5-Formylcytosin ist vermutlich eine semi-permanente Base an definierten Genompositionen

Meng Su, Angie Kirchner, Samuele Stazzoni, Markus Müller, Mirko Wagner, Arne Schröder und Thomas Carell*[a]

*Center for Integrated Protein Science am Department Chemie der Ludwig-Maximilians-Universität München, Butenandtstrasse 5–13, 81377 München. [a] Korrespondenz an Thomas Carell, thomas.carell@lmu.de http://www.carellgroup.de

Veröffentlicht am 19.09.2016 in *Angewandte Chemie*, **128**(39), 11974 - 11978

Abstract: Die epigenetische Base 5-Formyl-2'-desoxycytosin (fdC) wurde kürzlich im Genom von Stammzellen entdeckt. Ihre Funktion ist bisher jedoch unbekannt. Daten aus Genomsequenzierungen zeigen eine Anreicherung in CpG Inseln von Promotoren, weshalb eine Beteiligung an der Regulation der Transkription während der zellulären Differenzierung angenommen wird. Auch ist fdC dafür bekannt von dem Enzym Thymin-DNA-Glykosylase (Tdg) erkannt und ausgeschnitten zu werden. Folglich wird angenommen, dass fdC als Intermediat während der aktiven Demethylierung auftritt. Um die Funktion der neuen epigenetischen Base fdC zu verstehen, ist es von großer Bedeutung dessen Bildung und Entfernung an definierten genomischen Positionen analysieren zu können. Im Folgenden beschreiben wir eine neue derartige Methode, die sequenzspezifische, chemische Derivatisierung von fdC mit *Droplet Digital* PCR kombiniert. Erste Ergebnisse zeigen, dass das Reparaturprotein Tdg nur 50% der fdCs an einer bestimmten Position im Genom entfernt, was auf semi-permanente Eigenschaften dieser Base hinweist.

Die DNA beinhaltet, neben der Sequenzinformation, eine weitere epigenetische Informationsebene, die steuert, wie aktiv das kontrollierte Gen transkribiert wird.[1] Neben den vier kanonischen Basen sind bisher vier weitere epigenetische Basen bekannt.[2] Diese sind 5-Methylcytosin (5mdC), 5-Hydroxymethylcytosin (5hmdC),[3] 5-Formylcytosin (5fdC)[4] und 5-Carboxycytosin (5cadC).[5] (Abbildung 1a) In den vergangenen Jahren trugen sensitive massenspektrometrische Methoden dazu bei die globalen Werte der epigenetischen Basen in Stammzellen,[4,6] sowie in Geweben, einschließlich des Gehirns, zu bestimmen.[7] Um mehr über den Gehalt und die Verteilung der epigenetischen Basen an definierten Genompositionen zu lernen, wurden verschiedene Sequenziermethoden entwickelt,[8] wobei selektive chemische Umsetzungen der Basen unternommen wurden;[9] manchmal in Kombination mit Bisulfit-sequenzierung.[9c,10] So konnten Informationen über die Verteilung der Basen zu einem bestimmten Zeitpunkt geliefert werden. Es ist jedoch ein Merkmal epigenetischer Information sich kontinuierlich zu ändern. Um einen tieferen Einblick in die Dynamik dieser epigenetischen Information an einer definierten genomischen Position zu erhalten, ist es essentiell Methoden zu entwickeln, die diese Veränderungen verfolgbar machen, beispielsweise um die zeitliche Änderung des fdCs an einer bestimmten Stelle im Genom detektieren zu können. [11] Eine ideale Methode würde letztendlich erlauben die fdC Dynamik an unterschiedlichen genomischen Positionen simultan zu verfolgen. Als zentrale Frage stellt sich hierbei: Sind die bereits gemessenen globalen Werte, Mittelwerte verschiedener Prozesse an unterschiedlicher Positionen oder reflektieren sie was an einer individuellen Stelle passiert? Um diese Frage beantworten zu können, entwickelten wir eine sequenzspezifische, chemische Derivatisierungs-Methode. In Kombination mit *Droplet Digital* PCR, ermöglicht unsere Methode eine direkte Verfolgung der epigenetischen Base fdC an spezifischen Positionen im Genom von Stammzellen.

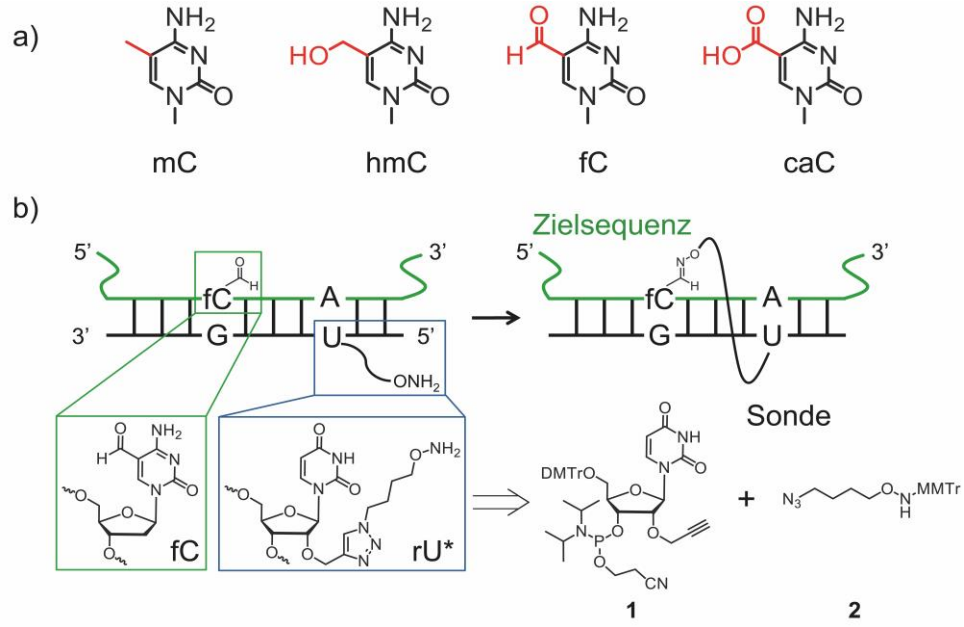


Abbildung 1. a) Strukturen der epigenetischen Cytosin-Modifikationen; b) Schematische Darstellung der fdC Detektionsstrategie und genutzte Bausteine, einschließlich der Click-Chemie basierten Assemblierung des rU* Sonden-Moleküls.

Zur sequenzspezifischen Lokalisierung von fdC im Genom benutzen wir ein kleines Sondenoligonukleotid (Abbildung 1b, Tabelle S1), das einen Hydroxylamin-Anker enthält, der eine kovalente Bindung mit fdC eingehen kann, sodass der Sondenstrang sofort fest mit dem Zielstrang verbunden ist. [4] Wir untersuchten systematisch verschiedene Linkerlängen, Ankerpunkte und Abstände. Die besten Ergebnisse erzielten wir mit einem Sondenoligonukleotid, in das wir ein 2'-*O*-propargyluridin (Phosphoramidit **1**) einfügten und das Azido-C4-hydroxylamin **2** mittels Cu(I)-Katalyse (Click-Reaktion) der Huisgen-Reaktion einbrachten.[12] Wir schützten den Hydroxylamin-Baustein für die Click-Reaktion mit einer Monomethoxytrityl-Gruppe (MMTr), die im Anschluss mit Essigsäure bei Raumtemperatur entfernt wurde. Diese kurze Säurebehandlung des Sondenoligonukleotids versursachte keine merkliche Depurinierung.

Nach Festphasensynthese, Click-Modifizierung und einem finalen Aufreinigungsschritt, erhielten wir Oligonukleotide mit unterschiedlichen Sequenzen und Längen, die eine rU-Hydroxylaminbase (rU*) an verschiedenen Positionen zur Reaktion mit der fdC-Base im Zielstrang, enthielten. Für nachfolgende Experimente wurden 13-nt lange Oligonukleotide angefertigt. Um festzustellen, an welcher Position der Linker am Besten mit dem fdC des Zielstranges reagiert, variierten wir die Position von rU* relativ zu fdC und untersuchten verschiedene Reaktionsbedingungen (Daten nicht gezeigt). Ausgezeichnete Ergebnisse wurden erreicht, als wir die Sonde **P1,** die rU*, exakt vier Basenpaare in 5´-Richtung relativ zu fdC enthielt, mit dem fdC Zielstrang in Gegenwart katalytischer Mengen an 4-Methoxyanilin (Abbildung 2a) hybridisierten. Mit diesem Katalysator, fand die *Crosslinking*-Reaktion bereits nach 24 h mit einer Ausbeute von über 95% statt. Ohne Katalysator konnte nur eine Ausbeute von 50% erhalten werden (Abbildung 2b).

Um die Reaktionsgeschwindigkeit zu steigern, wurden weitere Katalysatoren getestet. Die besten Ergebnisse lieferte hierbei 1,4-Diaminobenzol, wodurch die *Crosslinking*-Reaktion von **T1** und **P1** bereits nach 3 h erfolgte (Abbildung 2c). Die Ausbildung des Doppelstranges (**T1**:**P1**) wurde mittels denaturierendem PAGE Gel analysiert und über Fluoreszenz quantifiziert (Abbildung 2d).

Wird fdC eine Base weiter entfernt von rU* eingebaut, ohne die Sonde zu verändern, ist die Reaktion verlangsamt (Abbildung S2). Die Ergebnisse zeigen, dass rU*, vier oder fünf Basen in 5´-Richtung zu fdC entfernt, platziert, dem Anker erlaubt die Formylgruppe des fdC über die große Furche des Doppelstranges zu erreichen (Abbildung S3).

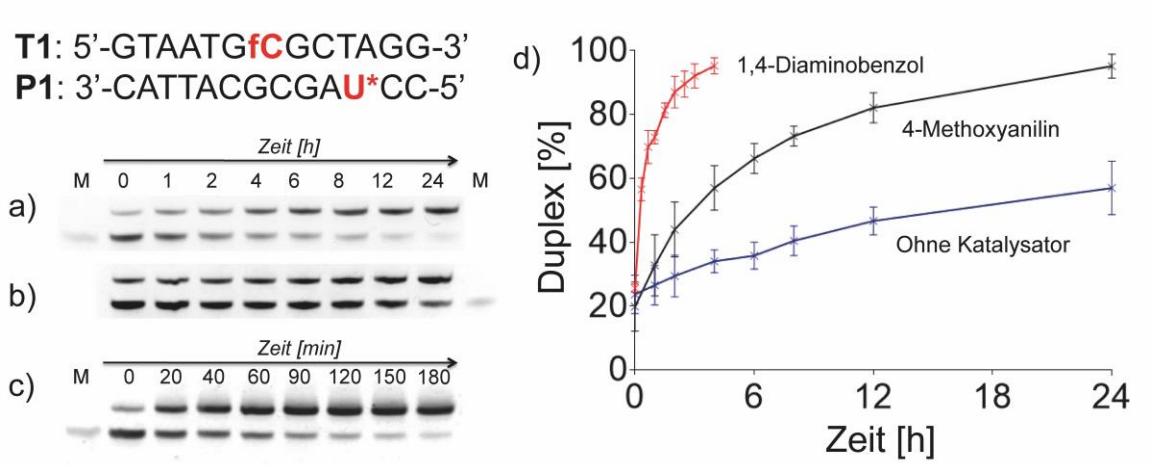


Abbildung 2. Denaturierendes PAGE Gel zeigt die Duplex-Bildung zwischen **T1** und **P1** bei 25 °C: a) Mit Katalysator 4-Methoxyanilin; b) Ohne Katalysator; c) Mit Katalysator 1,4-Diaminobenzol; d) Quantifizierung der DNA-Duplex Ausbildung während der Reaktion. Schwarz: Katalysator 4-Methoxyanilin, blau: ohne Katalysator, Rot: Katalysator 1,4-Diaminobenzol. Die Fehlerbalken repräsentieren die mittlere Standardabweichung, berechnet aus dem Mittelwert dreier Replikate. Bedingungen: 2 µM Oligonukleotide, 100 mM NaCl, 10 mM NaOAc Puffer pH 6.0, 10 mM 4-Methoxyanilin. M = Einzelstrangmarker. Der Zeitpunkt 0 befindet sich nach dem *Re-annealing*.

MALDI-TOF Daten bestätigten, dass sich der *Crosslink* wie erwartet bildete (Abbildung S4). Im Falle des reagierten Doppelstranges **T1:P1** erhielten wir die korrekte Molekülmasse mit $m/z_{gefunden}$ 8081,9 ($m/z_{berechnet}$ 8084,7). Wie erwartet, führte die Oxim-Bildungsreaktion zwischen **T1** und **P1** zu einer erhöhten Schmelztemperatur des hybridisierten und reagierten Doppelstranges (Abbildung S5). Der nicht durch *Crosslinking* verknüpfte 13-mer Doppelstrang schmolz typischerweise bei etwa 44 °C. Nach Ausbildung des *Crosslinks* zeigte der Doppelstrang dagegen eine Schmelztemperatur von über 80 °C.

