benzoylacetone. **11**, barbituric acid. **12**. benzoylacetonitrile. n.d., no detectable products as measured by MALDI-TOF mass spectrometry.



Supplementary Figure 2

MALDI-TOF characterizations of chemical labeling of 5fC in a 9-mer model DNA.

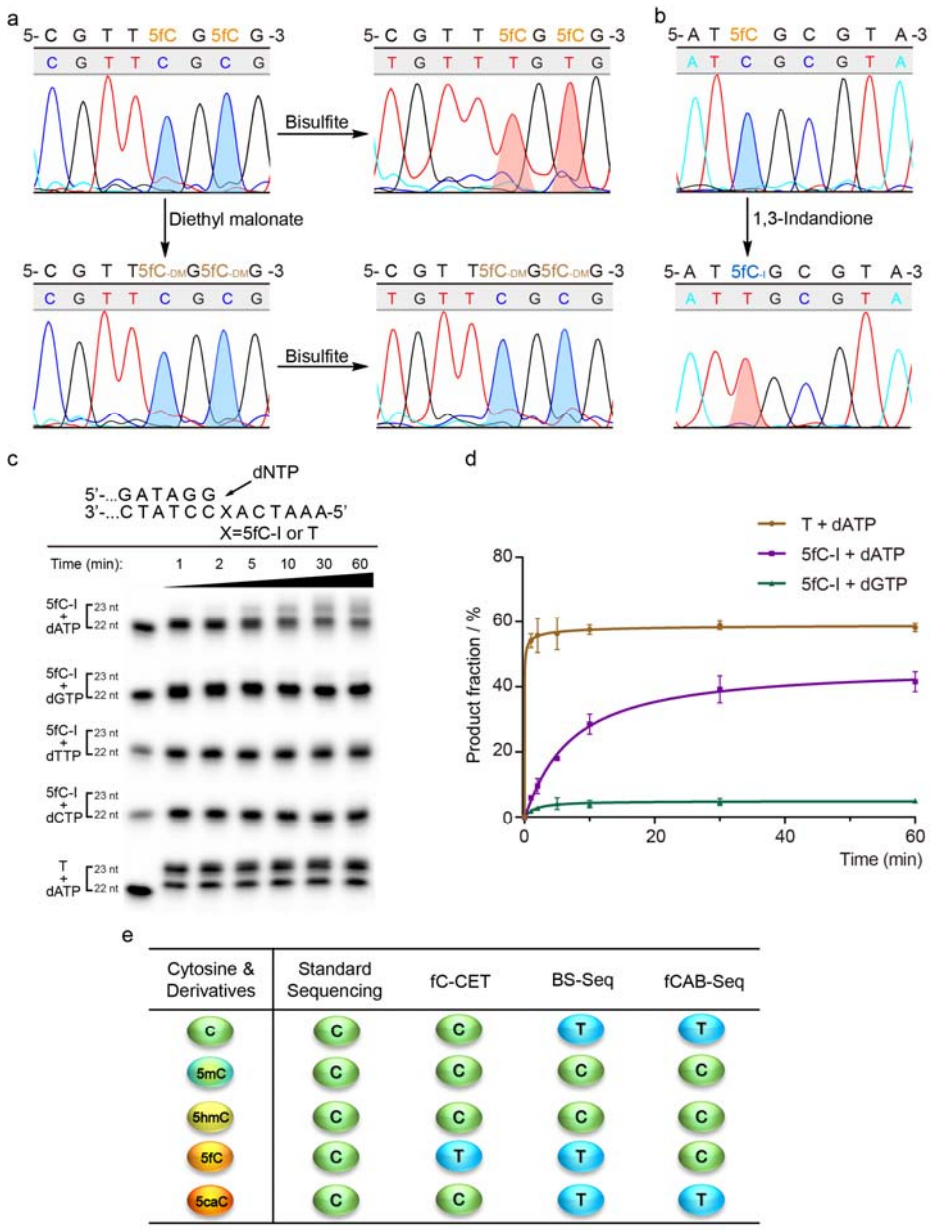
Calculated and observed molecular weight are shown. The proposed chemical structures of labeling product of 5fC are also drawn. *: Because this reaction was performed in alkaline methanol solution, the ethyl ester undergoes transesterification to form the methyl ester. 5fC-AA: reaction product between 5fC and ethyl or methyl acetoacetate. 5fC-EAO: reaction product between 5fC and ethyl 6-azido-3-oxohexanoate. 5fC-DM: reaction product between 5fC and diethyl malonate. 5fC-I: reaction product between 5fC and 1,3-indandione.



Supplementary Figure 3

HPLC analysis of 1,3-indanedione labeling on 5fC-containing 9-mer model DNA.

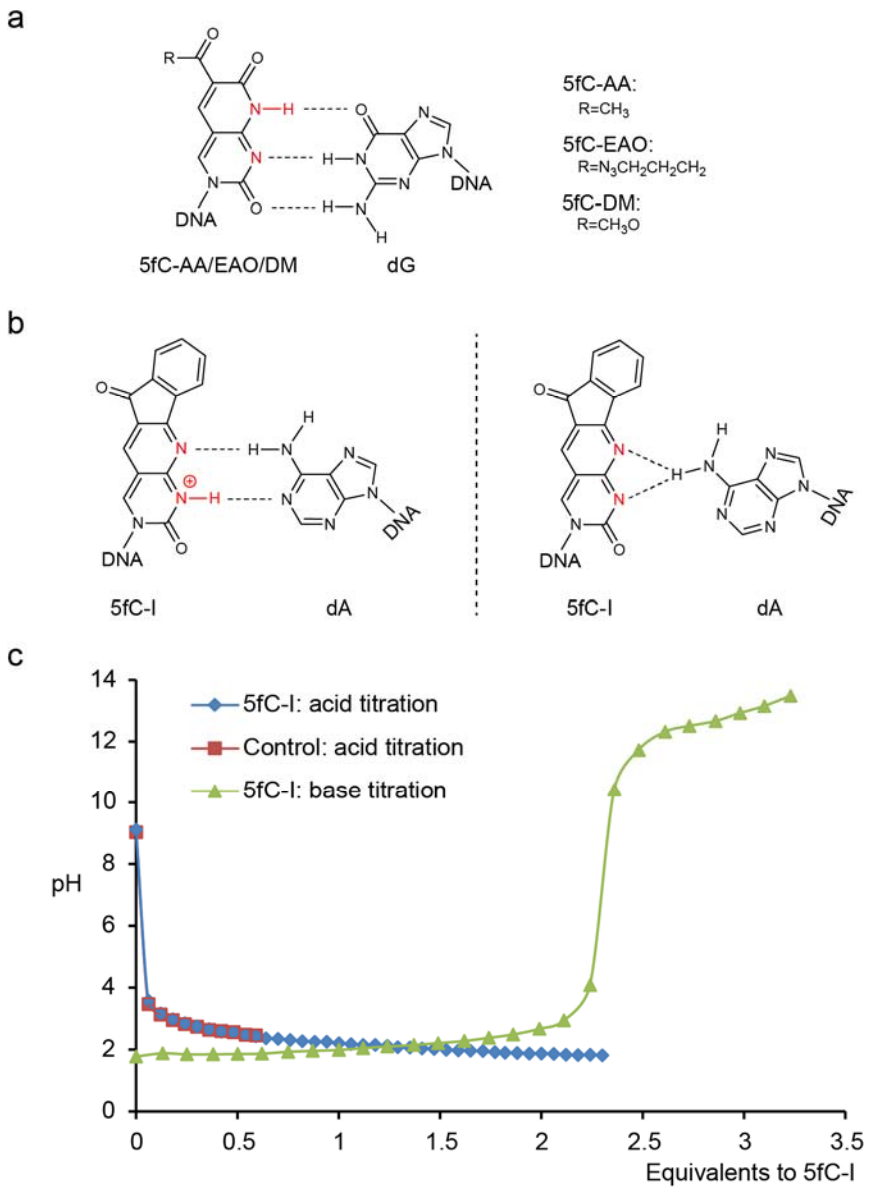
9-mer 5fC and 5fC-I oligonucleotides were digested to nucleosides and analyzed with a C-18 column of HPLC (260 nm and 310 nm). The peak for 5fC is completely undetectable after reaction, indicating full labeling. Authentic 5fC or 5fC-I nucleosides were also analyzed as positive controls.



Supplementary Figure 4

Properties of 5fC labeling products during PCR and/or bisulfite treatment.

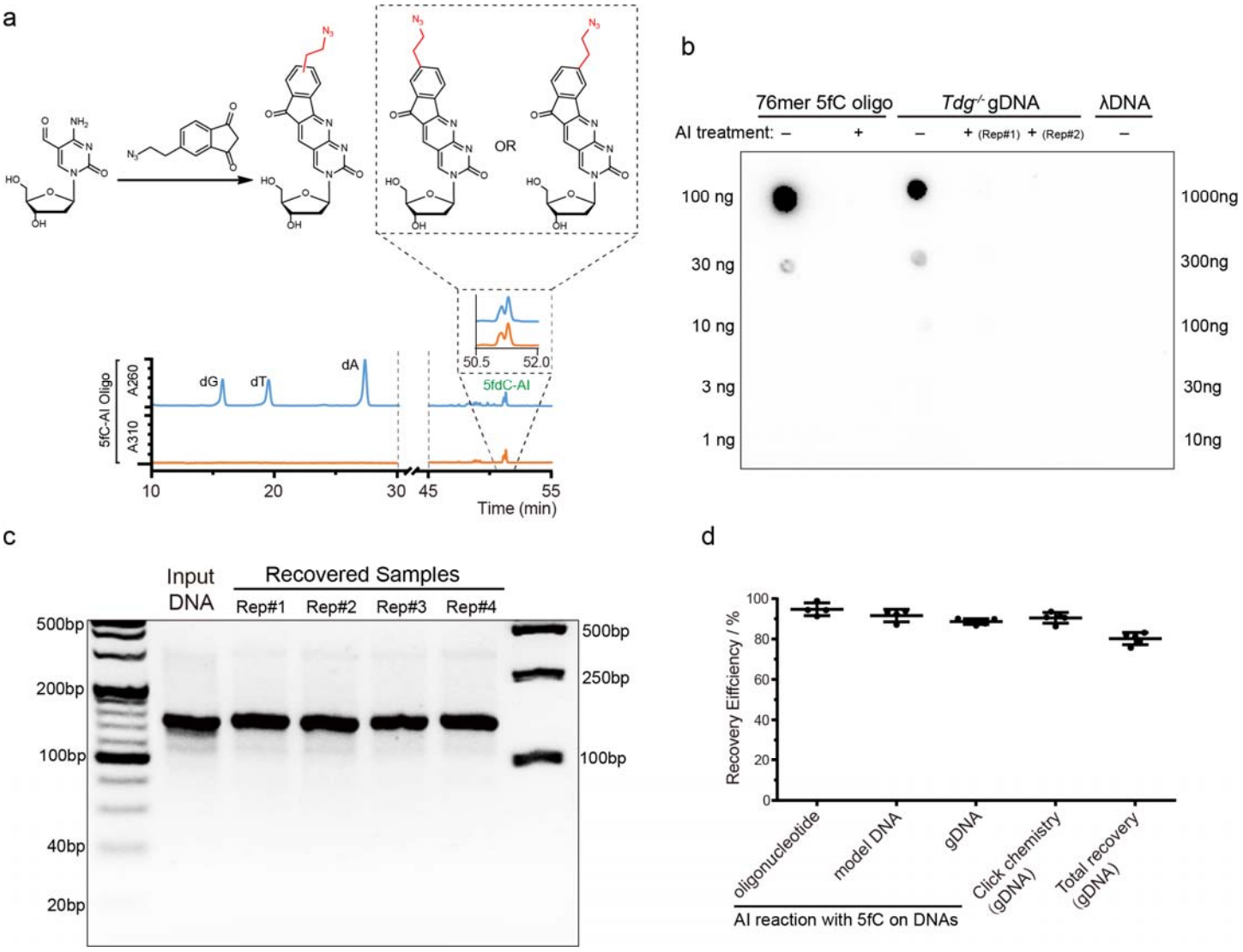
(a) 5fC labeling with diethyl malonate protects the product from bisulfite-mediated deamination and hence is read as “C” in canonical bisulfite sequencing, similar to our previously reported fCAB-Seq[8]. **(b)** The adduct between 5fC and 1,3-indandione results in C-to-T transition during PCR. **(c)** Single nucleotide incorporation of dATP, dGTP, dCTP and dTTP opposite the 5fC-I for 1, 2, 5, 10, 30 and 60 min, respectively. A template where X is a T was also included as a positive control. **(d)** Quantification of single nucleotide incorporation. Values were presented as mean ± SD (*n* = 3). **(e)** Schematic comparisons of fC-CET with fCAB-Seq.