Da Pyrimidine mit Nukleophilen an der C6 Position in einer Michael-artigen Reaktion reagieren können, die als Basis der Bisulfit-Sequenzierung dient, testeten wir als nächstes ob die Reaktion mit rU* auch mit anderen Pyrimidinen möglich ist (Abbildung 3). Zu unserer Zufriedenheit ergab eine Hybridisierung der rU*-Sonde mit einem dT bzw. dC, mdC, hmdC und cadC (**T3-8**) enthaltenden Zielstrang keine Reaktion. Eine Reaktion wurde jedoch mit einer abasischen Stelle festgestellt. Dies ist von Bedeutung, da fdC und cadC Substrate der Basenexzisionsreparatur sind und daher im Prinzip Vorläufer von abasischen Stellen sein könnten.[13] In diesem Sinne zeigt rU* immer die Präsenz von fdC und potentiell auch von abasische Stellen aus fdC und cadC an. Anschließend wandelten wir die Möglichkeit fdC zu detektieren in eine Methode um, einzelne fdC Basen an einer bestimmten Position in ganzen Genomen zu detektieren. Zu diesem Zweck verknüpften wir Chemie mit *Droplet Digital* PCR[14] basierender Amplifikation und Datenanalyse.

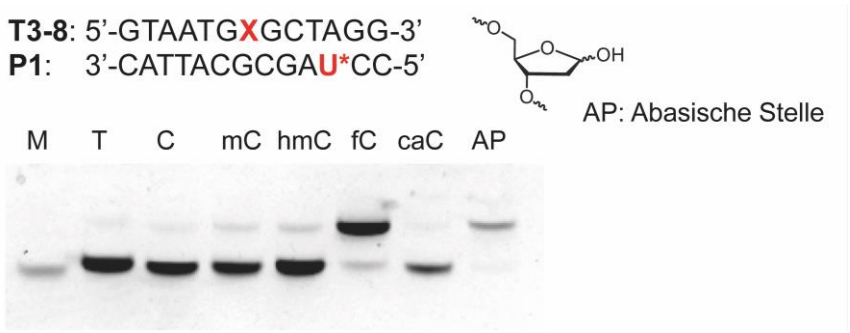


Abbildung 3. Denaturierendes PAGE Gel zeigt die Duplex-Ausbildung der Stränge **T3**-**8** mit **P1** bei 25 °C nach 24 h.

Zunächst wurde genomische DNA (**Tg**) aus murinen embryonalen Stammzellen (mESCs) zu verschiedenen Zeitpunkten während des sogenannten *Primings* aus naïven Stammzellen isoliert. Außerdem isolierten wir genomische DNA aus mESCs mit und ohne dem Reparaturenzym Tdg ($Tdg^{-/-}$), um so die Entfernung und Reparatur von fdC und cadC zu verhindern. Weiterhin isolierten wir genomische DNA aus Stammzellen ohne die drei DNA-Methyltransferasen (Dnmt1, 3a und 3b). In diesen Zellen entsteht kein mdC, weshalb sie nicht fähig sind auch die weiter oxidierten epigenetischen Basen xdC (x= hm, f und ca) herzustellen. Genomische DNA hieraus fungiert in unseren Experimenten als Negativkontrolle. Analysiert wurden zwei verschiedene bereits bekannte fdC Positionen, an denen ein unterschiedlich hoher fdC Gehalt beschrieben wurde.[10c]
Zunächst fokussierten wir uns auf die 30.020.539^{ste} Position des Chromosoms 16 *Mus musculus* (MM9) im Exon 3 von 632428C04Rik. RedBS-Sequenzierungen zufolge, enthält diese Stelle zu 23% fdC. Die zweite Stelle ist die 8.846.677^{ste} Position des Chromosoms 15 und befindet sich in nicht-kodierender DNA. Der gleichen Studie zufolge, enthält diese Position zu 32% fdC.

Zur Untersuchung der ersten Position, ließen wir eine rU* enthaltende 25-mer Sonde (**P2**, **SI**) mit **Tg** unter 1,4-Diaminobenzol Katalyse reagieren. In Abwesenheit von fdC bildet sich keine kovalente Bindung zwischen **P2** und **Tg** aus. Um den Überschuss an Sonde zu entfernen, luden wir den **Tg:P2** Komplex auf eine NEB Monarch DNA *Cleanup* Säule und spülten mit Waschpuffer, um Oligonukleotide kürzer als 50-nt zu eluieren, was ungebundenem **P2** entspricht. Nach dem Waschen, eluierten wir **Tg:P2** mit TE Puffer. UV/Vis Analyse des Eluats zeigte ein typisches gDNA Spektrum. Als nächstes gaben wir einen 70-mer 5'-phosphorylierten Reporterstrang (**R1**, **SI**) zu, der mit einem 18-nt langen Abschnitt in direkter Nachbarschaft zum Sondenstrang hybridisiert und ligierten diese Sonde mit dem Reporter bei 60 °C mit Ampligase, um den **R1-P2** Duplex herzustellen, wie in Abbildung 4a schematisch dargestellt. Als nächstes gaben wir zwei Primerpaare zum Assay (Blaue und grüne Pfeile, Abbildung 4a) zu, um das Ligationsprodukt relativ zum Zielduplex zu vervielfältigen. Von Bedeutung ist hierbei, dass die blauen Primer lediglich das **R1-P2** Hybrid der Sonde aus dem Ligationsschritt binden, während die grünen Primer die Gegenwart von gDNA signalisieren. Die Amplifikation wurde mit zwei verschiedenen TaqMan Sonden verfolgt, die eine Fluoreszenz bei 520 und 556 nm aufweisen. Diese relative Detektion ist notwendig, um auf die Menge an zugegebener gDNA normalisieren zu können. Da die konventionelle *Real-Time* PCR bei einer Differenz geringer als dem zehnfachen Unterschied ungenaue Ergebnisse liefert, nutzten wir für unsere Studien die Droplet Digital PCR. Bei dieser Methode werden kleine Tropfen gebildet, wobei im Optimalfall pro Tropfen maximal ein komplettes Analysekontrukt enthalten ist, wie in Abbildung 4a gezeigt. Die PCR Reaktionen finden in den Tropfen statt und generieren ein spezifisches Signal. Darauffolgendes Zählen der Farben jedes einzelnen Tropfens führt zu Werten, aus denen die exakte Menge an fdC, auch bei geringen Werten, berechnet werden kann. Ein repräsentativer Graph ist in Abbildung 4b gezeigt. Leere Tropfen ergeben kein PCR Signal (schwarze Punkte in Abbildung 4b). Tropfen, die nur **Tg** enthalten, ergeben das PCR Signal der grünen Primer (grüne Punkte in Abbildung 4b). Blaue Signale werden erhalten, wenn das ligierte Produkt **R1-P2** während des Ligationsprozesses (60 °C für 10 h) von **Tg** dissoziiert. Rote Signale werden schließlich aus Tropfen generiert, die beide PCR Produkte enthalten. Für die Berechnung der Werte wird auf die Hintergrundinformation verwiesen.

Mit dieser Methode untersuchten wir zunächst Stammzellen ohne Tdg Enzym ($Tdg^{-/-}$). Ein eher hoher Wert von 28.5% fdC konnte an der ersten Position gemessen werden (Abbildung 4c), was mit den Ergebnissen aus redBS-Sequenzierungen gut übereinstimmt.[10c] Als wir jedoch das Experiment in Stammzellen mit intaktem Tdg Reparaturenzym durchführten ($Tdg^{+/-}$), wurden an dieser Position nur noch 15.7% gemessen (Abbildung 4c). Dies ist von großer Bedeutung, da es zeigt, dass Tdg nur die Hälfte der fdCs an einer bestimmten Stelle entfernt und ungewöhnlich da Reparaturenzyme dafür bekannt sind quasi alle möglichen Substrate zu finden. Dieses Ergebnis unterstreicht die hohe Dynamik der fdC Bildung und Reparatur an einer bestimmten Position im Genom. Als wir den fdC Gehalt an dieser Stelle in Stammzellen ohne DNA-Methyltransferase (Dnmt TKO) untersuchten, fiel der fdC Wert wie erwartet auf etwa 5%. Dies zeigt, dass die gemessenen Werte in den $Tdg^{+/-}$ Zellen real und keine Artefakte sind. Um weiterhin aufzuklären, ob die fdC Werte einzelner Positionen (Abbildung 4c) den globalen genomischen fdC-Werten folgen, quantifizierten wir zudem die absoluten Werte an fdC in diesen Zellen (Abbildung 4d). Diese globalen Daten stehen in guter Übereinstimmung mit den Daten der fdC Quantifizierungen einzelner Positionen. Aufgrund unserer neuen Daten ist es eher unwahrscheinlich, dass fdC an einer Stelle komplett entfernt und an einer anderen gänzlich vor der Reparatur geschützt wird. Stattdessen wird fdC an einer Position in einer Zellpopulation eher teilweise entfernt. Alternativ, könnte Tdg fdC unterschiedlich auf den beiden Chromosomen entfernen, was jedoch weiterer Untersuchung bedarf.

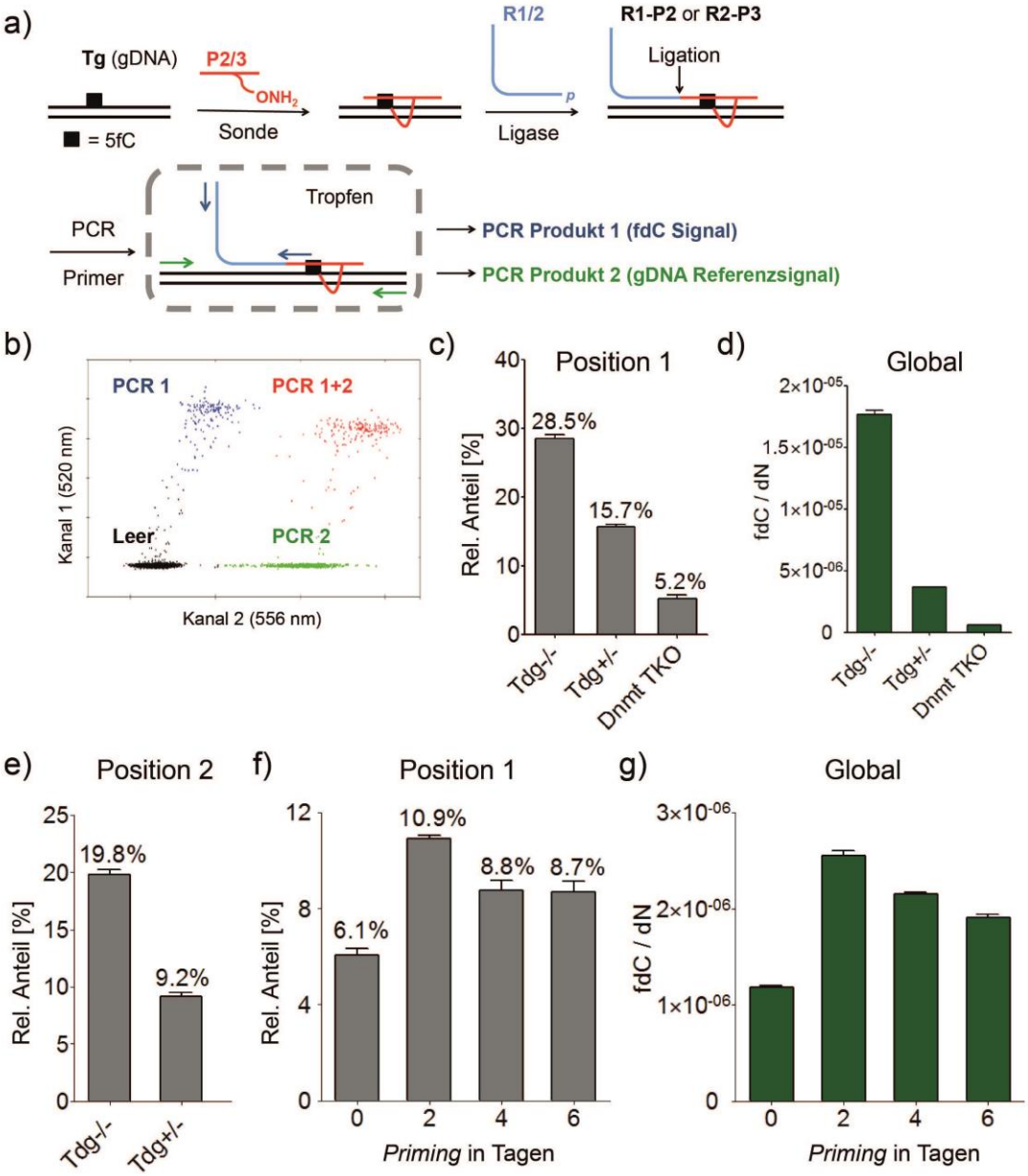


Abbildung 4. Die fdC Sequenzierungsmethode: a) Schematische Repräsentation der Methode, schwarze Linie: gDNA; rote Segmente: fdC Sonde; blaue Segmente: Reporterstränge; Pfeile: PCR Primerpaare; b) Typischer 2-D Plot der Tropfen-Fluoreszenz; c) Cluster-Verhältnisse der Position 1 in $Tdg^{-/-}$, $Tdg^{+/-}$ und Dnmt TKO mES Zellen; d) Globale fdC Quantifizierung in $Tdg^{-/-}$, $Tdg^{+/-}$ und Dnmt TKO mES Zellen mittels unserer LC-MS Methode; e) Cluster-Verhältnisse der Position 2 in $Tdg^{-/-}$ und $Tdg^{+/-}$ mES Zellen; f) Cluster-Verhältnisse der Position 1 in Wildtyp mES Zellen an verschiedenen Tagen während des *Primings*; g) Globale Quantifizierungsdaten der Wildtyp mES Zellen gemessen durch LC-MS.