Supplementary Figure 5

Potential base-pairing properties of different cyclic adducts.

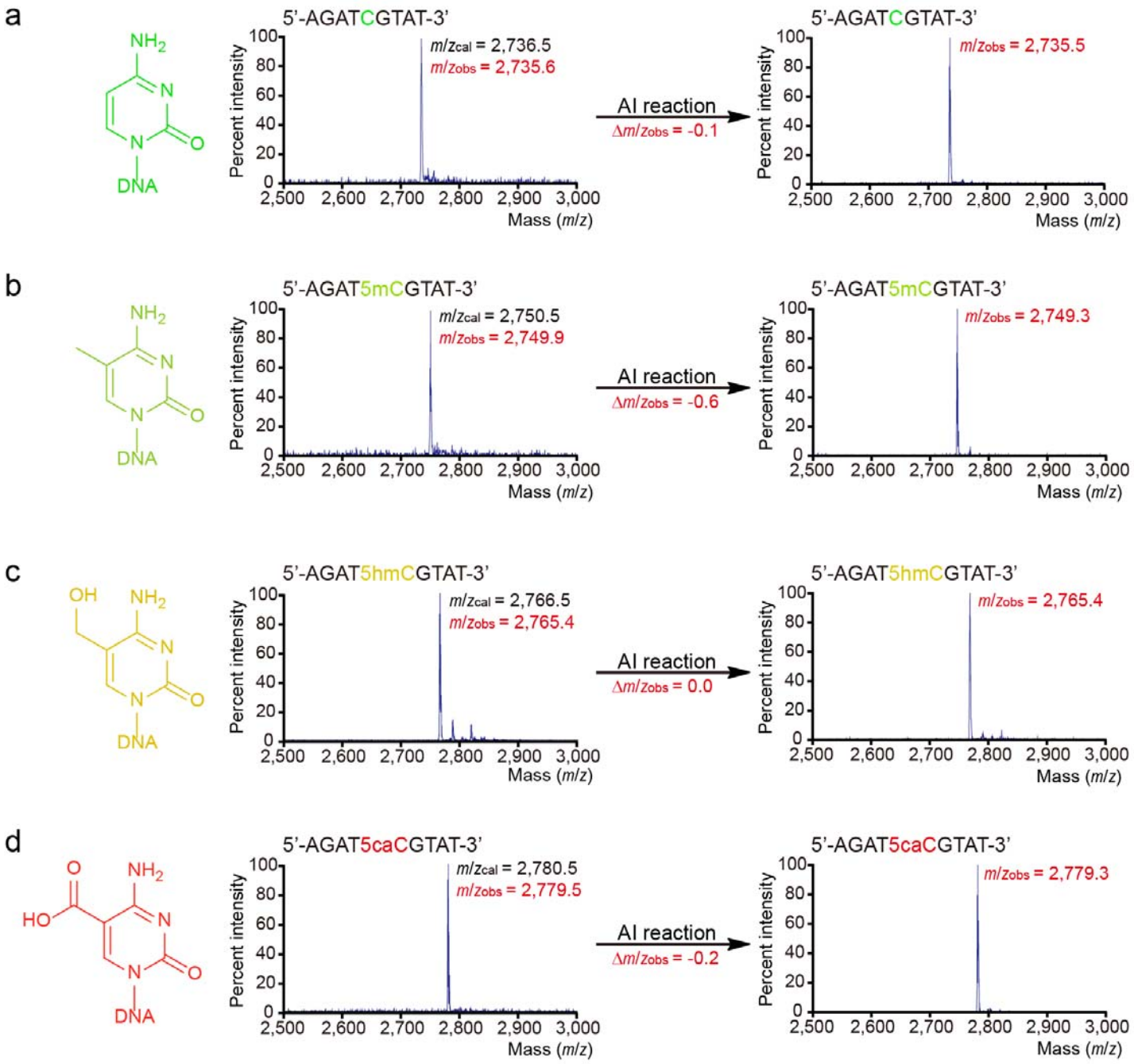
(a) Cyclic adducts between 5fC and ethyl acetoacetate, diethyl malonate or ethyl 6-azido-3-oxohexanoate. These adducts contain a "4-NH" group and are expected to pair with dG in a canonical Watson-Crick fashion. **(b)** Hypothesized base-pairing modes of 5fC-I with dA. Left: N3-protonated 5fC-I could base-pair with dA in the Watson-Crick manner; right: alternatively, the 6-amino group of dA could form a bidentate hydrogen-bond to 5fC-I. **(c)** Acid/base titration curves of 5fC-I solution (in DMSO:H_2O = 5.5:1 solvent). The ^{13}C NMR spectra of the free nucleoside and in 2 equivalents of hydrochoride acid (pH is ~2) are shown in the **Supplementary Note 2**. These data suggest that no protonation events occurred to the free 5fC-I nucleoside.



Supplementary Figure 6

Monitoring the efficiency of AI-mediated 5fC labeling on both model sequence and genomic DNA.

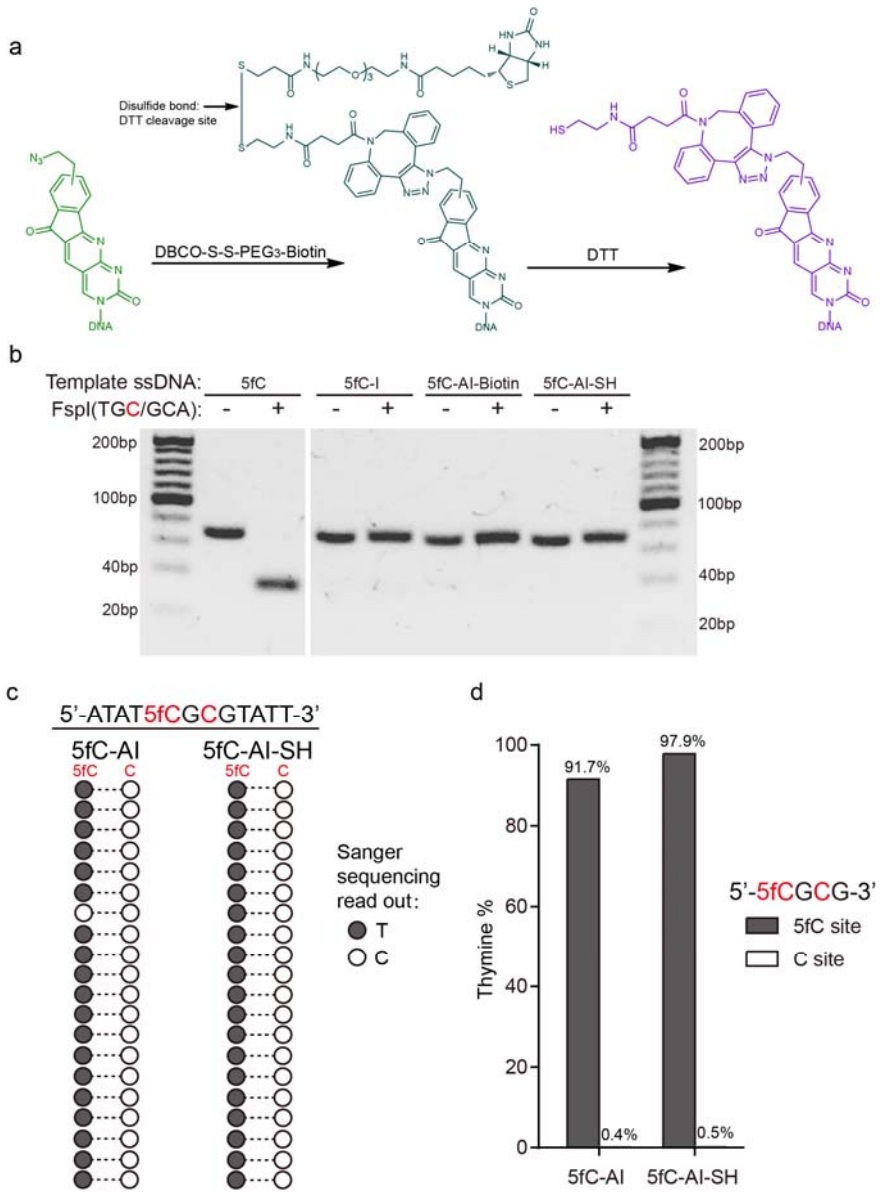
(a) AI labeling of 5fC generates a pair of isomers. **(b)** HPLC chromatograms (260 nm and 310 nm) of the nucleosides digested from the 9-mer 5fC-AI oligonucleotides. Similar to 1,3-indandione, the reaction goes to completion with no detectable 5fC signal after labeling. The split peaks of the two reaction products are also shown. **(c)** Dot-blot assay to monitor the efficiency of 5fC labeling on both model sequence (starting from 100 ng) and *Tdg*$^{-/-}$ mESC gDNA (starting from 1000 ng). Lambda DNA (λDNA, starting from 1000 ng) was used as a negative control. **(d)** Agarose gel analysis of AI labeled double-stranded 5fC model DNAs (4 replicates were shown). **(e)** Recovery efficiency for each step of fC-CET. 9-mer 5fC oligonucleotide, 5fC model DNA or gDNA samples were used to characterize the recovery efficiency. Values were presented as mean ± SD (*n* = 4 for oligonucleotide and model DNA; *n* = 5 for gDNA).



Supplementary Figure 7

AI-mediated labeling of 5fC is highly selective among cytosine derivatives.