Um unsere Daten verifizieren zu können, wiederholten wir die Tdg-Studie an einer zweiten Position (8,846,677^{ste} Nukleosid des Chromosoms 15). Für diese Stelle, entwarfen wir einen neuen Sondenstrang **P3**, einen neuen Reporterstrang **R2** und führten die *dd*PCR erneut mit zwei Primerpaaren durch (Abbildung S5). Vergleicht man die Daten aus $Tdg^{-/-}$ Zellen mit den Daten aus $Tdg^{+/-}$ Zellen, ist an dieser Position wieder ein 50%iger Rückgang des fdC Gehalts erkennbar. Dies ist in voller Übereinstimmung mit den Daten der ersten Position (Abbildung 4e).

Schließlich führten wir eine Untersuchung durch, bei der wir die Entwicklung von fdC an der ersten Position während des *Primings* von Stammzellen aus naïven Zellen beobachteten (Abbildung 4f). An dieser Position ist ein Anstieg der fdC Werte, besonders in der frühen *Priming*-Phase, zu beobachten. Es folgen ein geringer Abfall und schließlich stabile Werte (Abbildung 4f), was wiederrum in Übereinstimmung mit den globalen Messwerten steht, die wir durch unsere bereits beschriebene Methode bestimmten (Abbildung 4g).

Die Tatsache, dass unsere Methode in der Lage ist, die beschriebenen globalen Werte an einer einzelnen genomischen Position zu reproduzieren, stimmt uns positiv, dass unsere Methode erlaubt, verlässlich wiederzugeben was an einer definierten Stelle geschieht. Da die Daten der individuellen Positionen und die globalen Werte korrelieren, haben wir erste Hinweise, dass die beschriebenen globalen Trends tatsächlich wiedergeben, was an einer einzigen fdC enthaltenden Position geschieht, anstatt den Mittelwert unterschiedlicher Dynamiken an verschiedenen Positionen abzubilden. Ein interessantes Ergebnis dieser Studie ist hierbei, dass das Reparaturenzym Tdg nur die Hälfte der fdC Basen an einer bestimmten Stelle im Genom einer mESC Population entfernt, was auf semi-permanente Eigenschaften der Base an definierten genomischen Positionen schließen lässt.

Experimenteller Teil

***Crosslink* der Sonde** gDNA Lösung (1.2 µg), fdC Sonde (1 µM, 2 µL), NaH_2PO_4-Na_2HPO_4 Puffer (200 mM, pH = 6.0, 2 µL), NaCl aq. (1.5 M, 2 µL) und *dd*H_2O wurden zu einem Endvolumen von 18 µL zusammen gegeben. Die Mischung wurde 3 min bei 95 °C erhitzt und schnell auf 25 °C abgekühlt. 1,4-Benzoldiamin aq. (10 mM, 2 µL) wurde zugegeben und das Reaktionsgefäß 6 h bei 25 °C geschüttelt. Vor Aufreinigung mit dem NEB Monarch PCR DNA *Cleanup* Kit, wurde die Reaktionsmischung mit Na_2HPO_4 aq. (200 mM, 40 µL) neutralisiert.
Ligation Die zuvor beschriebene gDNA Lösung (300 ng), Reporterstrang (20 nM, 1 µL), Ampligase Reaktionspuffer (10×, 2 µL), Ampligase von Epicentre (5 U/µL, 2 µL, 10 U) und *dd*H_2O wurden bis zu einem Endvolumen von 20 µL zusammen gegeben. Die Mischung wurde für 3 min bei 95 °C erhitzt, dann 1 min bei 94 °C, 1 h für 60 °C und wieder auf 94 °C (10 Zyklen) temperiert. Danach wurde das Reaktionsgemisch mit Tris-HCl Puffer (200 mM, pH = 7.6, 50 µL) verdünnt, bevor es mittels NEB Monarch PCR DNA *Cleanup* Kit gereinigt wurde.
***Droplet Digital* PCR** Die *dd*PCR wurde auf einem Bio-Rad QX100 *dd*PCR System durchgeführt. Für die Reaktion wurden gDNA (6 ng), vier Primer (jeweils 18 µM, 1 µL), zwei TaqMan Sonden (jeweils 5 µM, 1 µL), *digital* PCR Supermix für Sonden (kein dUTP, 2×, 10 µL) und *dd*H_2O gemischt; das Endvolumen betrug 20 µL. PCR-Zyklus: 95 °C für 10 min, 94 °C für 30 sec und 64 °C für 1 min (35 Zyklen), dann folgten 98 °C für 10 min und schließlich wurde mit 2 °C/s auf 12 °C abgekühlt. Eine detaillierte Beschreibung ist den Hintergrundinformationen zu entnehmen.

Danksagung

Wir danken K. Hufnagel für die Herstellung der Phosphoramidite der verwendeten epigenetischen Basen. Dieses Projekt wurde durch die Deutsche Forschungsgemeinschaft SFB1032 und den Excellence Cluster CiPS^M (EXC114) gefördert. Weiterhin wurde das Projekt durch den *European Union's Horizon 2020 Research* und das *Innovation Program Under Grant Agreement* Nr. 642023 (ITN clickgene) unterstützt.

Stichworte: Epigenetische Base • Click-Chemie • 5-Formylcytosine • Genomische DNA • *Droplet Digital* PCR

[1] P. A. Jones, Nat. Rev. Genet. 2012, 13, 484-492.
[2] T. Carell, C. Brandmayr, A. Hienzsch, M. Müller, D. Pearson, V. Reiter, I. Thoma, P. Thumbs, M. Wagner, Angew. Chem., Int. Ed. 2012, 51, 7110-7131; Angew. Chem. 2012, 124, 7220-7242.
[3] a) S. Kriaucionis, N. Heintz, Science 2009, 324, 929-930; b) M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, A. Rao, Science 2009, 324, 930-935.
[4] T. Pfaffeneder, B. Hackner, M. Truss, M. Münzel, M. Müller, C. Deiml, C. Hagemeier, T. Carell, Angew. Chem., Int. Ed. 2011, 50, 7008-7012; Angew. Chem. 2011, 123, 7146-7150.
[5] a) S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, Science 2011, 333, 1300-1303; b) Y.-F. He, B.-Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, Y. Sun, X. Li, Q. Dai, C.-X. Song, K. Zhang, C. He, G.-L. Xu, Science 2011, 333, 1303-1307.
[6] S. Schiesser, B. Hackner, T. Pfaffeneder, M. Müller, C. Hagemeier, M. Truss, T. Carell, Angew. Chem., Int. Ed. 2012, 51, 6516-6520; Angew. Chem. 2012, 124, 6622-6626.
[7] a) M. Münzel, D. Globisch, T. Brückl, M. Wagner, V. Welzmiller, S. Michalakis, M. Müller, M. Biel, T. Carell, Angew. Chem., Int. Ed. 2010, 49, 5375-5377; Angew. Chem. 2010, 122, 5503-5505; b) D. Globisch, M. Münzel, M. Müller, S. Michalakis, M. Wagner, S. Koch, T. Brückl, M. Biel, T. Carell, PLoS ONE 2010, 5, e15367; c) M. Wagner, J. Steinbacher, T. F. J. Kraus, S. Michalakis, B. Hackner, T. Pfaffeneder, A. Perera, M. Müller, A. Giese, H. A. Kretzschmar, T. Carell, Angew. Chem., Int. Ed. 2015, 54, 12511-12514; Angew. Chem. 2015, 127, 12691-12695.
[8] a) N. Plongthongkum, D. H. Diep, K. Zhang, Nat. Rev. Genet. 2014, 15, 647-661; b) M. J. Booth, E.-A. Raiber, S. Balasubramanian, Chem. Rev. 2015, 115, 2240-2254.
[9] a) W. A. Pastor, U. J. Pape, Y. Huang, H. R. Henderson, R. Lister, M. Ko, E. M. McLoughlin, Y. Brudno, S. Mahapatra, P. Kapranov, M. Tahiliani, G. Q. Daley, X. S. Liu, J. R. Ecker, P. M. Milos, S. Agarwal, A. Rao, Nature 2011, 473, 394-397; b) E.-A. Raiber, D. Beraldi, G. Ficz, H. Burgess, M. Branco, P. Murat, D. Oxley, M. Booth, W. Reik, S. Balasubramanian, Genome Biol. 2012, 13, R69; c) C.-X. Song, Keith E. Szulwach, Q. Dai, Y. Fu, S.-Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, J. Gao, P. Liu, L. Li, G.-L. Xu, P. Jin, C. He, Cell 2013, 153, 678-691; d) B. Xia, D. Han, X. Lu, Z. Sun, A. Zhou, Q. Yin, H. Zeng, M. Liu, X. Jiang, W. Xie, C. He, C. Yi, Nat. Methods 2015, 12, 1047-1050.
[10] a) M. Yu, Gary C. Hon, Keith E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren, C. He, Cell 2012, 149, 1368-1380; b) M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, S. Balasubramanian, Science 2012, 336, 934-937; c) M. J. Booth, G. Marsico, M. Bachman, D. Beraldi, S. Balasubramanian, Nat. Chem. 2014, 6, 435-440; d) X. Lu, C.-X. Song, K. Szulwach, Z. Wang, P. Weidenbacher, P. Jin, C. He, J. Am. Chem. Soc. 2013, 135, 9315-9317.
[11] a) A. Nomura, K. Sugizaki, H. Yanagisawa, A. Okamoto, Chem. Commun. 2011, 47, 8277-8279; b) J. Duprey, G. A. Bullen, Z.-Y. Zhao, D. M. Bassani, A. F. A. Peacock, J. Wilkie, J. H. R. Tucker, ACS Chem. Bio. 2016, 11, 717-721.
[12] a) P. M. E. Gramlich, S. Warncke, J. Gierlich, T. Carell, Angew. Chem., Int. Ed. 2008, 47, 3442-3444; Angew. Chem. 2008, 120, 3491-3493.; b) J. Willibald, J. Harder, K. Sparrer, K.-K. Conzelmann, T. Carell, J. Am. Chem. Soc. 2012, 134, 12330-12333.
[13] A. Maiti, A. C. Drohat, J. Biol. Chem. 2011, 286, 35334-35338.
[14] B. J. Hindson, K. D. Ness, D. A. Masquelier, P. Belgrader, N. J. Heredia, A. J. Makarewicz, I. J. Bright, M. Y. Lucero, A. L. Hiddessen, T. C. Legler, T. K. Kitano, M. R. Hodel, J. F. Petersen, P. W. Wyatt, E. R. Steenblock, P. H. Shah, L. J. Bousse, C. B. Troup, J. C. Mellen, D. K. Wittmann, N. G. Erndt, T. H. Cauley, R. T. Koehler, A. P. So, S. Dube, K. A. Rose, L. Montesclaros, S. Wang, D. P. Stumbo, S. P. Hodges, S. Romine, F. P. Milanovich, H. E. White, J. F. Regan, G. A. Karlin-Neumann, C. M. Hindson, S. Saxonov, B. W. Colston, Anal. Chem. 2011, 83, 8604-8610.
[15] M. Wendeler, L. Grinberg, X. Wang, P. E. Dawson, M. Baca, Bioconjugate Chem. 2014, 25, 93-101.

Table of contents

1. General methods of organic synthesis
2. Synthesis of the hydroxylamine linker
3. ^{1}H-NMR spectra of the linker
4. General methods for oligonucleotide synthesis
5. Crosslinking studies with the synthesized strands
6. Experimental details of the genomic fdC profiling study
7. Quantification modeling
8. Droplet digital PCR data

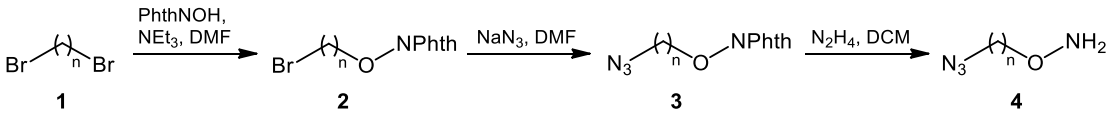
1. General methods of organic synthesis

Chemicals were purchased from *Sigma-Aldrich* and used without further purification. The solvents for organic synthesis were of reagent grade and purified by distillation. Solutions were concentrated *in vacuo* on a *Heidolph* rotary evaporator with a *Vario PC2001* diaphragm pump by *Vacuubrand*. All mixed solvent systems are reported as v/v solutions. All reactions were monitored by thin-layer chromatography (TLC), performed on *Merck* 60 (silica gel F_{254}) plates. Chromatographic purification of products was accomplished using flash column chromatography on silica gel (230-400 mesh) purchased from Merck.

^{1}H- and ^{13}C-NMR spectra were recorded in deuterated solvents on *Bruker ARX* 400 spectrometers and calibrated to the residual solvent peak. Chemical shifts (δ, ppm) are quoted relative to the residual solvent peak as internal standard and coupling constants (*J*) are corrected and quoted to the nearest 0.1 Hz. Multiplicities are abbreviated as follows: s = singlet, d = doublet, t = triplet, m = multiplet.

2. Synthesis of the hydroxylamine linker

O-(4-Azidobutyl)hydroxylamine 4



1,4-Dibromobutane **1** (5.9 mL, 49.4 mmol, 2.0 eq.) was added to a solution of *N*-hydroxyphthalimide (PhthNOH, 4.0 g, 24.5 mmol, 1.0 eq.) and triethylamine (7.5 mL, 53.6 mmol, 2.2 eq.) in anhydrous dimethylformamide. The mixture was stirred at room temperature for 24 h. The reaction was diluted with water, and the aqueous phase was extracted three times with ethyl acetate. The combined organic phases were dried over $MgSO_4$, filtered and concentrated to give the crude product **2** (5.02 g, 16.9 mmol, 0.69 eq.) as a white solid. The residue was dissolved in anhydrous dimethylformamide and sodium azide (1.32 g, 20.6 mmol, 0.85 eq.) was added. The mixture was stirred at room temperature for 2 h, diluted with water and extracted with ethyl acetate three times. The combined organic phases were dried over $MgSO_4$, filtered and concentrated. The crude was purified by flash chromatography on silica gel (*iso*-hexane/ethyl acetate 10:1→ 2:1) to give **3** (3.97 g, 15.2 mmol, 0.62 eq.) as a yellow oil. The oil was redissolved in hydrazine monohydrate (1.1 mL, 22.8 mmol, 0.93 eq.) and dichloromethane (10 mL). The mixture was stirred at room temperature for 24 h and then filtered. The solution was diluted with dichloromethane and washed with NaCl aq. three times. The combined organic phases were dried over $MgSO_4$,

filtered and concentrated to give **4** (1.81 g, 13.9 mmol, 57% yield over three steps) as a colorless oil.

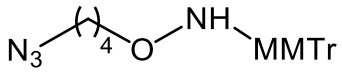
$\boldsymbol{R}_f$ = 0.42 (DCM/ MeOH 10:1).

^{1}H-NMR (400 MHz, $CDCl_3$): δ = 3.70 (t, $^3J_{H,H}$ = 5.6 Hz, 2H, O-C**H$_2$**-CH_2), 3.38 (t, $^3J_{H,H}$ = 6.4 Hz, 2H, C**H$_2$**-N_3), 1.68−1.64 (m, 4H, CH_2-C**H$_2$**-C**H$_2$**-CH_2).

^{13}C-NMR (100 MHz, $CDCl_3$): δ = 75.1 (-O-**C**H_2-CH_2), 51.2 (**C**H_2-N_3), 25.6 (**C**H_2), 25.5 (**C**H_2).

HRMS (ESI+): calculated for $C_4H_{11}ON_4^+$ $[M+H]^+$: 131.0927, found: 131.0928.

***O*-(4-Azidobutyl)-*N*-[(4-methoxyphenyl)diphenylmethyl]hydroxylamine** (**5**)



5

O-(4-Azidobutyl)hydroxylamine **4** (1.88 g, 14.4 mmol, 1.0 eq.) was dissolved in anhydrous dichloromethane (40 mL). 4-Monomethoxytritylchloride (MMTr-Cl, 4.91 g, 15.9 mmol, 1.1 eq.) and diisopropylethylamine (5.0 mL, 28.9 mmol, 2.0 eq.) was added to the mixture at 0°C. The reaction was stirred at room temperature for 2 h, diluted with dichloromethane, washed with saturated $NaHCO_3$, dried over $MgSO_4$, filtered and concentrated. The crude was purified by flash chromatography on silica gel (*iso*-hexanes/ethyl acetate 15:1 + 3% triethylamine) to give **5** (4.89 4g, 12.2 mmol, 84%) as a yellowish oil.

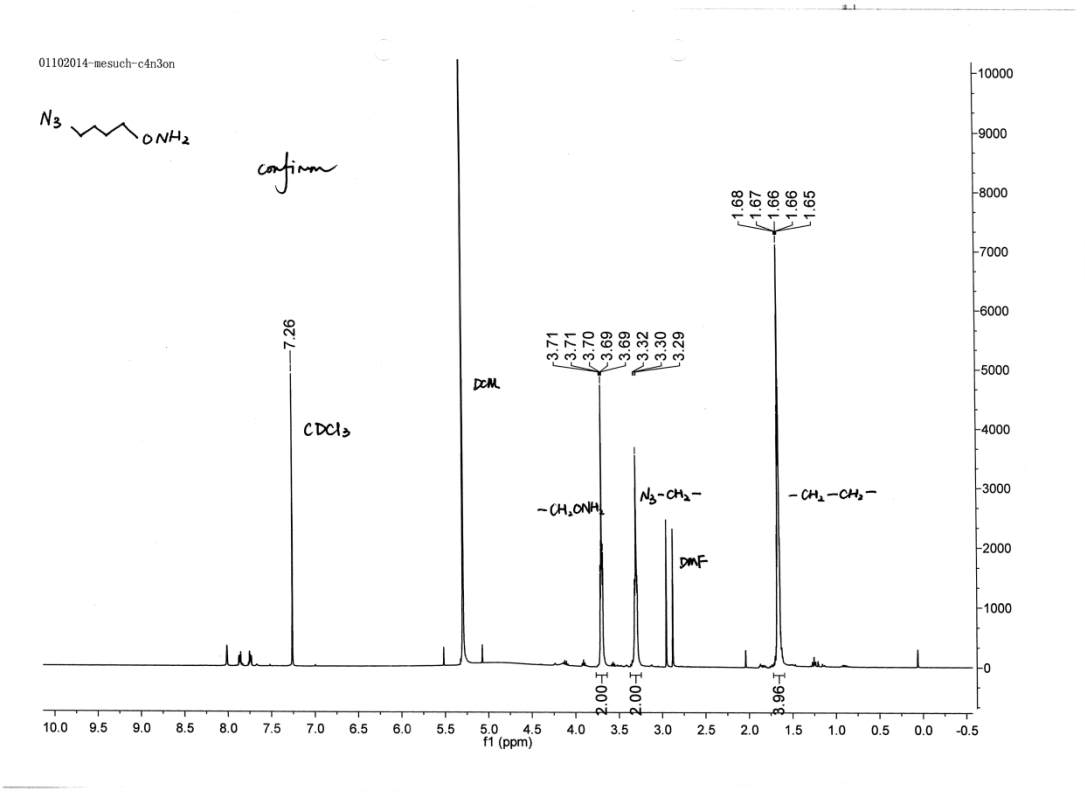
$\boldsymbol{R}_f$ = 0.64 (*iso*-hexane/ ethyl acetate 10:1 + 3% triethylamine).

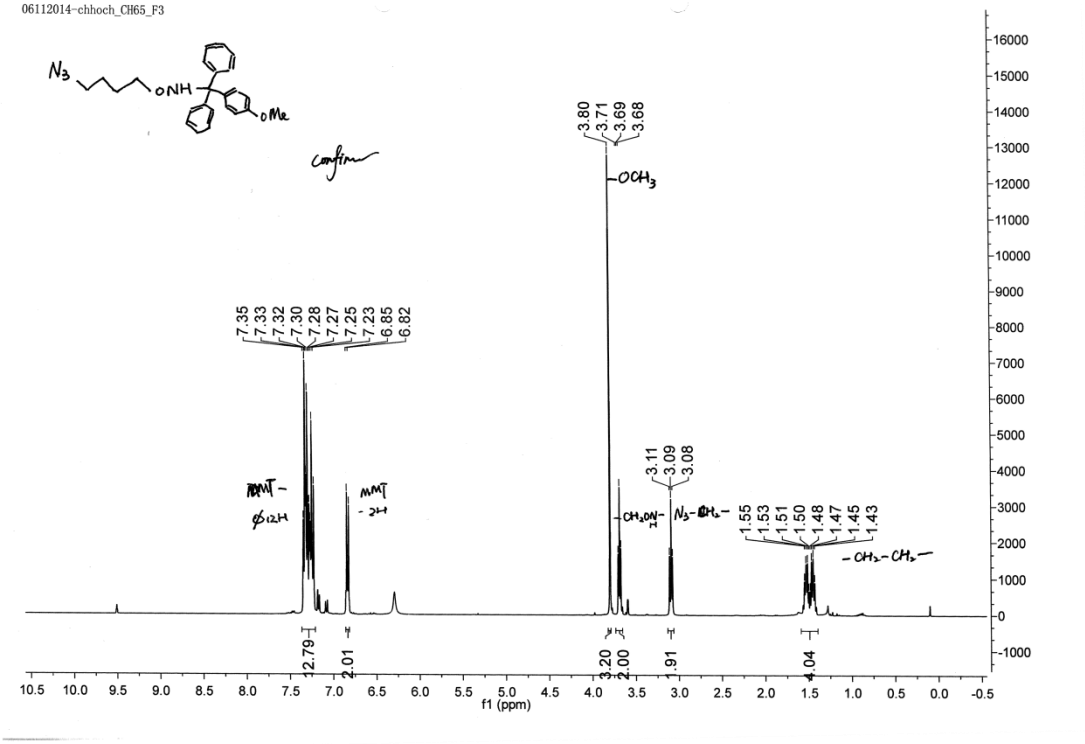
^{1}H-NMR (400 MHz, $CDCl_3$): δ = 7.31−7.18 (m, 12H, 12 × C_{Ar}**H**), 6.79 (d, $^3J_{H,H}$ = 8.8 Hz, 2H, 2 × CH_3-O-C-C**H**), 3.76 (s, 3H, O-C**H$_3$**), 3.65 (t, $^3J_{H,H}$ = 6.0 Hz, 2H, O-C**H$_2$**-CH_2), 3.05 (t, $^3J_{H,H}$ = 6.8 Hz, 2H, N_3-C**H$_2$**), 1.52−1.37 (m, 4H, N_3-CH_2-C**H$_2$**-C**H$_2$**-CH_2).

^{13}C-NMR (100 MHz, $CDCl_3$): δ = 158.3 (CH_3-O-**C**), 144.6 (2 × **C**$_{Ar}$), 136.6 (**C**$_{Ar}$), 130.2 (2 × **C**$_{Ar}$), 129.0 (4 × **C**$_{Ar}$), 127.6 (4 × **C**$_{Ar}$), 126.7 (2 × **C**$_{Ar}$), 112.9 (2 × **C**$_{Ar}$), 77.2 (O-NH-**C**), 73.2 (O-**C**H_2), 55.2 (O-**C**H_3), 51.1 (N_3-**C**H_2), 25.8 (N_3-CH_2-**C**H_2), 25.5 (O-CH_2-**C**H_2).

HRMS (ESI-): calculated for $C_{25}H_{27}O_4N_4^-$ $[M+HCO_2]^-$: 447.2038, found: 447.2040.

3. ^{1}H-NMR spectra of the linker





4. General methods for oligonucleotide synthesis

DNA Oligonucleotide synthesis was performed on an Applied Biosystems Incorporated 394 automated synthesizer. Phosphoramidites and solid supports columns were purchased from *Glen Research*, *Link Technology*, and *ChemGene Corporation*. Oligodeoxynucleotides were synthesized in a 1 μmol scale with standard DNA synthesis cycles (trityl off mode). Coupling time for modified nucleosides was extended to 10 min. The oligonucleotide was cleaved using conc. ammonium hydroxide aq. at 25 °C for 17 h. The aqueous solution was then collected and evaporated in a *SpeedVac* concentrator, and the pellet was redissolved in ddH_2O.

Analytical RP-HPLC was performed using a *Macherey-Nagel* Nucleodur 100-3 C18ec column on 2695 Separation Module equipped with a Waters Alliance 2996 Photodiode Array Detector using a flow of 0.5 mL/min. Semi-preparative RP-HPLC was performed using a *Macherey-Nagel* C18 column (5 mm, 9.4 × 250 mm) on *Waters* Breeze 2487 Dual λ Array Detector, 1525 Binary HPLC Pump. Conditions: Buffer A, 0.1 M TEAA (triethylammonium acetate) in water; buffer B, 0.1 M TEAA in 80% acetonitrile.