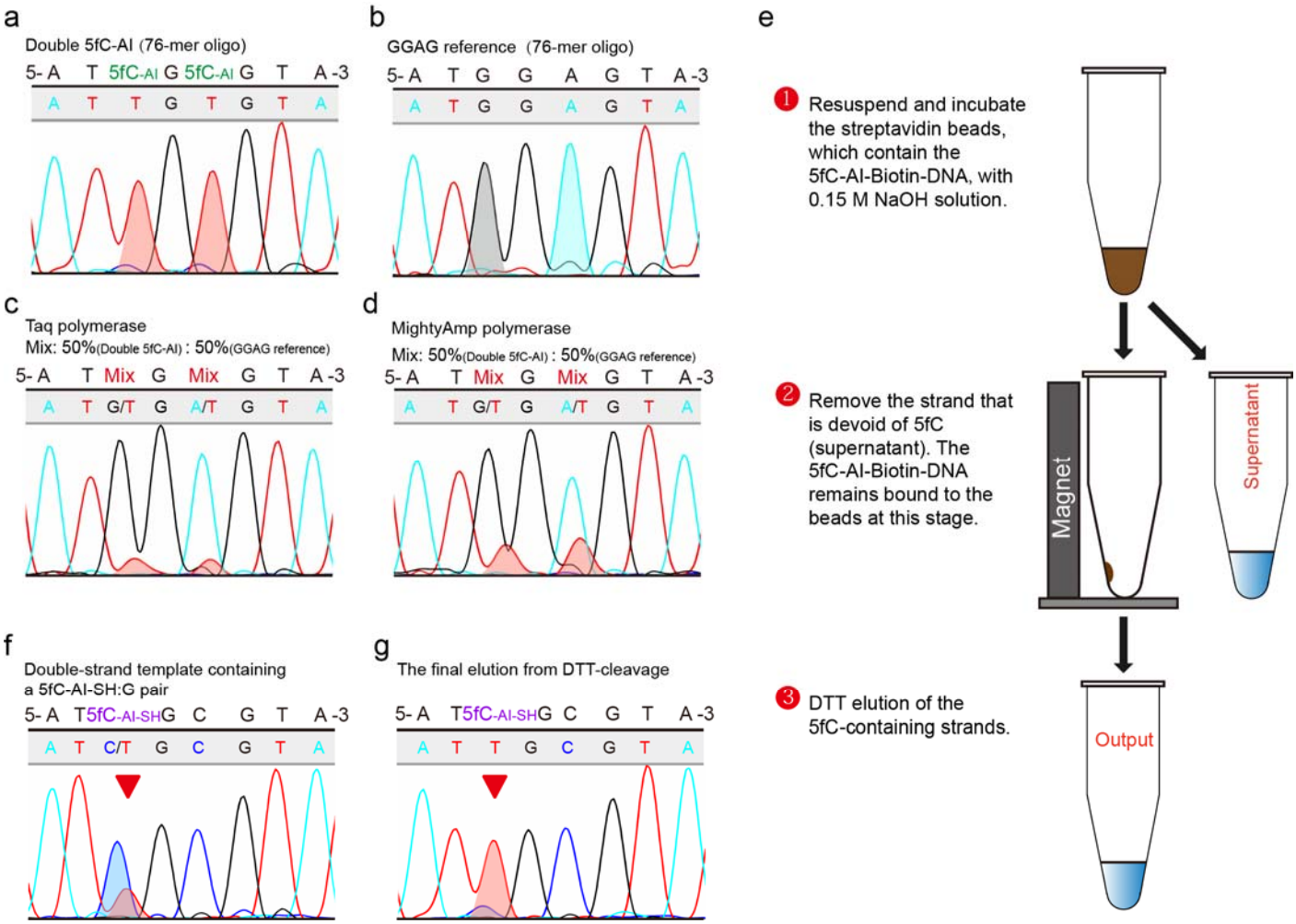
9-mer DNAs with a central C **(a)**, 5mC **(b)**, 5hmC **(c)** and 5caC **(d)** were labeled under the same conditions as 5fC and then analyzed with MALDI-TOF. No cross-reactivity was observed for these cytosines.



Supplementary Figure 8

AI labeling of 5fC and subsequent C-to-T transition.

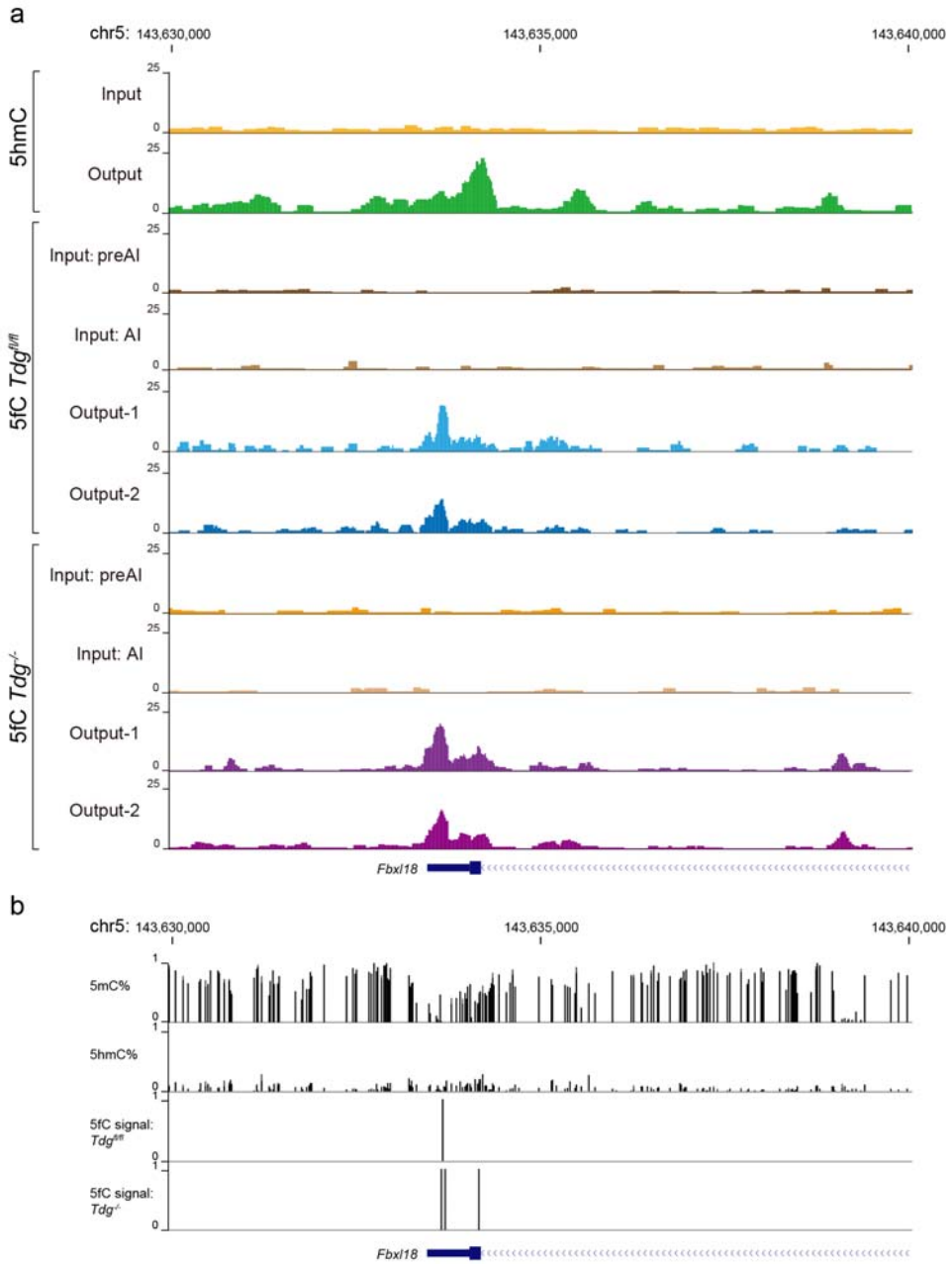
(a) Biotin conjugation for pull-down and DTT cleavage. **(b)** FspI-digestion tests to confirm the C-to-T transition caused by AI-mediated 5fC labeling. PCR-amplified input 70-mer 5fC oligo with FspI restriction site in the middle can be cleaved, while amplified products of 5fC-AI-Biotin or 5fC-AI-SH stay intact, suggesting loss of restriction site and hence C-to-T transition. **(c,d)** Efficiency of 5fC labeling on a model DNA. T%, calculated from TOPO-cloning **(c)** or high-throughput sequencing using Mi-Seq **(d)**, respectively, was used to measure C-to-T transition rate of 5fC after AI labeling and AI-mediated pulldown. In both cases, a normal C next to the target 5fCpG was used as a control.



Supplementary Figure 9

Choice of polymerase to maximize C-to-T signals during PCR.

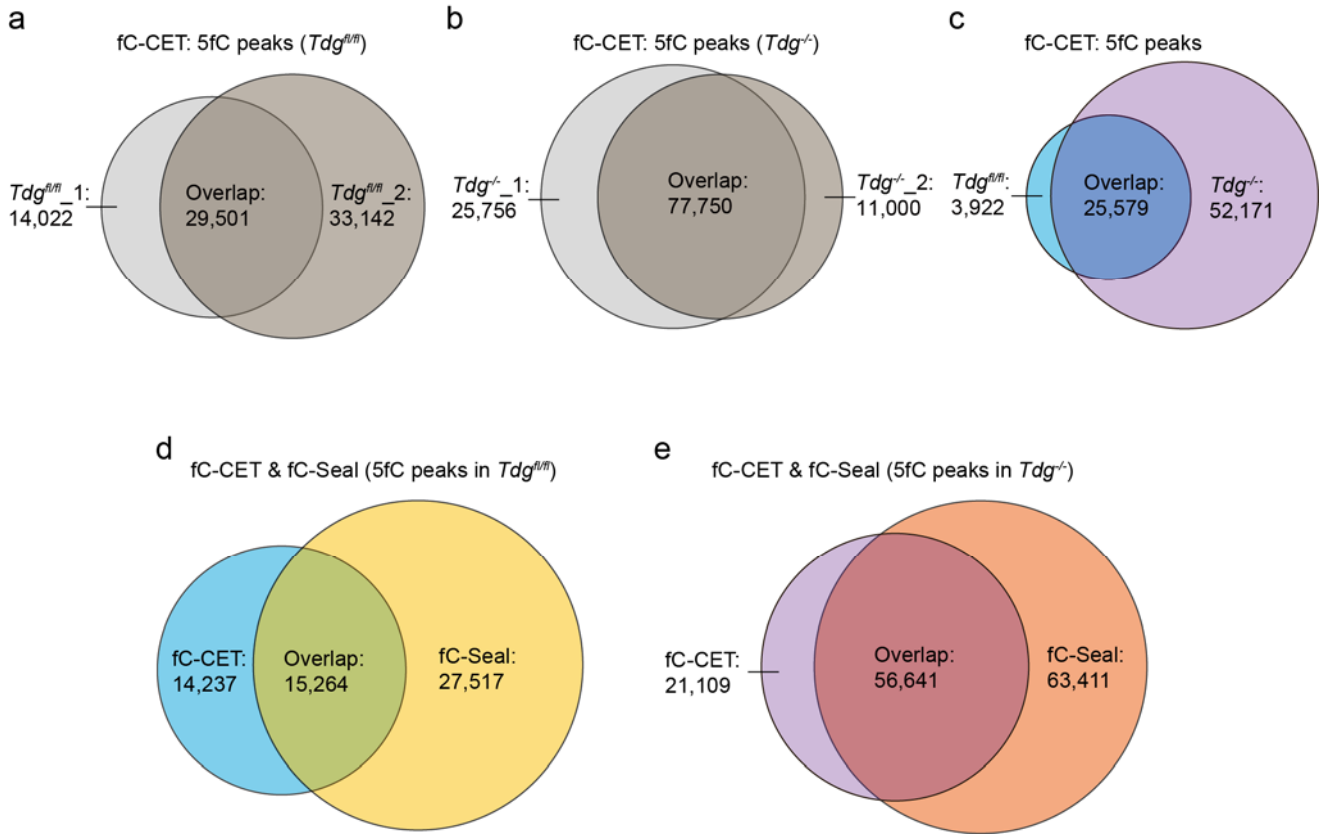
(a) 76-mer oligonucleotide with two 5fC sites labeled with AI, PCR amplified and then subjected to Sanger sequencing. The graph shows that the polymerase can readily read through the modified sequence even with two successive 5fC-AI adducts next to each other (in a 5'-fCGfCG-3' context). **(b)** 76-mer control oligonucleotide with G and A replacing the two 5fCs in the sequence in (a). **(c)** Sanger sequencing results of Taq polymerase reading through the adduct between 5fC and AI. PCR bias will diminish the effective C-to-T signals; hence different polymerases were tested to maximize the signal for 5fC detection. Alternatively, new labeling probes could be explored to minimize chemical scarring on the 5fC base[31]. **(d)** Commercially available MightyAmp DNA polymerase best overcomes the PCR bias and hence allows maximal C-to-T transition during PCR. **(e)** Scheme diagram of the on-bead wash step (with NaOH solution) to remove any strands that do not contain 5fC ("supernatant" depicted in "Step 2"). Therefore, the output (depicted in "Step 3") is enriched for 5fC-AI-SH. **(f)** AI labelled 76bp dsDNA with a central 5fC-AI-SH:G pair. The C-T transition peak can be observed at the 5fC site, although the signal of C (resulting from the G in the complementary strand) is higher than T (from 5fC-AI-SH). **(g)** Elution from the DTT-cleavage step. The eluted fraction is enriched for 5fC-AI-SH, and hence a near complete C-to-T transition peak can be observed at the 5fC site.