The purified fractions were concentrated and characterized by Matrix Assisted Laser Desorption Ionization Time of Flight (MALDI-TOF) on a Bruker Daltonics Autoflex II instrument. The concentration of the oligonucleotide solutions was calculated from the UV absorbance at 260 nm on a Nanodrop ND-1000 spectrophotometer. Extinction coefficients of the oligonucleotides at 260 nm were calculated by addition of the extinction coefficients of the individual nucleobases: dA 15.0 L/(mmol·cm), dC 7.1 L/(mmol·cm), dG 12.0 L/(mmol·cm), dT 8.4 L/(mmol·cm), mdC 7.8 L/(mmol·cm), hmdC 8.7 L/(mmol·cm), fdC 11.3 L/(mmol·cm), cadC 7.1 L/(mmol·cm). The 1,2,3-triazole and abasic monomer are transparent at 260 nm.

Click reaction on the solid support with azide linkers and deprotection

After solid phase synthesis (0.2 μmol scale, approx. 50% yield for 13 mer, calculated as 0.1 μmol), the solid support was suspended in dimethyl sulfoxide (80 μL) and acetonitrile (25 μL). To the mixture, $CuSO_4$ aq. (100 mM, 50 μL, 5.0 μmol, 50 eq.), sodium ascorbate aq. (500 mM, 20 μL, 10 μmol, 100 eq.), *N*,*N*-diisopropylethylamine solution in acetonitrile (200 mM, 75 μL, 15 μmol, 150 eq.), solution of **5** in dimethyl sulfoxide (100 mM, 50 μL, 5.0 μmol, 50 eq.) were added. The reaction was conducted at 25 °C for 24 h. Afterwards, the solid support was washed with dimethyl sulfoxide, dilute $NaHCO_3$ aq., acetonitrile, ether and air-dried to a powder. The solid phase was then cleaved with conc. aqueous NH_3 at 25 °C for 17 h, purified by HPLC and concentrated. The removal of the MMTr protection group on the hydroxylamine was carried out by dissolving the obtained oligonucleotide in acetic acid aq. (20%, 200 μL) at 25 °C for 30 min, precipitated by addition of sodium acetate solution (3 M, 60 μL) and EtOH (1040 μL), and then purified again with HPLC.

Schiff base formation between fdC-oligonucleotides and probe strands

A mixture of fdC containing oligonucleotides (**T1**,**2**), or T/C/mC/hmC/caC/Abasic containing oligonucleotides (**T3-8**) in control experiments (15 μM, 20 μL, 0.3 nmol, 1 eq.), probe strands (15 μM, 20 μL, 0.3 nmol, 1 eq.), NaCl aq. (1 M, 15 μL), NaOAc aq. (pH = 6.0, 100 mM, 15 μL) and ddH_2O (80 μL) was prepared to make a final volume of 150 μL (oligonucleotide working concentration 2 μM). The mixture was heated to 85 °C for 5 min then slowly cooled down to 25 °C in 3 h. A first aliquot (15 μL) was taken and quenched before 5.4 μL of 4-methoxyaniline solution (250 mM, ddH_2O/DMSO, v/v 9/1, acidified to pH = 5.5 with acetate acid) was added to give a catalyst working concentration of 10 mM. The reaction was conducted at 25 °C, 500 rpm for 24 h. Aliquots (15 μL) were taken at 1, 2, 4, 6, 8, 12, 24 h and quenched by addition of loading buffer. All the samples were heated at 85 °C for 3 min followed by PAGE assay as mentioned above.

When using 1,4-benzenediamine as the catalyst, a 10 mM stock solution of 10 mM in 0.5% acetic acid aq. was prepared.

Melting point experiments

Melting profiles were measured on a JASCO V-650 spectrometer using quartz glass cuvettes with 10 mm path length. The samples contained 100 mM NaCl, 10 mM NaOAc buffer pH 6.0 and 1 μM of each strand in a final volume of 200 μL. The measurement was repeated three times with independent sample. Before the measurement, the oligonucleotides were hybridized by slowly cooling down the samples from 85 °C to room temperature. The solutions were covered with silicon oil and tightly plugged. Absorbance was recorded in the forward and reverse direction at temperatures from 25 °C (or 15 °C) to 85 °C with a slope of 1 °C/min. T_M values were calculated as the zero-crossing of 2^{nd} derivate of the 339 nm background corrected change in hyperchromicity at 260 nm.

Table S1. Synthesized oligonucleotides in this study. (Letters in bold and italic stand for 2'-*O*-propargyl nucleosides or epigenetic modifications.)

Entry	Description	5'----------3'	Calc.	Exptl.
T1	ODN-fC	GTAATG**fC**GCTAGG	4040.9	4036.2
T2	fC-shift	GTAAT**fC**CGCTAGG	4000.9	3999.6
T3	fC-T	GTAATG**T**GCTAGG	4027.7	4024.7
T4	fC-C	GTAATG**C**GCTAGG	4012.7	4008.6
T5	fC-mC	GTAATG**mC**GCTAGG	4026.7	4021.6
T6	fC-hmC	GTAATG**hmC**GCTAGG	4042.7	4039.6
T7	fC-caC	GTAATG**caC**GCTAGG	4056.7	4054.9
T8	fC-Abasic	GTAATG**AP**GCTAGG	3919.6	3916.4
P1	Probe-1-alkyne	CC***U***AGCGCATTAC	3932.7	3930.5
	Probe-1-MMTr	CC***U***AGCGCATTAC	4335.2	4335.1
	Probe-1-ONH_2	CC***U****AGCGCATTAC	4084.8[a]	4084.2
P2	Probe-2-MMTr	CC***U***ATCGCATTAC	4310.2	4310.3
	Probe-2-ONH_2	CC***U****ATCGCATTAC	4059.8[a]	4064.3

[a] contains one sodium ion

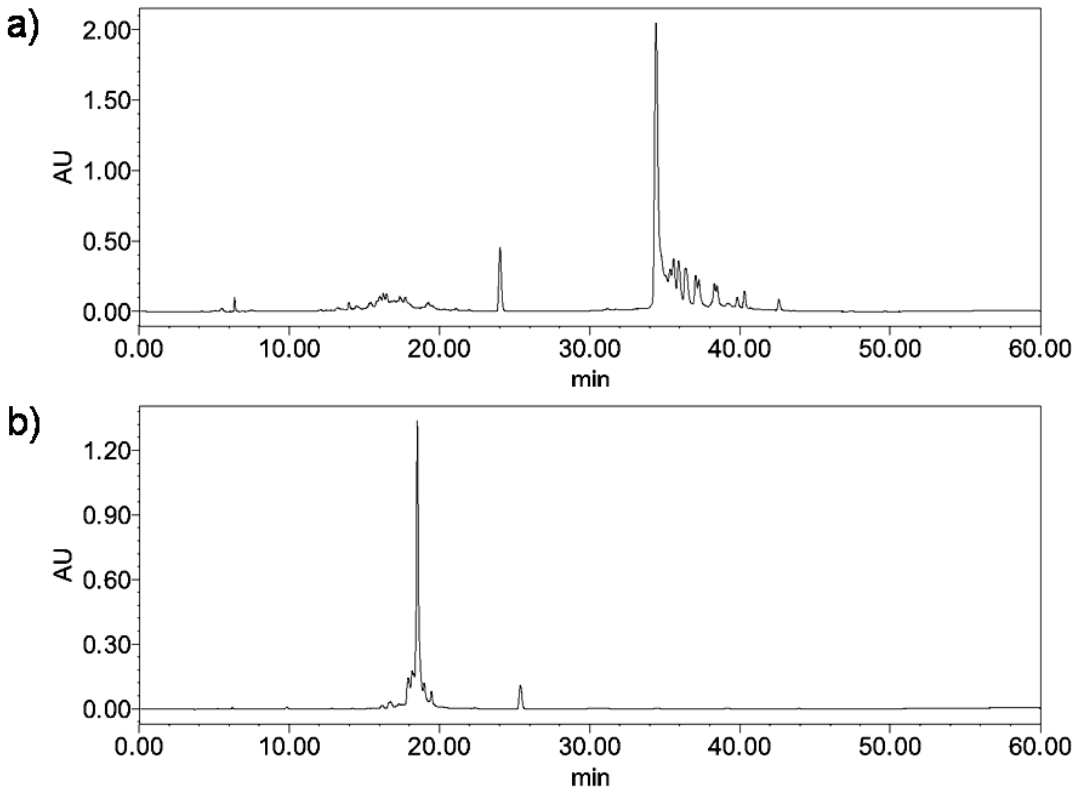


Figure S1. Typical HPLC trace of crude product a) **P1-MMTr** and b) **P1-ONH2**. Conditions: buffer A, 0.1 M TEAA; buffer B, 0.1 M TEAA in 80% acetonitrile, linear gradient from 0% to 60% B over 45 min. Retation time: (a) 34.4 min, (b) 18.5 min. AU = arbitrary unit.

5. Crosslinking studies with the synthesized strands

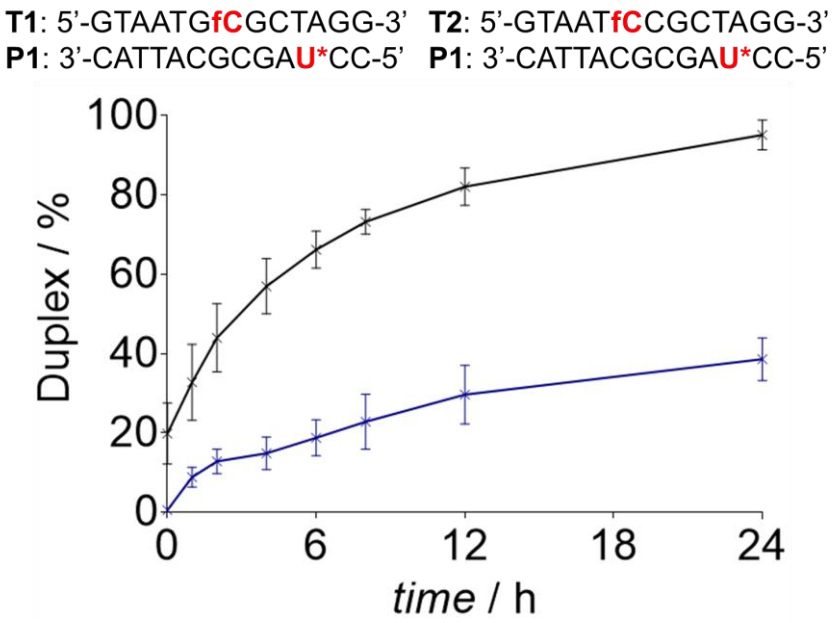


Figure S2. Quantification of the DNA duplex formation during the reaction using the catalyst 4-methoxyaniline. Black line: duplex formation between **T1** and **P1**, blue line: duplex formation between **T2** and **P1**. Error bars represent the standard error of the mean calculated from three replicates.