Supplementary Figure 10

Genome browser view of 5fC at the Fbxl 18 gene.

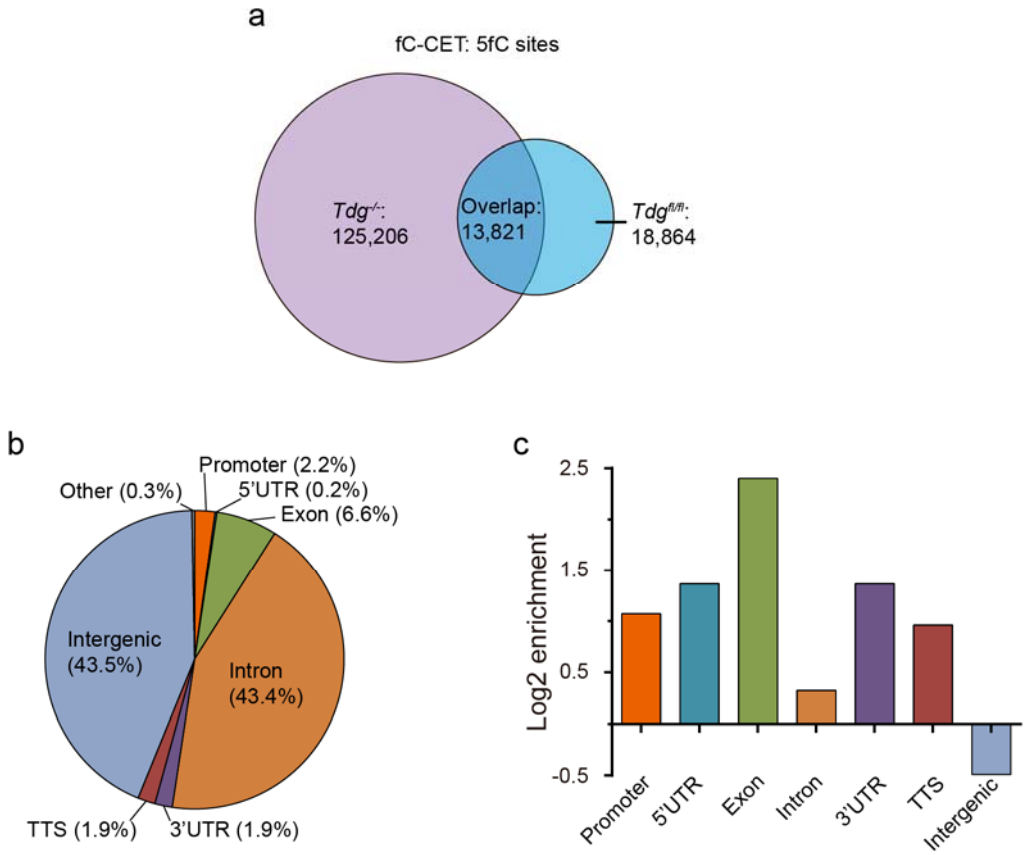
(a) 5fC-enriched regions in both *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mESCs. Results from two replicates are shown, demonstrating high reproducibility of fC-CET. 5hmC-enriched peaks by hmC-Seal are also shown. **(b)** Single-base 5fC sites, along with 5mC and 5hmC, in *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mESCs are shown. The peaks of 5fC corresponds to 5mC sites with low abundance.



Supplementary Figure 11

Venn diagrams of the 5fC-enriched regions.

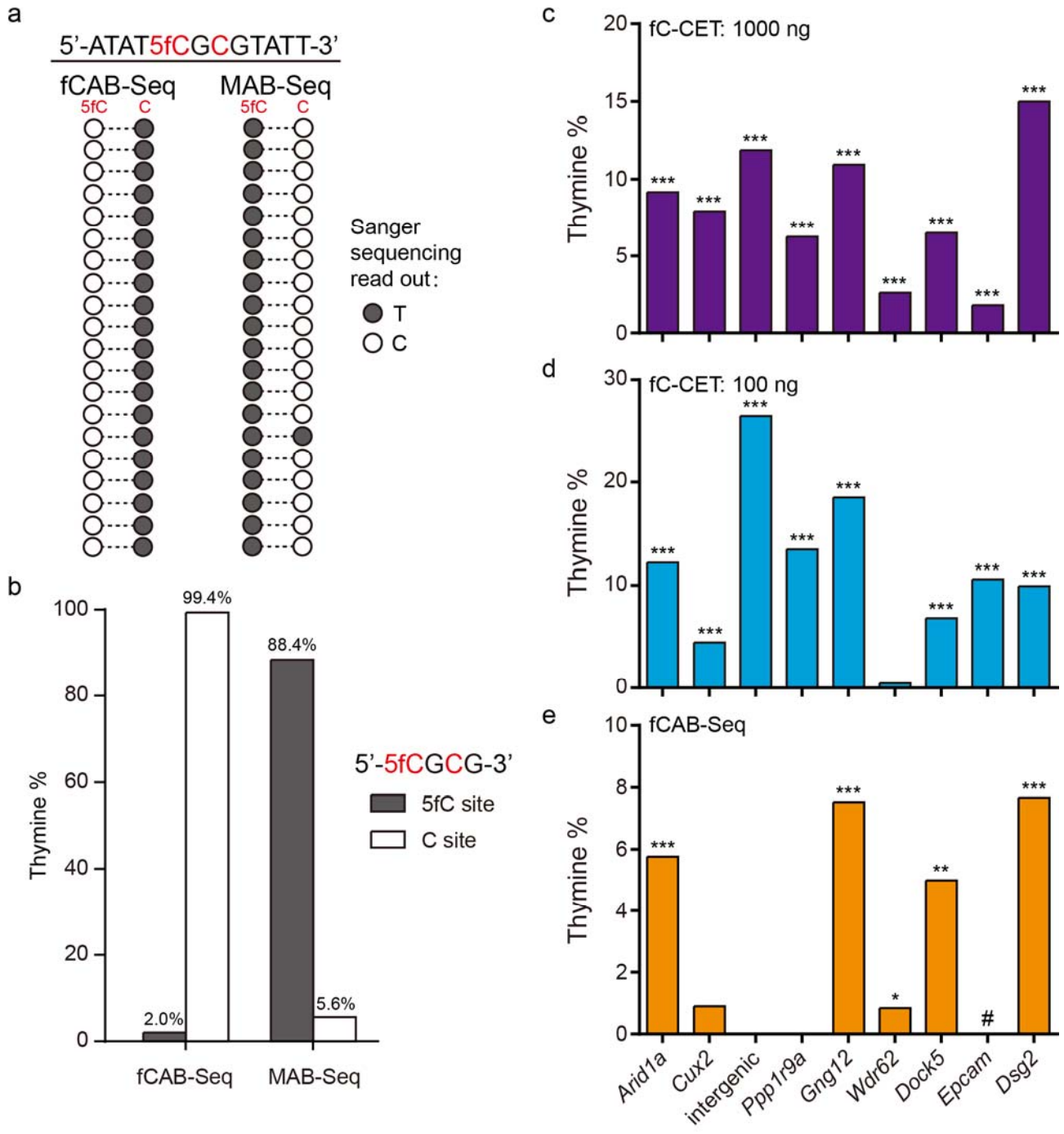
(a, b) 5fC-enriched regions in two biological replicates of *Tdg*$^{fl/fl}$ (a) and *Tdg*$^{-/-}$ (b) mESCs, respectively. **(c)** Majority of 5fC-marked regions in the *Tdg*$^{fl/fl}$ mESCs fall within those in *Tdg*$^{-/-}$ mESCs. **(d,e)** Comparisons of 5fC-enriched regions detected from fC-CET with results from fC-Seal in the *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mESCs.



Supplementary Figure 12

5fC sites at single-base resolution.

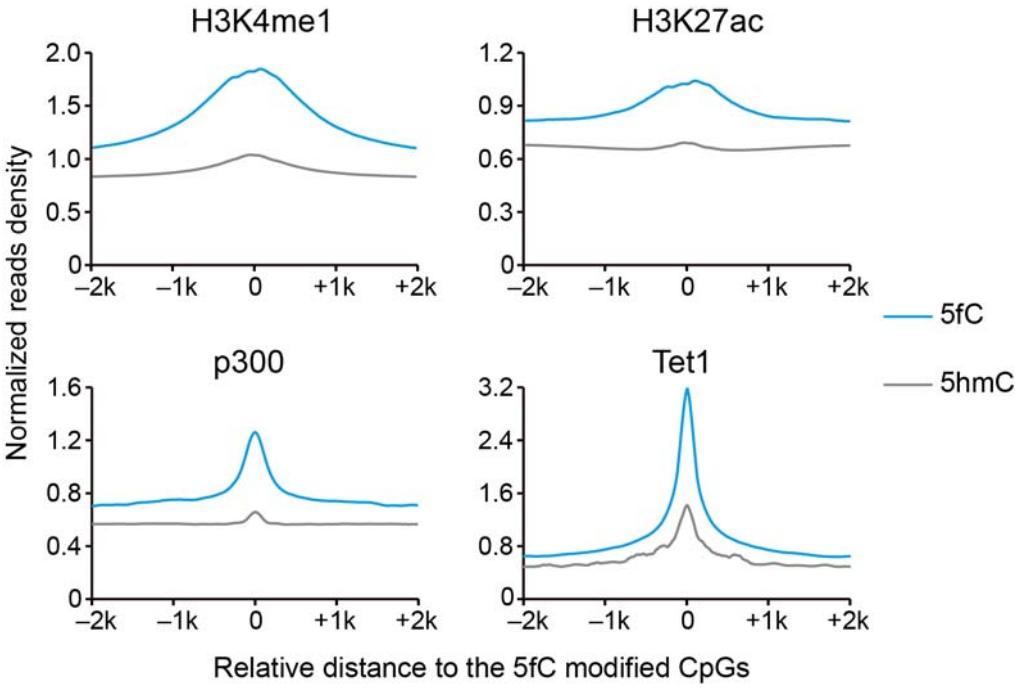
(a) Comparisons of 5fC sites in the *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mESCs. The smaller overlap of 5fC sites on the single-base level than on the enriched-region level hints that TDG-mediated excision of 5fC might be a very dynamic process. **(b)** 5fC sites in *Tdg*$^{-/-}$ mESCs are grouped based on genomic elements. **(c)** The relative enrichment of *Tdg*$^{-/-}$ 5fC sites in different genomic elements.