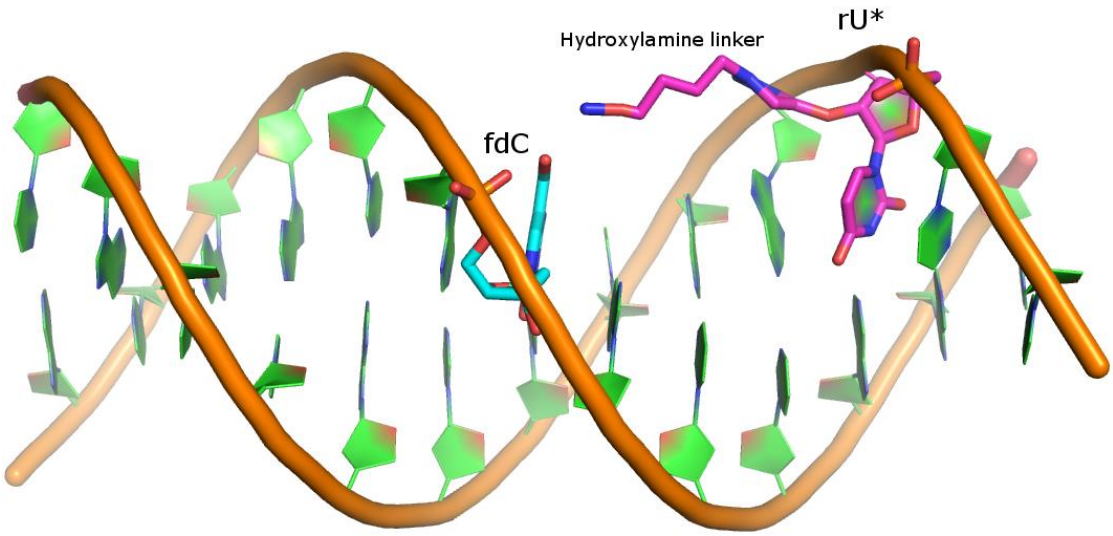
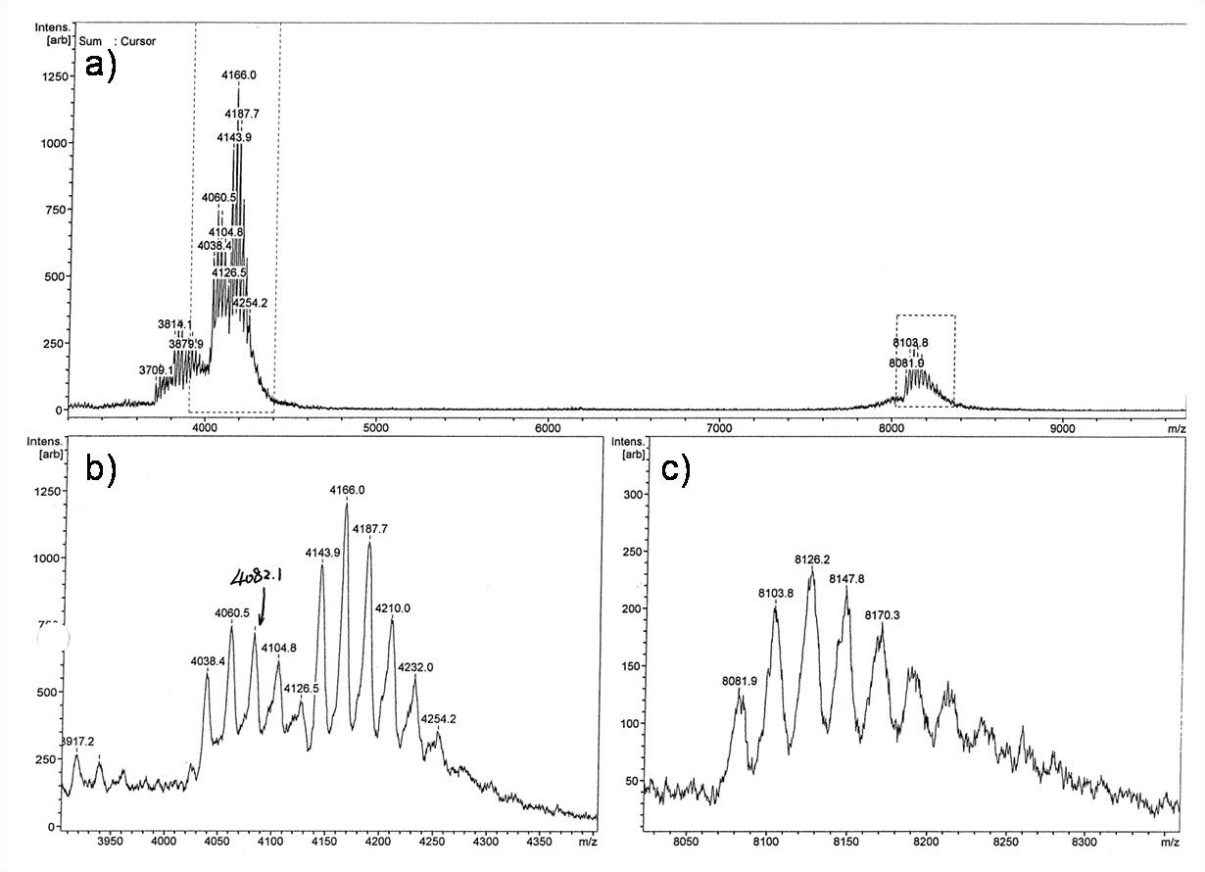


Figure S3. Model representation of a duplex showing the position of the hydroxylamine linker relative to the fdC on the complementary strand.



No.	5'----------3'	Calc.	Exptl.
P1	CC***U*****AGCGCATTAC	4061.8	4060.5
T1	GTAATG**fC**GCTAGG	4040.9	4038.4
T1	GTAATG**fC**GCTAGG	4146.0[a]	4143.9
T1:P1		8084.7	8081.9

[a] conjugate with 4-methoxyaniline

Figure S4. MALDI-TOF mass spectrum of the crosslinked duplex **T1:P1** and single strands. a) Overall MALDI-TOF spectrum; b) peaks corresponding to single strands **P1** and **P1**- 4-methoxyaniline conjugate; c) peaks corresponding to linked duplex **T1:P1.** Conditions: 10 µM oligonucleotides, 100 mM NaCl, 10 mM NaOAc buffer pH 6.0.

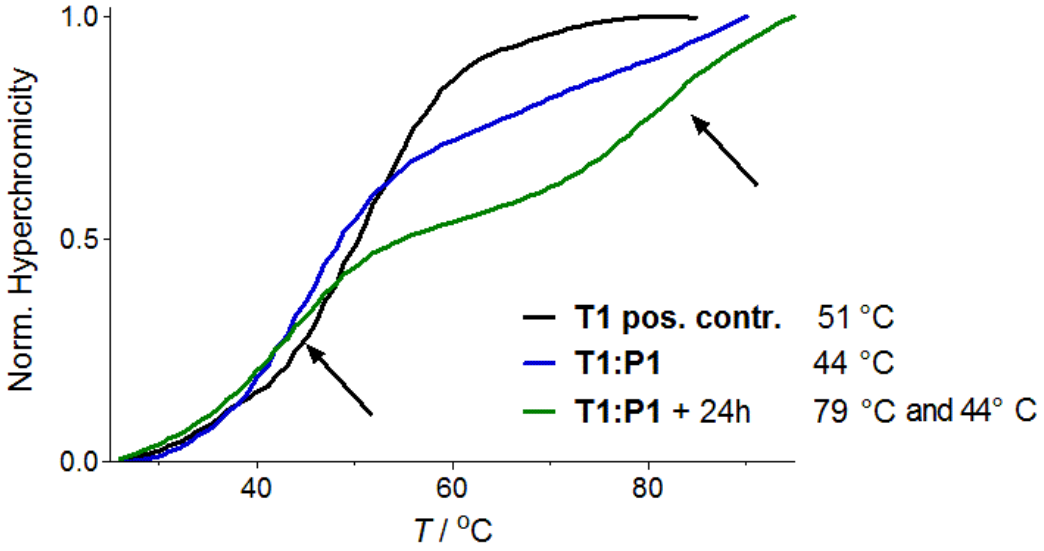


Figure S5. Melting curves of duplex **T1**:**P1** after reannealing or after 24 h incubation without catalyst compared with duplex **T1** and its counter strand (positive control). Conditions: 1 μM oligonucleotides, 100 mM NaCl, 10 mM NaOAc buffer pH 6.0, the final volume of 200 μL.

6. Experimental details of the genomic fdC profiling study

Cell culture and genomic DNA isolation

J1 wild type stem cells (strain 129/SvJae),[1] Dnmt TKO (J1, strain 129/SvJae),[2] $Tdg^{+/-}$ (E14, strain 129/Ola) and the $Tdg^{-/-}$ cell line (E14, strain 129/Ola),[3] were routinely maintained on gelatinized plates in DMEM (Sigma-Aldrich) supplemented with 10% FBS (PAN Biotech), 1x MEM-nonessential amino acids (NEAA), 0.2 mM L-alanyl-L-glutamine, 1x penicillin-streptomycin, 0.1 mM ß-mercaptoethanol (all from Sigma-Aldrich), 1000 U/ml mouse recombinant LIF (ORF Genetics), 1 µM PD 0325901 and 3 µM CHIR 99021 (2i; both from Axon Medchem). In these conditions, global genomic mC levels are very low and its oxidized derivatives are even lower, as we described previously.[4] For the experiments, the cultures were passaged twice (over five days), in DMEM supplemented with FBS and LIF, but lacking 2i. With this strategy, primed mESC cultures were obtained and oxidized cytosine derivatives reached reproducibly higher and stable levels.[4] In case of the experiment using J1 wild type and Dnmt TKO cells, the cultures were passaged every second day over a period of six days.

Mouse embryonic stem cells were lysed directly in the plates with RLT-buffer (Qiagen). The lysates were homogenized with a TissueLyser MM400 (Retsch) for 1 min at 30 Hz and centrifuged for 5 min at 21000 xg. Then genomic DNA was isolated using the Zymo Quick gDNA Midi Kit according to the manufacturer's instruction. The concentration was measured using a Nanodrop ND-1000 (Peqlab).

Probe crosslinking

The gDNA solution obtained above (1.2 µg), the fdC probe (**P**) (1 µM, 2 µL), NaH_2PO_4-Na_2HPO_4 buffer (200 mM, pH = 6.0, 2 µL), NaCl aq. (1.5 M, 2µL), and ddH_2O were mixed to a final volume of 18 µL. The mixture was heated to 95 °C for 3 min, and then cooled down rapidly to 25 °C. 1,4-Benzenediamine aq. (10 mM, 2 µL) was added and the reaction vial shaken (300 rpm) for 6 h at 25 °C. First, the mixture was neutralized with Na_2HPO_4 aq. (200 mM, 40 µL), and then purification with *NEB Monarch* PCR DNA Cleanup Kit using the binding buffer (120 µL) and eluting with the elution buffer (Tris-EDTA) (30 µL). The eluted solution was quantified with the Nanodrop and 22-32 ng/µL was obtained. UV spectra confirmed the main peak centered at 260 nm.

Ligation

The crosslinked gDNA solution (300 ng), reporter strand (**R**) (20 nM, 1 μL), Ampligase reaction buffer (10×, 2 μL), Ampligase from *Epicenter* (5 U/μL, 2 μL, 10 U) and ddH_2O were mixed to a final volume of 20 μL. The mixture was heated to 95 °C for 3 min, and then to 94 °C for 1 min, 60 °C for 1 h and back to 94 °C for 10 cycles. Then, the reaction mixture was diluted with Tris-HCl buffer (200 mM, pH = 7.6, 50 μL) before purification with *NEB Monarch* PCR DNA Cleanup Kit using the binding buffer (140 μL) and eluting with the elution buffer (10 μL). The eluted solution was quantified with the Nanodrop, obtaining 20-32 ng/μL. UV spectra confirmed the main peak centered at 260 nm.

Droplet digital PCR

ddPCR experiments were performed on a *Bio-Rad* QX100 ddPCR System. For one reaction, gDNA (6 ng), four primers (18 μM each, 1 μL), two TaqMan probes (5 μM each, 1 μL), digital PCR Supermix for Probes (no dUTP, 2×, 10 μL), and ddH_2O were mixed to a final volume of 20 μL with primer working concentration of 900 nM and TaqMan probe working concentration of 250 nM.

PCR cycles were conducted on a *Bio-Rad* T100 Thermal cycler. PCR cycle: 95°C for 10 min, 94°C for 30 sec and specific annealing temperature (64°C) for 1 min for 35 or 40 cycles, then 98°C for 10 min and cooled down to 12°C. A temperature ramp of 2°C/s was used. Droplet generation and counting were conducted according to the manufacturer's instructions, i.e. reaction mixture prepared as above (20 μL) and ddPCR droplet generate oil (70 μL) were used per reaction. The accounted droplet number was retained in 10000-18000. FAM for detection amplicon was set to channel 1; HEX for reference amplicon was set to channel 2.

Each percentage value represents the averages and standard deviations from the mean of at least two technical replicates and two biological replicates. LC-MS quantification were conducted according to the previous report.[5]

References

[1] E. Li, T. H. Bestor, R. Jaenisch, *Cell* **1992**, *69*, 915-926.

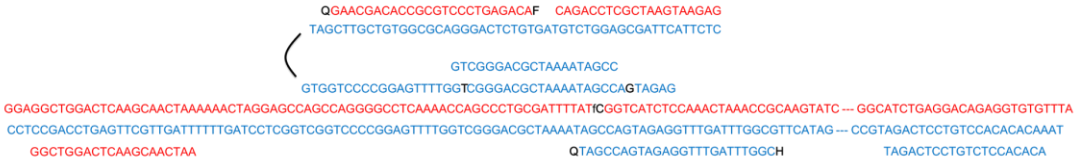
[2] A. Tsumura, T. Hayakawa, Y. Kumaki, S. Takebayashi, M. Sakaue, C. Matsuoka, K. Shimotohno, F. Ishikawa, E. Li, H. R. Ueda, J. Nakayama, M. Okano, *Genes Cells* **2006**, *11*, 805-814.

[3] D. Cortazar, C. Kunz, J. Selfridge, T. Lettieri, Y. Saito, E. MacDougall, A. Wirz, D. Schuermann, A. L. Jacobs, F. Siegrist, R. Steinacher, J. Jiricny, A. Bird, P. Schar, *Nature* **2011**, *470*, 419-423.

[4] T. Pfaffeneder, F. Spada, M. Wagner, C. Brandmayr, S. K. Laube, D. Eisen, M. Truss, J. Steinbacher, B. Hackner, O. Kotljarova, D. Schuermann, S. Michalakis, O. Kosmatchev, S. Schiesser, B. Steigenberger, N. Raddaoui, G. Kashiwazaki, U. Muller, C. G. Spruijt, M. Vermeulen, H. Leonhardt, P. Schar, M. Muller, T. Carell, *Nat Chem Biol* **2014**, *10*, 574-581.