Supplementary Figure 13

Loci-specific validation of 5fC sites.

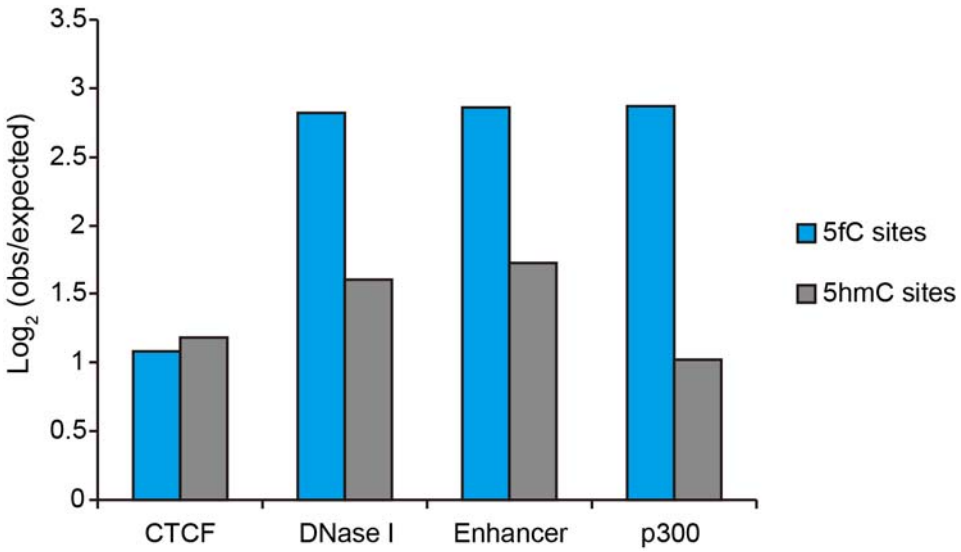
(a,b) fCAB-Seq[8] and MAB-Seq[13] were performed (in parallel with fC-CET) on the 76-mer 5fC model sequence. Results from TOPO-cloning assay **(a)** and high-throughput sequencing **(b)** are shown. In both cases, a normal C on the 5fCpG nearby was used as a control. **(c–e)** Loci-specific validation of 5fC sites with fC-CET **(c,d)** and fCAB-Seq **(e)**. The columns represent the percentage of "T" at the target 5fC sites. For each site, a p value was given to show whether it is a statistically significant 5fC site (binomial test for **c** and **d**, and Fisher's test for differences between fCAB and BS datasets in **e**). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; **#**, failed to detect because of technical issues.



Supplementary Figure 14

5fC sites in *Tdg*$^{-/-}$ mESCs at different regulatory elements.

Normalized read densities of 5fC in *Tdg*$^{-/-}$ mESCs at the H3K4me1, H3K27ac, p300 and Tet1 regions, respectively. The signals of 5fC at such regions are higher than 5hmC, suggesting 5fC represents a more active marker compared to 5hmC.



Supplementary Figure 15

5fC is more enriched than 5hmC at different genomic elements.

The genome loci analyzed include CTCF binding region, Dnase I hypersensitive region, enhancer and p300 binding region; the relative enrichment (observed over expected) is used for comparison. The 5fC sites are more enriched than the 5hmC sites in the enhancer domain and p300 binding region, suggesting the more active nature of 5fC marked regions. This is further confirmed by the DNase I hypersensitive region, which represents the more de-condensed genomic regions.

Supplementary Tables

Supplementary Table 1. Model sequences

ID NO.	Sequence (5'-3')	Annotation
1	AGAT5fCGTAT	5fC-9mer
2	AGATCGTAT	C-9mer
3	AGAT5mCGTAT	5mC-9mer
4	AGAT5hmCGTAT	5hmC-9mer
5	AGAT5caCGTAT	5caC-9mer
6	CCTCACCATCTCAACCAATATTATATTACGCGTATAT5fC GCGTATTTCGCGTTATAATATTGAGGGAGAAGTGGTGA	76mer single 5fC
7	CCTCACCATCTCAACCAATATTATATTACGCGTATAT5fC G5fCGTATTTCGCGTTATAATATTGAGGGAGAAGTGGTG A	76mer double 5fC
8	CCTCACCATCTCAACCAATATTATATTATGTCTACACGT T5fCG5fCGTTCCGTGTTATAATATTGAGGGAGAAGTGGT GA	dsDNA for BS-seq
9	CCTCACCATCTCAACCAATATTATATTATGTCTACACGT TGGAGTTCCGTGTTATAATATTGAGGGAGAAGTGGTGA	76mer control DNA
10	CCTCACCATCTCAACCAATATTATATTAGTATTG5fCGCA TACGCGTTATTATATTGAGGGAGAAGTGGTGA	FspI digest
11	CCTCACCATCTCAACCAATA	Model-F
12	CCCTTT TATTATTTTAATTAATATTATATT	Model-BS-F
13	TCACCACTTCTCCCTCAAT	Model-R
14	CTCCGACATTATCACTACCATCAACCACCCATCCTACCT GGACTACATTCTTATTCAGTATTCACCACTTCTCCCTCA AT	Model-Seq-R
15	CTCCGACATTATCACTACCA	Sanger Sequencing Primer
16	AAATCAXCCTATCCTCCTTCAGGACCAACGTAC X = 5fC-I or T	Primer extension: templates
17	CGTTGGTCCTGAAGGAGGATAGG	Primer extension: primer

Supplementary Table 2. Model sequences for quantitative PCR

ID NO.	Sequence (5'-3')	Annotation
18	CATGAGTGCCCTCAGCAGTAAGTAACTGACCAGATCTC TCGTGCCTCTTGAGGCTACTGAGTTATCCAACCTTTAG GAGCCATGCATCGATAGCATCCG5fCCACAGGCAGTGA GGCTACTGAGTCATGCACGCAGAAAGAAATAGC	qPCR-Single 5fC, dsDNA
19	CATGAGTGCCCTCAGCAGTAAGTAACTGACCAGATCTC TCGTGCCTCTTGAGGCTACTGAGTTATCCAACCTTTAG GAGCCATGCATCGATAGCATCCGCCACAGGCAGTGAG GCTACTGAGTCATGCACGCAGAAAGAAATAGC PCR amplified with 100% dATP, 100% dTTP, 100% dGTP, 90% dCTP and 10% dfCTP.	qPCR-10% 5fC, dsDNA
20	ATTCACTCCCACTGAGACTGTGGATCAGGCCAACATAC ATGCCTTCAGTAACTGACCAGATCTCTTAGTTCTCTTGA GGCTACTGAGTTAGAATGGCAGAGTCAAGGAGC PCR amplified with 100% dATP, 100% dTTP, 100% dGTP, 70% dCTP, 15% dmCTP，10% dhmCTP and 5% dcaCTP.	qPCR-Ctl, dsDNA
21	CTACGCAAACTGGCTGTCAAAGTAACTGACCAGATCTC TCGGCTCTCTTGAGGCTACTGAGTTATCATGGACGCTA CCTCACAG	qPCR-Ref, dsDNA
22	CATGAGTGCCCTCAGCAGTA	qPCR-5fC-F
23	TCCAACCTTTAGGAGCCATG	qPCR-5fC-R
24	AGGCCAACATACATGCCTTC	qPCR-Ctl-F
25	GAATGGCAGAGTCAAGGAGC	qPCR-Ctl-R
26	CTACGCAAACTGGCTGTCAA	qPCR-Ref-F
27	CTGTGAGGTAGCGTCCATGA	qPCR-Ref-R

Supplementary Table 3. Primers for loci-specific 5fC sites validation

Chr	5fC Position	Genes	PCR primers (5'-3')	
chr4	133272266 (-)	*Arid1a*	BS-F	ATTGTAGGGAAGGAGTTTAGGATAG
			BS-R	ACTTCAAACAATCCTCTAAAAAAAA
			F	GATCAGTTTATAGCCCATCCTG
			R	GTCCCGGCTAGTGTTTCATC
chr5	122444219 (-) *	*Cux2*	BS-F	TTTTTTGTTGGTTTTGTGTTAGAGT
			BS-R	AACACACCTCTCCACCTCACTATAC
			F	GTTGACGGCTGGATTAGAATG
			R	AGAACCGGTGTTCTAAAGCT
chr6	12065399 (-)	Intergenic	BS-F	GGGGGTAGAATATGTAAATGGAATT
			BS-R	ACCCAAATAATCCTTAAAACTCCTC
			F	GCAGAACATGTAAATGGAACCT
			R	AGCTTACGTGAAGTGGAAAC
chr6	4918473 (+)	*Ppp1r9a*	BS-F	GGTATTTTTGTGTTGTAGTTTGTTG
			BS-R	AAATCTACATAAAAAACCATTAAATC
			F	AACACCTTTACGCAGGGTAG
			R	AATCTGCATGGGGAACCATT
chr6	66854376 (-) *	*Gng12*	BS-F	TGTAGATTATTTGTGGGGATATATT
			BS-R	CATCAACTTCCCATTATACAAATAC
			F	ATACTTCCCGAAGGTGCTC
			R	CTGGGGATTGTATCTGATCCT
chr7	31027944 (-)	*Wdr62*	BS-F	GGGTGAGTTTGGATTTTTTTATTG
			BS-R	AACCATTACCCAAACCTAACAACT
			F	CGGGTGAGTCTGGATTTCTC
			R	CTACCTTCAGTAGGGGCATC
chr14	68371503 (-) *	*Dock5*	BS-F	GAGGGTTAGGTATAGTTTTAGTTTTTTG
			BS-R	AATAACCATTATCCCATCTCAACTC
			F	GCCTGCTATACAATTCACGC
			R	CCTTGTTTTACCCAGGGGAT
chr17	88039777 (-) *	*Epcam*	BS-F	(Failed to amplify)
			BS-R	(Failed to amplify)
			F	CTTGTCGGTTCTTCGGACT
			R	CATTGGGCGTTACTGTCATC
chr18	20748620 (+) *	*Dsg2*	BS-F	AAGTAGTTTTTTATAAGTTTATTTTGAGT
			BS-R	AAATCTATCCATTACCTCACTAACTC
			F	GTCAAGAACGTGGTTGAAGG
			R	ACTTTAGCTGCTTGACCAGT

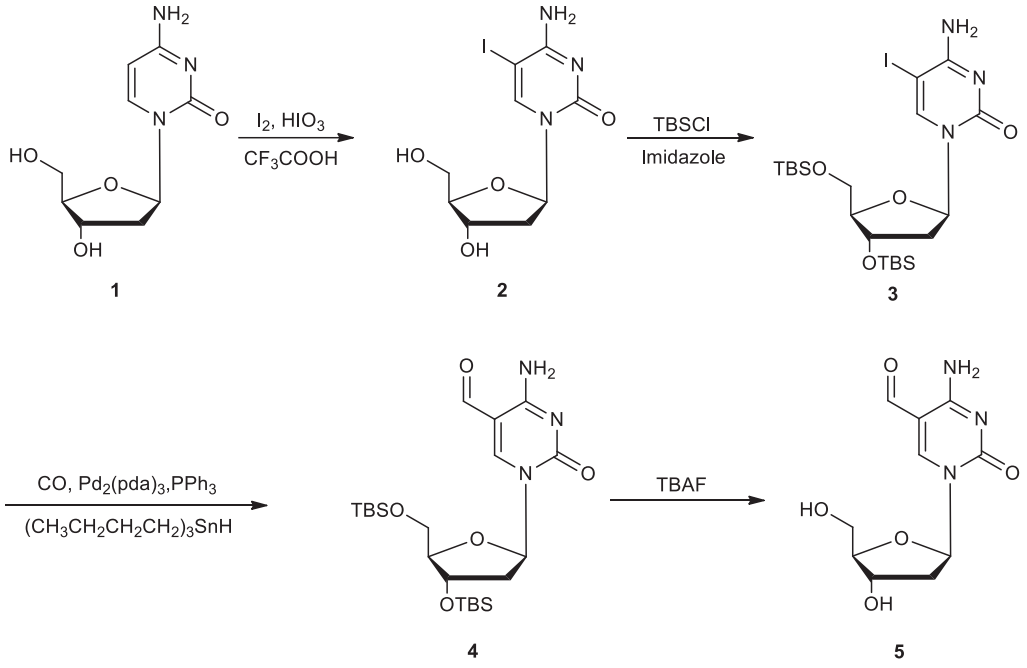
*, 5fC sites detected by previous methods.

Supplementary Notes

Supplementary Note 1: Synthesis of the 5-formyl-2'-deoxycytidine characterizations of related compounds

Synthesis of 5-formyl-2'-deoxycytidine

The synthesis of the 5-formyl-2'-deoxycytidine was based on a previous report[32].



5-iodo-2'-deoxycytidine (2): To a solution of **1** (2'-deoxycytidine, 5.0 g, 22 mmol) in the mixture of CF_3COOH (12.5 mL) and water (25 mL), HIO_3 (2.0 g, 11.3 mmol), I_2 (3.0 g, 11.8 mmol) and dichloromethane (25 mL) were added sequentially. The mixture was stirred for 10 h at 20 °C. The aqueous layer was separated and washed with dichloromethane (25 mL). The residue after evaporation was purified by silica gel chromatography, eluting with 10% methanol and 0.5% ammonium hydroxide in dichloromethane to give **2** (5.6 g, 72.2%) as a white powder. ^{1}N NMR (500 MHz, DMSO-*d6*): δ 8.28 (s, 1H), 7.77 (s, 1H), 6.57 (s, 1H), 6.08 (t, *J* = 6.4 Hz, 1H), 5.18(d, *J* = 4.2 Hz, 1H), 5.07 (t, *J* = 6.4 Hz, 1H), 5.07

(t, *J* = 4.90 Hz, 1H), 4.21 (m, 1H), 3.79 (q, *J* = 3.4 Hz, 1H), 3.63 (m, 1H), 3.55 (m, 1H), 2.14 (m, 1H), 2.00 (m, 1H). ^{13}C NMR (125.8 MHz, DMSO-*d6*): δ 164.1, 154.3, 147.8, 87.9, 85.8, 70.4, 61.3, 56.9, 41.3. HRMS (*m*/*z*): $[M+H]^+$ calcd. for $C_9H_{13}IN_3O_4$, 353.99449; found 353.99453.

5-iodo-3',5'-(tert-butyldimethylsilyl)-2'-deoxycytidine (3): A solution of **2** (2.0 g, 5.6 mmol) in *N*,*N*-dimethylformamide (30 mL), imidazole (2.0 g, 29 mmol) and tert-butyldimethylchlorosilane (2.7 g, 17.9 mmol) was stirred for 2h at 80 °C. The reaction was quenched with methanol (5 mL) and stirred for an additional 10 min. The reaction mixture was poured into water, extracted with ethyl acetate (2 × 125 mL) and dried over anhydrous $MgSO_4$. After the organic phase was concentrated to dryness, the residue was recrystallized from the mixture of petroleum ether and ethyl acetate (1:1) to give **3** (2.64 g, 79.9%) as a white powder. ^{1}N NMR (500 MHz, $CDCl_3$): δ 8.06 (s, 1H), 7.87 (s, 1H), 6.23 (t, *J* = 6.5 Hz, 1H), 5.51 (s, 1H), 4.35 (m, 1H), 3.97 (d, *J* = 2.6 Hz, 1H), 3.88 (dd, *J* = 11.4 Hz, 2.5 Hz, 1H), 3.75 (dd, *J* = 11.4 Hz, 2.5 Hz, 1H), 2.46 (ddd, *J* = 13.2 Hz, 5.8 Hz, 3.0 Hz, 1H), 1.95 (m, 1H), 0.93 (s, 9H), 0.88 (s, 9H), 0.13 (s, 3 H),0.12 (s, 3H), 0.07 (s, 3H), 0.06 (s, 3H). ^{13}C NMR (125.8 MHz, $CDCl_3$): δ 163.6, 154.7, 146.6, 88.3, 86.8, 72.2, 62.9, 55.8, 42.7, 26.1, 25.9, 25.7, 18.5, 18.0, -4.60,-4.88, -5.14, -5.26. HRMS (*m*/*z*): $[M+H]^+$ calcd. for $C_{21}H_{41}IN_3O_4Si_2$, 582.16748; found 582.16711.

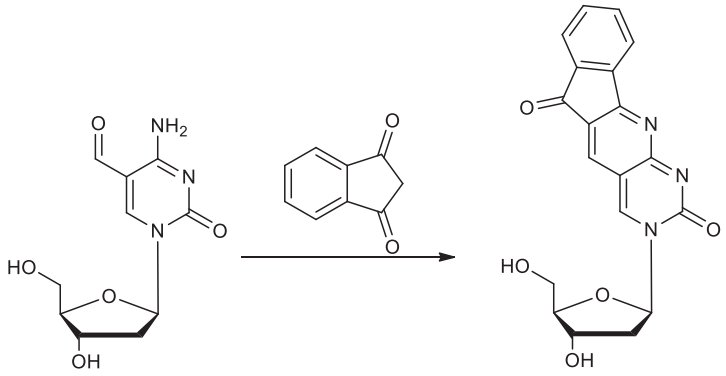
5-formyl-3',5'-(tert-butyldimethylsilyl)-2'-deoxycytidine (4): To a solution of **3** (1.0 g, 1.7 mmol) in toluene (40 mL), $Pd_2(dba)_3{\cdot}CHCl_3$ (0.35 g,0.34 mmol) and

triphenylphosphine (0.53 g, 2.0 mmol) were added in a high pressure glass autoclave. The autoclave was flushed with CO to remove residual air, and the reaction was stirred under a CO pressure of 3.0 bar for 10h at 60 °C. Then, tributyltin hydride (1.2 mL, 3.4 mmol) was added and stirred within 4 h. The product was evaporated and then purified by silica gel column chromatography, eluting with 1% methanol in dichloromethane to give **4** (0.76 g, 92%). ^{1}N NMR (500 MHz, $CDCl_3$): δ 9.48 (s, 1H), 8.53 (s, 1H), 8.17 (s, 1H), 6.62 (s, 1H), 6.20 (t, *J* = 6.2 Hz, 1H), 4.35 (m, 1H), 4.04 (m, 1H), 3.95 (dd, *J* = 11.5 Hz, 2.6 Hz, 1H), 3.78 (dd, *J* = 11.5 Hz, 2.5 Hz, 1H), 2.61(ddd, *J* = 13.5 Hz, 6.1 Hz, 3.9 Hz, 1H),2.07 (m, 1H), 0.90 (s, 9H), 0.89 (s, 9H), 0.10 (s, 3 H),0.09 (s, 3H), 0.08 (s, 3H), 0.07 (s, 3H). ^{13}C NMR (125.8 MHz, $CDCl_3$): δ 187.1, 162.7, 153.3, 153.0, 105.0, 88.7, 87.8, 71.6, 62.6, 55.8, 42.9, 26.0, 25.7, 25.5, 18.5, 18.0, -4.50,-4.88, -5.24, -5.31. HRMS (*m*/*z*): $[M+H]^+$ calcd. for $C_{22}H_{42}N_3O_5Si_2$, 484.26575; found 484.26651.

5-formyl-2'-deoxycytidine (5): A solution of **4** (0.5g, 1.03 mmol) and tetrabutylammonium fluoride trihydrate (0.57 g, 2.16 mmol) was stirred for 2h at 10 °C. Reaction mixture was then evaporated to dryness and purified by silica gel column chromatography, eluting with 1% methanol in dichloromethane to give **5** (0.21 g, 81.8%) as a white solid. ^{1}N NMR (500 MHz, DMSO-*d6*): δ 9.48 (s, 1H), 8.83 (s, 1H), 8.07 (s, 1 H), 7.85 (s, 1 H), 6.07 (t, *J* = 6.1 Hz, 1 H), 5.24 (d, *J* = 4.4 Hz, 1 H), 5.08 (t, *J* = 5.2 Hz, 1 H), 4.23 (dq, *J* = 5.9 Hz, 4.2 Hz, 1H), 3.87 (q, *J* = 3.8 Hz, 1 H), 3.68 (ddd, *J* = 12.0 Hz, 5.3 Hz, 3.8 Hz, 1 H), 2.31 (ddd, *J* = 13.4 Hz, 6.2 Hz, 4.4 Hz, 1 H), 2.10 (dt, *J* = 13.4 Hz, 6.1 Hz, 1H). ^{13}C NMR (125.8 MHz, DMSO-*d6*): δ 189.1, 162.7, 154.9, 153.1, 105.1, 88.4, 86.9,

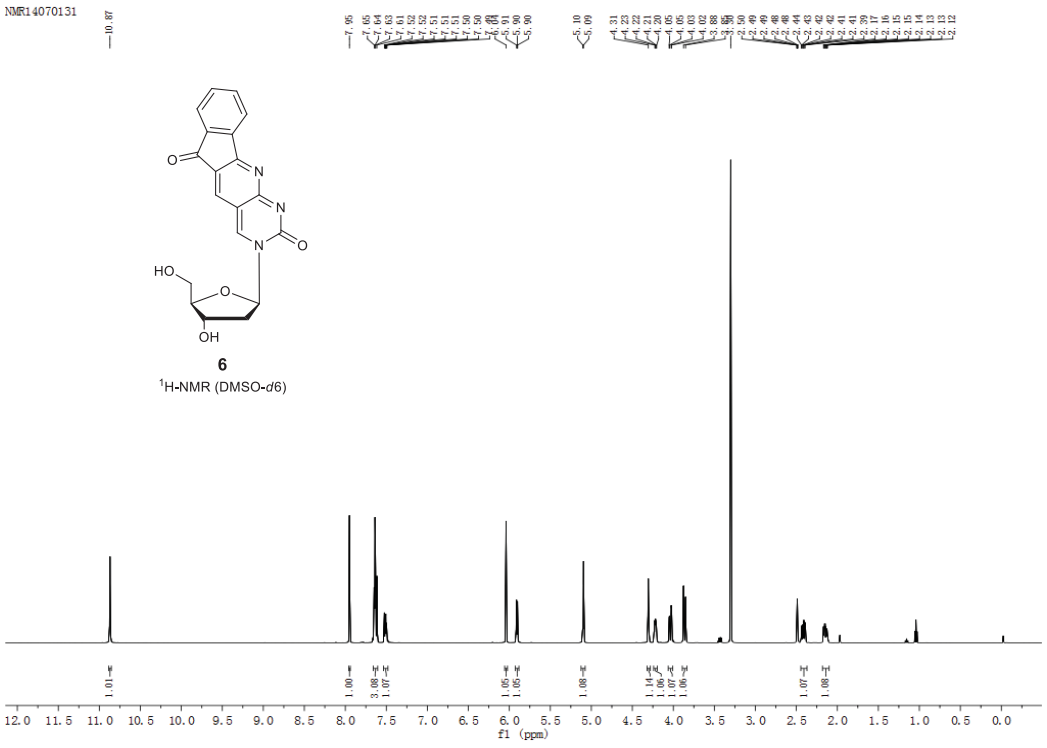
70.0, 61.1, 41.5. HRMS (*m*/*z*): $[M+H]^+$ calcd. for $C_{10}H_{14}N_3O_5$, 256.09280; found 256.09331.