[5] M. Wagner, J. Steinbacher, T. F. J. Kraus, S. Michalakis, B. Hackner, T. Pfaffeneder, A. Perera, M. Müller, A. Giese, H. A. Kretzschmar, T. Carell, *Angew. Chem., Int. Ed.* **2015**, *54*, 12511-12514.

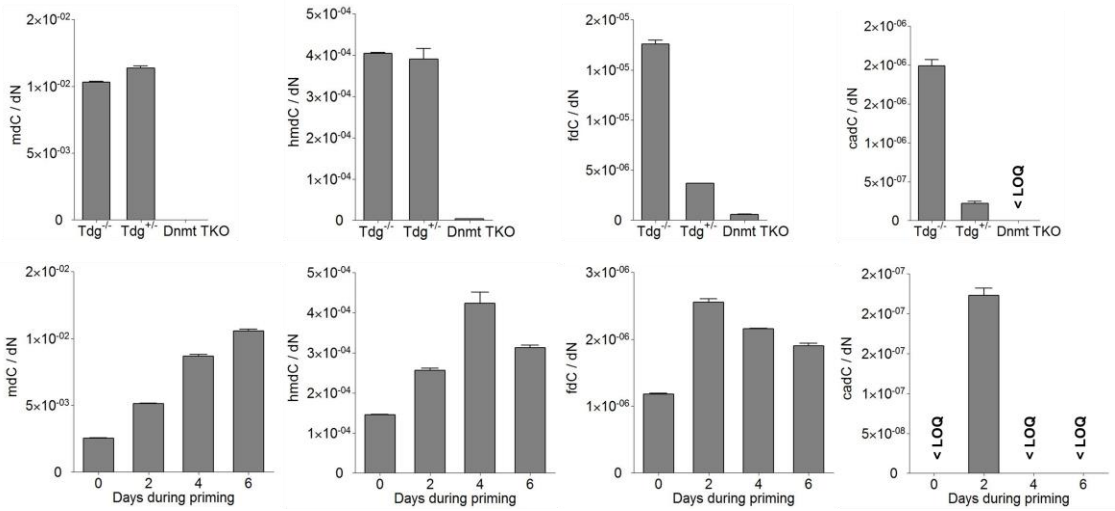
a)



b)

Entry	Description	5' ---------- 3'	mer
P2	probe for locus 1	GAGAT**G***ACCGATAAAATCGCAGGGC	25
R1	report strand for locus 1	**p**TGGTTTTGAGGCCCCTGGTGTAGCTTGCTGTGGCG CAGGGACTCTGTGATGTCTGGAGCGATTCATTCTC	70
locus 1	detection forward primer	CCGATAAAATCGCAGGGCTG	20
	detection reverse primer	GAGAATGAATCGCTCCAGAC	20
	detection TaqMan probe	**F**-ACAGAGTCCCTGCGCCACAGCAAG-**Q**	24
	reference forward primer	GGCTGGACTCAAGCAACTAA	20
	reference reverse primer	ACACACCTCTGTCCTCAGAT	20
	reference TaqMan probe	**H**-CGGTTTAGTTTGGAGATGACCGAT-**Q**	24
P3	probe for locus 2	C**A***AGCGAAGCAGGGCAAATGGCGA	24
R2	report strand for locus 2	**p**TCTCGAACCTCTGCCAGCCTAGCTTGCTGTGGCGC AGGGACTCTGTGATGTCTGGAGCGATTCATTCTC	69
locus 2	detection forward primer	AAGCAGGGCAAATGGCGA	19
	detection reverse primer	GAGAATGAATCGCTCCAGAC	20
	detection TaqMan probe	**F**-ACAGAGTCCCTGCGCCACAGCAAG-**Q**	24
	reference forward primer	CCAGGGAGCATCTGTGAAAA	20
	reference reverse primer	AAAAGCCCATCTGGGAAACA	20
	reference TaqMan probe	**H**-CCCCTTCAGACGCAAGCGAAGCAG-**Q**	24

Figure S6. Sequence detail of the detection strategy: a) Illustration of the detection strategy with sequence details for locus 1, i.e. 30,020,539th position on chromosome 16 (MM9); b) synthesized and purchased oligonucleotides for locus 1 and 2. Bold and red letters represent nucleoside modifications or functional group: p, phosphate group at 5' terminus; F, FAM; H, HEX; Q, BHQ-1. MALDI-TOF: **P2**: calc.7951.5, found 7948.5, **P3** calc. 7672.5, found 7671.0, contain one sodium ion.



	mdC		hmdC		fdC		cadC	
	pro dN	STABW%	pro dN	STABW%	pro dN	STABW%	pro dN	STABW%
Tdg-/-	1.0E-02	1.1	4.0E-04	1.1	1.8E-05	3.6	2.0E-06	7.3
Tdg+/-	1.1E-02	2.0	3.9E-04	11.3	3.7E-06	0.8	2.2E-07	22.3
Dnmt TKO	3.0E-06	28.0	3.8E-06	18.8	6.0E-07	6.8	*	*
WT0	2.5E-03	1.7	1.5E-04	0.6	1.2E-06	1.9	*	*
WT2	5.1E-03	1.6	2.6E-04	3.6	2.6E-06	3.3	2.2E-07	6.8
WT4	8.7E-03	2.4	4.2E-04	11.2	2.2E-06	0.8	*	*
WT6	1.1E-02	1.7	3.1E-04	3.1	1.9E-06	3.4	*	*

Figure S7. Global 5mC, 5hmC, fdC, and 5caC quantification using LC-MS: $Tdg^{-/-}$, $Tdg^{+/-}$, and Dnmt TKO mES cells and 0, 2, 4, 6 days during priming of wild-type mES cells. *: <LOQ, below the limit of quantification.

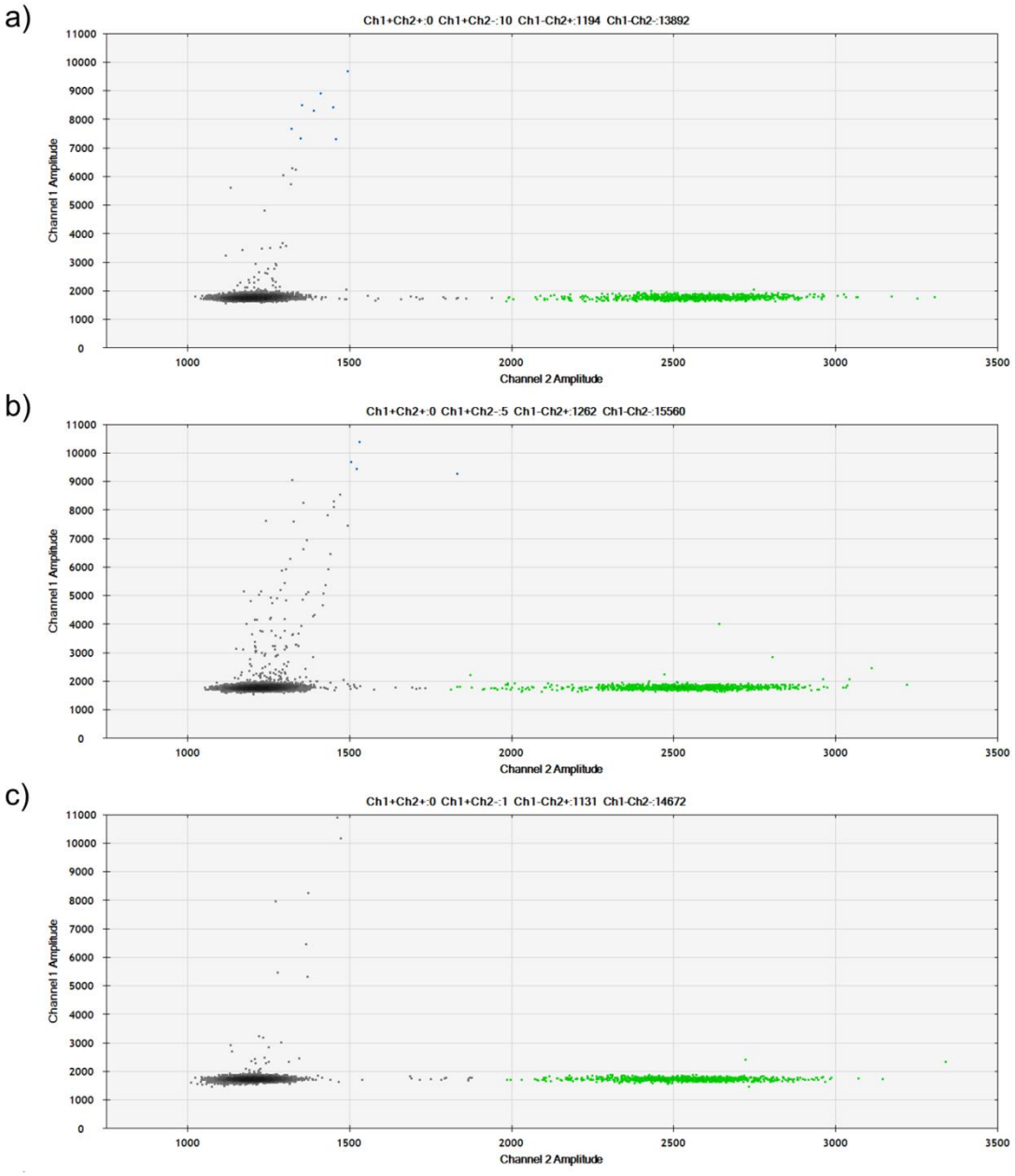


Figure S8. 2-D plot of droplet fluorescence for negative control of locus 1: a) **P3**, instead of **P2**, was used; b) no reporter stand **R1**; c) no Ampligase.

7. Quantification modeling

The encapsulation maximum of one target amplicon in one droplet to generate a positive or negative signal is the ideal scenario for our situation. If a droplet contains more than one detection amplicon, for example, one contains one fdC and one cytosine, it will show a positive signal, and the negative cytosine signal vanishes.

The probability for two or more detection amplicons to get into one droplet can be calculated according to the Poisson distribution, i.e. a discrete random variable X complies the Poisson distribution with parameter $\lambda > 0$, if, for k = 0, 1, 2, …, the probability mass function of X is given by:

$$f\ (k;\lambda) = \Pr\ (\mathrm{X} = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where e is Euler's number and *k*! is the factorial of *k*.

Table S2 Poisson distribution probabilities of genome copies in the droplet.

input ng	λ	*k*				
		1	2	3	4	5
3	0.05	4.8%	0.1%	0.0%	0.0%	0.0%
6	0.10	9.0%	0.5%	0.0%	0.0%	0.0%
9	0.15	12.9%	1.0%	0.0%	0.0%	0.0%
10	0.17	14.1%	1.2%	0.1%	0.0%	0.0%
15	0.25	19.5%	2.4%	0.2%	0.0%	0.0%
20	0.33	23.9%	4.0%	0.4%	0.0%	0.0%
30	0.50	30.3%	7.6%	1.3%	0.2%	0.0%
40	0.67	34.2%	11.4%	2.5%	0.4%	0.1%

For example, the mass of a mouse genome is approximately 3.0 pg (3.0×10^{-12} g). If 30 ng for a 20 μL reaction is used, 10,000 genomes will be distributed into 20,000 droplets. So, λ equals to 10,000 / 20,000 = 0.50. Let X = 1, then *f* (1; 0.50) = 0.303; let X = 2, then *f* (2; 0.50) = 0.076. This means 30.3% of the droplets, instead 50% of the droplet, contain a single copy while 7.6% of the droplets contain two copies. Extensive distribution probabilities are listed in Table S2. If less than 9 ng is settled in a 20 μL reaction, the probability to have two copies inside one droplet will be lower than 1%. For the ease of calculation, 6 ng gDNA is used for each reaction of 20 μL, corresponding to ca. 90 copy/μL.

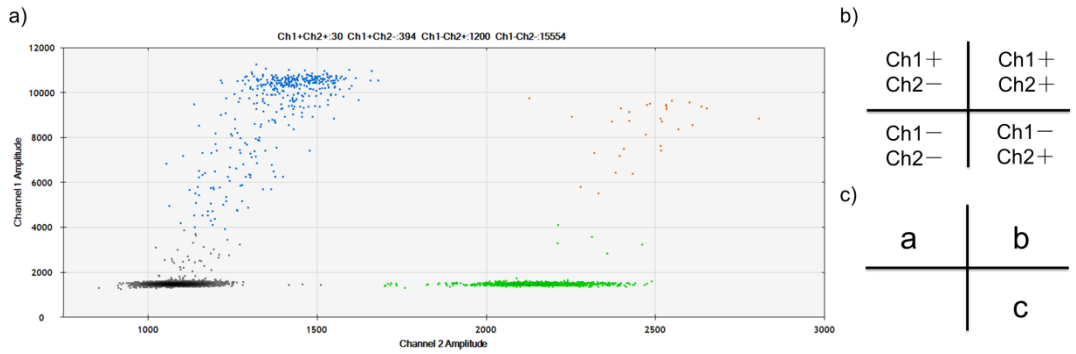


Figure S9. ddPCR output and modeling: a) 2-D plot of droplet fluorescence for illustration; b) clusters separation in four quadrants; c) algebraic simplification of counting numbers of the clusters.