Characterizations of the 5-formyl-2'-deoxycytidine-1,3-indandione adduct (5fC-I)

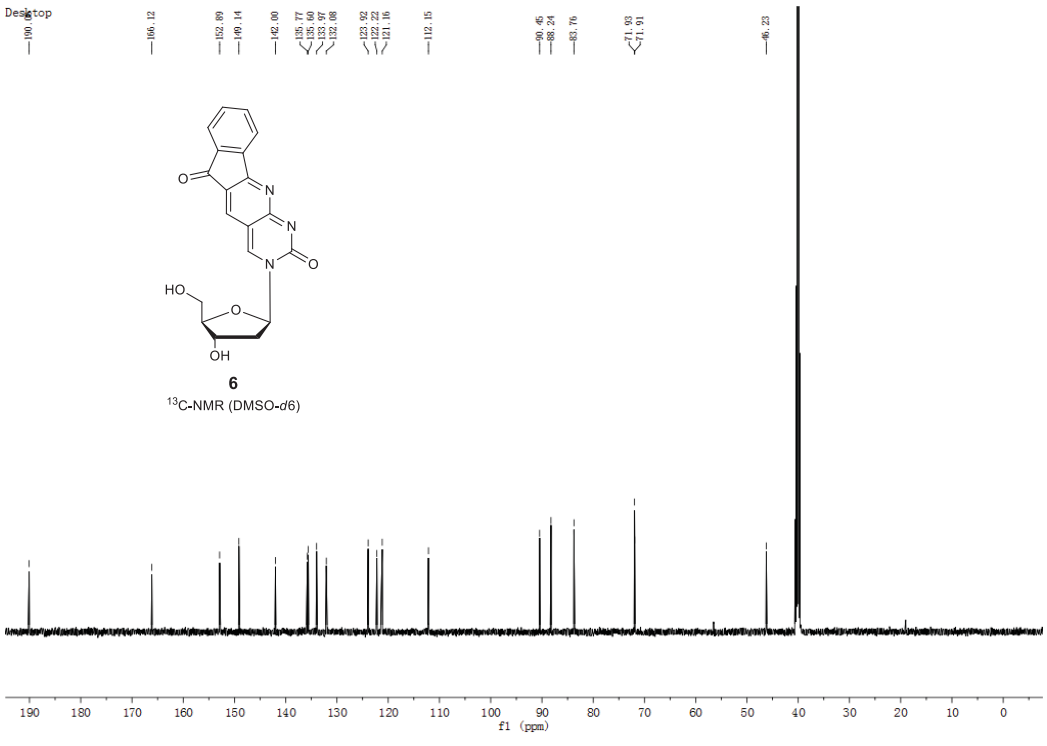


5 **6**

5fC-I (6): A solution of **5** (130 mg, 0.5 mmol), 1,3-indanedione (89 mg, 0.6 mmol) and sodium hydroxide (25 mg, 0.6 mmol) in ethanol (8 mL) was stirred for 12h at 78 °C. The reaction mixture was evaporated to dryness and purified by silica gel column chromatography, eluting with 5% methanol in dichloromethane to give **6** (59.5 mg, 32.1%) as a brown solid. ^{1}N NMR (500 MHz, DMSO-*d6*): δ 10.88 (s, 1H), 7.95 (s, 1H), 7.61-7.65(m, 3H), 7.50(m,1H), 6.04(s,1H), 5.90 (m,1H), 5.10 (d, *J* = 4.4 Hz, 1 H),4.31 (d, *J* = 5.2 Hz, 1 H), 4.22(dt, *J* = 7.3 Hz, 3.6 Hz 1 H), 4.04 (dd, *J* = 12.5 Hz, 2.5 Hz, 1 H), 3.86 (d, *J* = 12.4 Hz, 1 H), 2.42 (m, 1H), 2.14 (ddd, *J* = 14.1 Hz, 6.6 Hz, 3.6 Hz, 1H). ^{13}C NMR (125.8 MHz, DMSO-*d6*): δ 190.1, 166.1, 152.9, 149.1, 142.0, 135.8, 135.6, 134.0, 132.1, 123.9, 122.2, 121.2, 112.1, 90.5, 88.2, 83.7, 71.9, 71.9, 46.2. HRMS (*m*/*z*): $[M+H]^+$ calcd. for $C_{19}H_{16}N_3O_5$, 366.10845; found 366.10881.

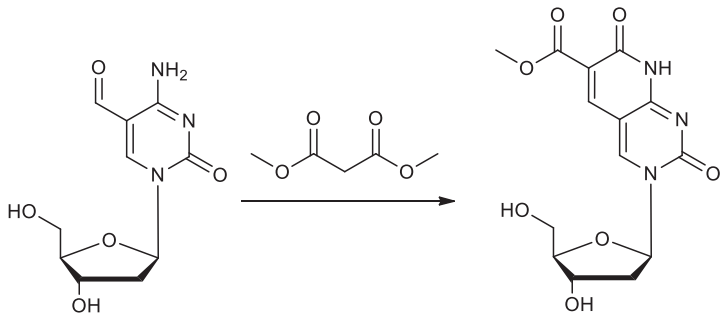


^{1}H NMR (in DMSO-*d6*)



^{13}C NMR (in DMSO-*d6*)

Characterizations of 5-formyl-2'-deoxycytidine-diethyl malonate adduct (5fC-DM)

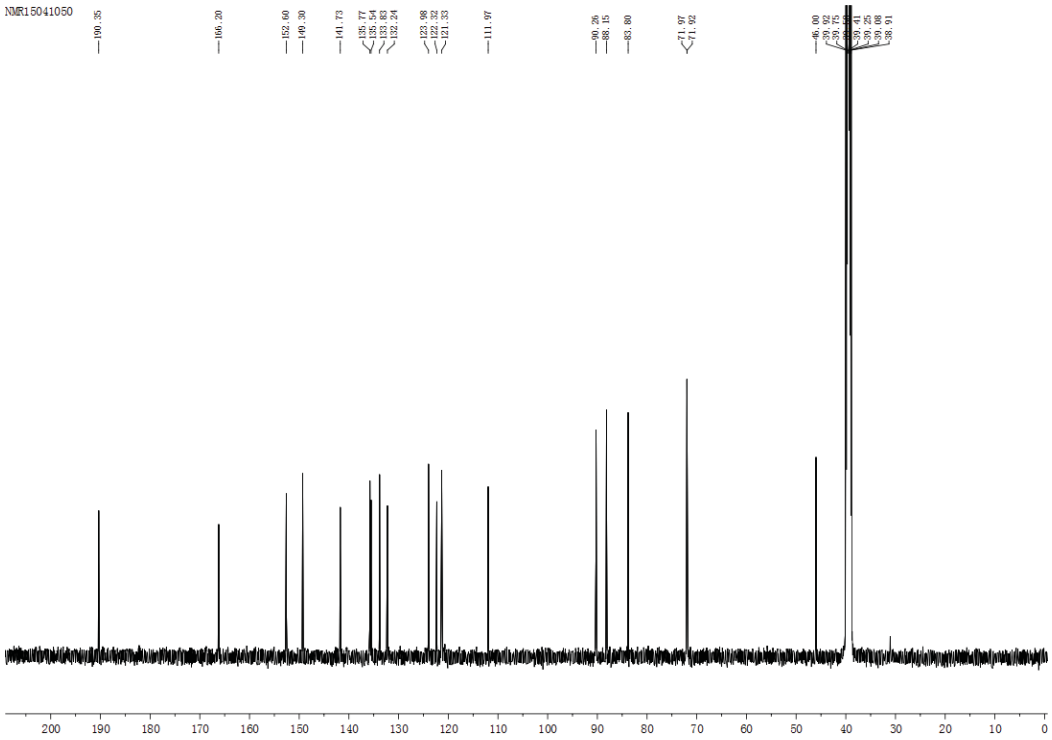


5 **7**

5fC-DM (7): A solution of **5** (30 mg, 0.1 mmol), dimethyl malonate (23 mg, 0.2 mmol) and piperidine (10 mg, 0.1 mmol) in methanol (6 mL) was stirred for 12 h at 68 °C. After the organic phase was concentrated to dryness, the residue was recrystallized from the mixture of ethyl acetate and ethanol (5:1) to give **7** (22 mg, 56.1%) as a light yellow powder. ^{1}N NMR (500 MHz, DMSO-*d6*): δ 11.39 (s, 1H), 9.10 (s, 1H), 8.42 (s, 1H), 6.06 (t, *J* = 6.0 Hz, 1 H), 5.28 (s, 1H), 5.08 (s, 1H), 4.22 (d, *J* = 5.1 Hz, 1 H), 3.95 (d, *J* = 3.8 Hz, 1 H), 3.77 (s, 3H), 3.69 (dd, *J* = 12.0 Hz, 3.3 Hz 1 H), 3.63 (dd, *J* = 12.0 Hz, 3.3 Hz, 1 H), 2.42 (m, 1H), 2.12 (m, 1H). ^{13}C NMR (125.8 MHz, DMSO-*d6*): δ 164.3, 160.5, 159.6, 153.4, 149.3, 144.3, 118.2, 101.1, 89.0, 88.3, 70.0, 63.6, 61.1, 52.4, 41.6. HRMS (*m*/*z*): $[M+H]^+$ calcd. for $C_{14}H_{16}N_3O_5$, 338.09828; found 338.09875.

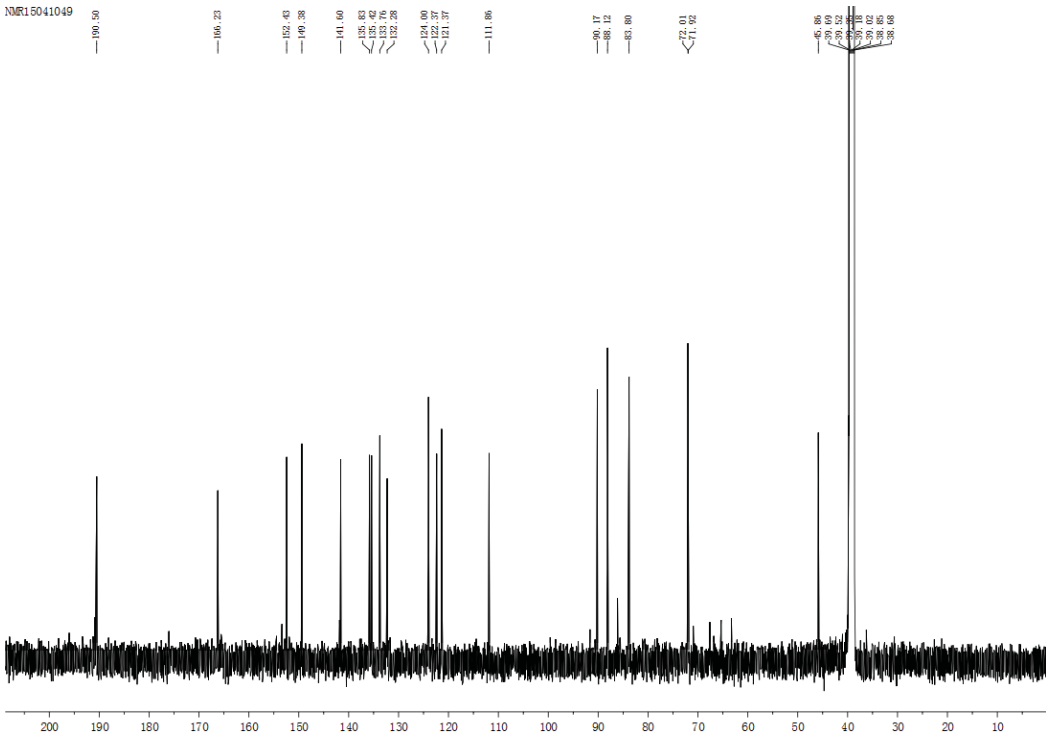
Supplementary Note 2: Acid titration experiments of 5fC-I

Acid titration of 5fC-I solution (in DMSO:H_2O=5.5:1 solvent). The ^{13}C NMR spectra of the 5fC-I free nucleoside in neutral condition or in titrated 2 equivalents of hydrochloride acid (pH is ~2) were shown as below. The unchanged spectra suggest that no protonation events occurred to the free 5fC-I nucleoside.



^{13}C NMR, neutral control

^{13}C NMR (125.8 MHz, DMSO-*d6*:D_2O=5.5:1): δ 190.4, 166.2, 152.6, 149.3, 141.7, 135.8, 135.5, 133.8, 132.2, 124.0, 122.3, 121.3, 112.0, 90.3, 88.2, 83.8, 72.0, 71.9, 46.0.

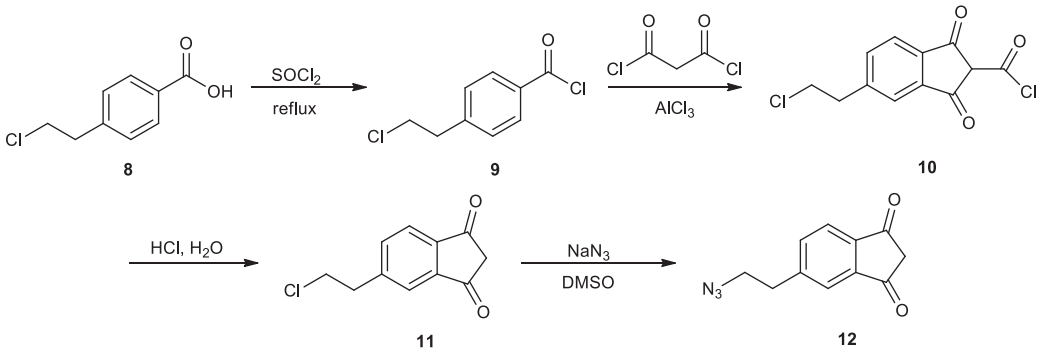


^{13}C NMR, in the presence of 2 equivalents of HCl (pH: ~2)

^{13}C NMR (125.8 MHz, DMSO-*d6*:D_2O=5.5:1): δ 190.5, 166.2, 152.4, 149.4, 141.6, 135.8, 135.4, 133.8, 132.3, 124.0, 122.4, 121.4, 111.9, 90.2, 88.1, 83.8, 72.0, 71.9, 45.9.

Supplementary Note 3: Synthesis of azido chemicals

Synthesis of 5-(2-azidoethyl)-1,3-indandione (AI)



4-(2-chloroethyl)-benzoyl chloride (9): The mixture of **8** (10 g, 108 mmol) and 50 mL $SOCl_2$ was added with 0.5 mL DMF, and then heated to reflux for 2h. The excess $SOCl_2$ was evaporated to give a yellow liquid containing **9** (10.8 g，96%). The liquid was directly used for the next step of synthesis.

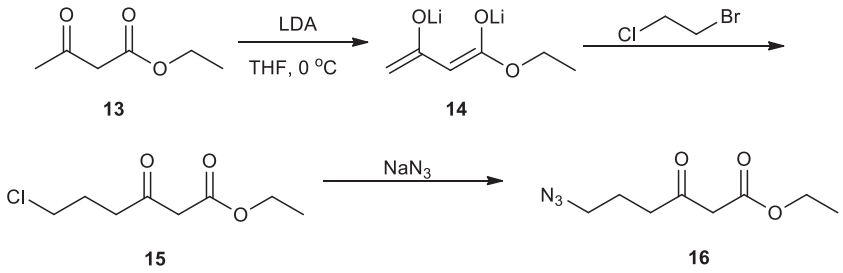
5-(2-chloroethyl)-1,3-indandione (11): The liquid of **9** was resolved in 100 mL CH_2Cl_2, and $AlCl_3$ (14 g, 106 mmol, 1 eq.) was added under nitrogen atmosphere. The mixture was cooled to 0 ^oC, and then the freshly prepared malonyl chloride (16.5 g, 117 mmol, 1.1 eq.) was added slowly to give a dark brown liquid. The reaction was mildly stirred for 2 h at room temperature to give a mixture containing **10**. After reaction, the mixture was poured onto ice, followed by adding 250 mL 10% HCl and vigorous stirring for 1 h. The mixture was extracted with $CHCl_3$ (3 × 400 mL), dried over anhydrous $MgSO_4$ and concentrated for purification with silica gel column chromatography, eluting with 2:1 petroleum ether / CH_2Cl_2 (v/v) to give a light yellow solid **11** (7.9 g，36%). ^1H NMR (300 MHz, $CDCl_3$): δ 7.93 (d, *J* = 7.8 Hz, 1H), 7.83 (s, 1H), 7.71 (d, *J* = 7.8 Hz, 1H), 3.80 (t, *J* = 6.6 Hz, 2H), 3.25 (t, *J*

= 6.6 Hz, 2H), 3.24 (s, 2H). ^{13}C NMR (75 MHz, $CDCl_3$): δ 197.3, 196.8, 146.8, 143.8, 142.2, 136.5, 123.4, 123.1, 45.3, 43.9, 38.9. ESI-MS: found 209.0 $[M+H]^+$.

5-(2-azidoethyl)-1,3-indandione (12): NaN_3 (2.3g, 36 mmol, 2 eq.) and **11** (3.7 g，18 mmol) were added to the 100 mL dried DMSO, and the reaction was mildly stirred for 20 min at 80 °C followed by adding 300 mL water. The reaction mixture was sequentially extracted with ether (3 × 400 mL), dried over anhydrous $MgSO_4$ and concentrated for purification with silica gel column chromatography, eluting with a 1:1 petroleum ether / CH_2Cl_2 (v/v) to give a brown solid **12** (680 mg，18%). ^{1}H NMR (300 MHz, $CDCl_3$): δ 7.94 (d, *J* = 7.8 Hz, 1H), 7.82 (s, 1H), 7.70 (d, *J* = 7.8 Hz, 1H), 3.62 (t, *J* = 6.6 Hz, 2H), 3.24 (s, 2H), 3.06 (t, *J* = 6.6 Hz, 2H). ^{13}C NMR (75 MHz, $CDCl_3$): δ 197.6, 197.1, 147.4, 144.1, 142.4, 136.7, 123.8, 123.4, 51.9, 45.6, 35.9. ESI-MS: found 216.1 $[M+H]^+$.

Synthesis of ethyl 6-azido-3-oxohexanoate (EAO)

The synthesis of 6-azido-3-oxohexanoate was adapted from the previous report[33].



Ethyl 6-chloro-3-oxohexanoate (15): LDA (320 mmol, 2 M, 2.5 eq.) was resolved in 155 mL THF and cooled to 0 °C, and **13** (16 g, 123 mmol) was added under nitrogen atmosphere and stirred for 1 h at 0 °C. Then the 1-bromo-2-chloroethane (17.4 g, 135

mmol, 1.1 eq.) was slowly added to the reaction solution at -78 °C. The solution was stirred and slowly heated up to 20 °C, and stirring continued for 2 h before quenching by adding 500 mL 5 M HCl. The mixture was sequentially extracted with ether (3 × 500 mL), washed with saturated NaCl solution, dried over anhydrous $MgSO_4$ and concentrated for purification with silica gel column chromatography, eluting with a 20:1 petroleum ether / ether (v/v) to give a light yellow liquid **15** (12.8 g，56%). ^{1}H NMR (300 MHz, $CDCl_3$): δ 4.20 (q, *J* = 7.2 Hz, 2H), 3.58 (t, *J* = 6.3 Hz, 2H), 3.45 (s, 2H), 2.76 (t, *J* = 6.6 Hz, 2H), 1.92-2.12 (m, 2H), 1.28 (t, *J* = 7.2 Hz, 3H). ESI-MS: found 193.1 $[M+H]^+$.

Ethyl 6-azido-3-oxohexanoate (16): The liquid **15** (6.8 g, 35 mmol) was resolved in 80 mL 3:1 acetone / H_2O. Then NaN_3 (3.5 g, 53 mmol, 1.5 eq.) was added, and the reaction was stirred and heated to reflux for 18 h. After reaction, the excess acetone was evaporated, and the mixture was sequentially extracted with ethyl acetate (3 × 50 mL), washed with saturated NaCl solution, dried over anhydrous $MgSO_4$ and concentrated for purification with silica gel column chromatography, eluting with a 20:1 petroleum ether / ether (v/v) to give a light yellow and transparent liquid **16** (4.1 g，58%). ^{1}H NMR (300 MHz, $CDCl_3$): δ 4.21 (q, *J* = 7.2 Hz, 2H), 3.46 (s, 2H), 3.34 (t, *J* = 6.6 Hz, 2H), 2.67 (t, *J* = 7.2 Hz, 2H), 1.85-1.94 (m, 2H), 1.29 (t, *J* = 7.2 Hz, 3H). ^{13}C NMR (75 MHz, $CDCl_3$): δ 201.9, 167.3, 61.7, 50.7, 49.6, 39.8, 22.9, 14.3. ESI-MS: $[M+H]^+$ found 200.1。

References

30 Shen, Q. *et al.* Synthesis of Quinolines via Friedlander Reaction in Water and under Catalyst-Free Conditions. *Synthesis* **44**, 389-392 (2012).

31 McInroy, G. R., Raiber, E. A. & Balasubramanian, S. Chemical biology of genomic DNA: minimizing PCR bias. *Chem. Comm.* **50**, 12047-12049 (2014).

32 Dai, Q. & He, C. A. Syntheses of 5-Formyl- and 5-Carboxyl-dC Containing DNA Oligos as Potential Oxidation Products of 5-Hydroxymethylcytosine in DNA. *Org. Let.* **13**, 3446-3449 (2011).

33 Lambert, P. H., Vaultier, M. & Carrie, R. Application of the Intramolecular Aza-Wittig Reaction to the Synthesis of Vinylogous Urethanes and Amides. *J. Org. Chem.* **50**, 5352-5356 (1985).