As shown in Figure S8, Ch1+Ch2+ (yellow) refers to droplets with both positive signals; Ch1+Ch2- (blue) refers to droplets with only detection (report strand) signal; Ch1-Ch2+ (green) refers to droplets with only reference (gDNA) signal; Ch1-Ch2- (black) refers to droplets without target locus and ligated product; AD refers to all the droplets accepted; resolution refers to the separation of the clusters.

In principle, Ch1+Ch2+ shows the droplets that contain fdC sites in the target locus; Ch1+Ch2- indicates false-positive signals due to unspecific amplification and the dissociated ligated products; Ch1-Ch2+ shows the droplets containing only the target gDNA.

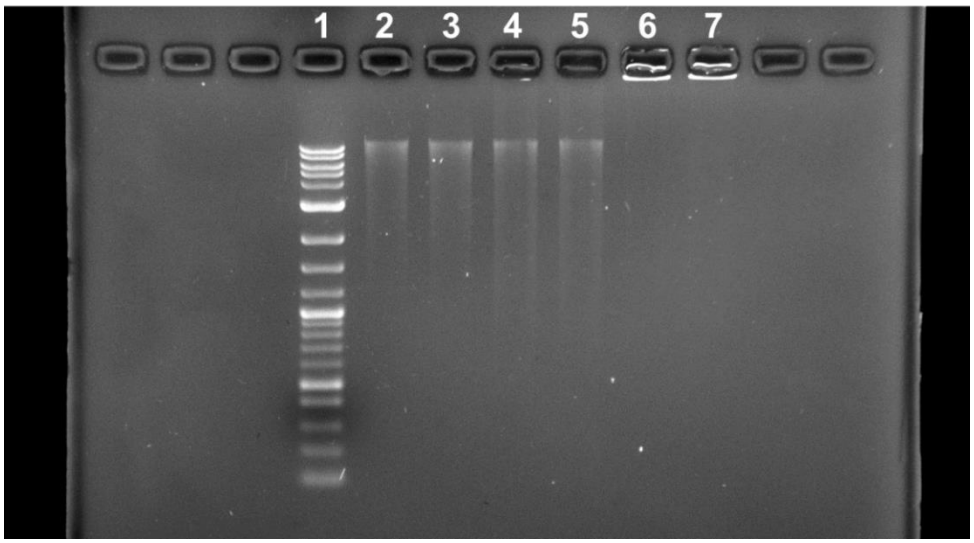


Figure S10. Agarose gel showing gDNA degradation: Line 1, log 2 marker; line 2,3, gDNA (150 ng) after crosslinking; line 4,5, gDNA (150 ng) after ligation cycle, 95°C for 3min, then 10 cycles of 94°C for 1min and 60°C for 1h; line 6,7, gDNA (150 ng).

Without considering the dissociation of the ligated products, the unreacted probe which remained in the system will cause unspecific amplification, i.e. Ch1+Ch2- signals. Catalyst, acid buffer, and ligation cycles will cause gDNA degradation (Figure S9, giving more Ch1+Ch2- false-negative signals. However, Ch1+Ch2- and Ch1+Ch2- / Ch1-Ch2- resolution do not play a role in the mathematical modelling that we used.

Assuming that all fdC at the target site is converted to the reporter strand via crosslinking and ligation, the yield is 100%. Assume that there are less than 150 copies in 1 μL so that the Poisson distribution is exclusive in our model.

Let a = Ch1+Ch2-, b = Ch1+Ch2+, c = Ch1-Ch2+, (Figure S8) A = Accepted droplet for the experiment entry, respectively, a', b', c', and A' for the control, i.e. TET knockout cell line.

Let η = fdC content of the target site.

Then,

$$\eta = \frac{\mathrm{a} + \mathrm{b} - \frac{A}{A'}(a' + b')}{b + c + (a - \frac{A}{A'}a')/\eta}$$

where $\mathrm{a} - \frac{\mathrm{A}}{\mathrm{A}'}a'$ refers to the degraded gDNA copy containing fdC at the target site that does not show in Ch2, $(\mathrm{a} - \frac{\mathrm{A}}{A'}a')/\eta$ refers to all the degraded gDNA copies.

So,

$$\eta = \frac{(\mathrm{b} - \frac{A}{A'}\mathrm{b}')}{b + c}$$

Herein, in this ideal model, without considering the dissociation of the ligated products, η is independent of a, a', and c', i.e. genome degradation and unspecific ligation do not affect fdC percentage. Ch1+Ch2- only be resulted from the dissociation of the ligated products. Also, as shown in Figure S6 a-c, b' can be omitted. Simplified η'

$$\eta' = \frac{\mathrm{a} + \mathrm{b}}{b + c}$$

is calculated to indicate relative abundance of fdC at the target site.

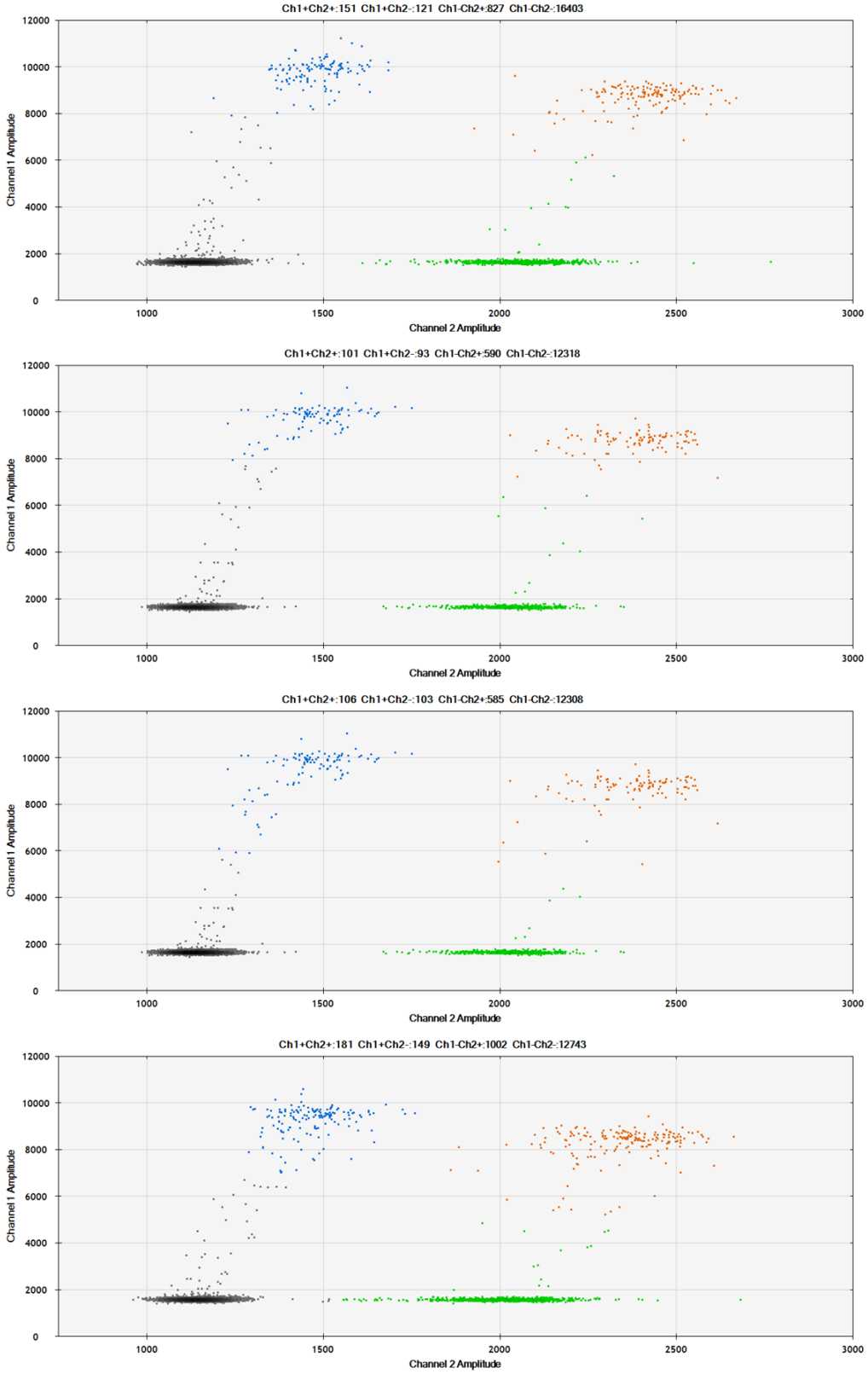
In reality, neither the fdC probe covers all the target sites nor the reporter strand ligates to all the target-linked probe. Therefore, only relative quantification is possible in our model.

8. Droplet digital PCR data

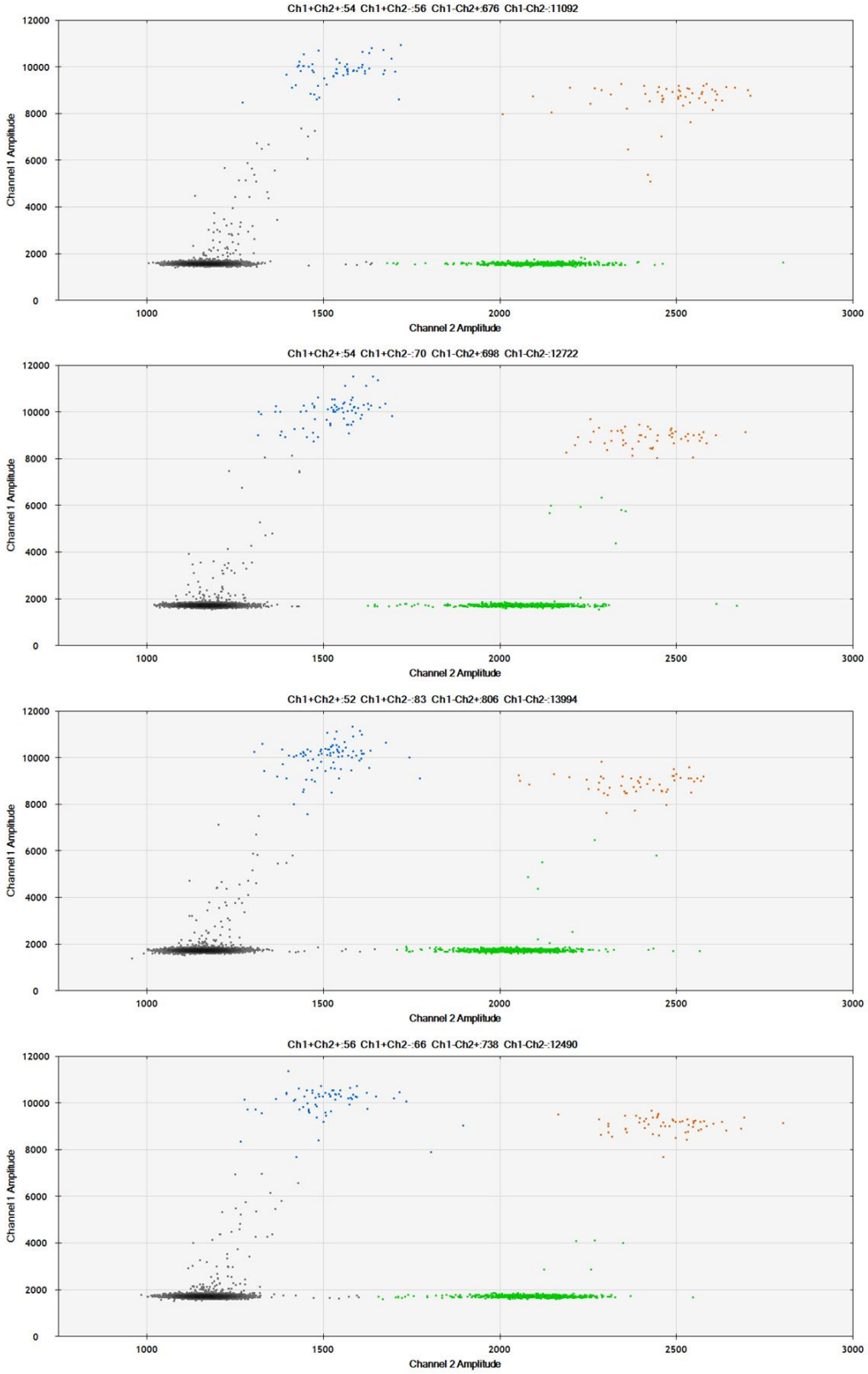
Raw data of fdC detection in $Tdg^{-/-}$, $Tdg^{+/-}$, Dnmt TKO cells for locus 1. (AD: accepted droplets)

	Ch1	Ch2	1+2+	1+2-	1-2+	1-2-	AD	η	Average
Tdg-/-	17.6	67.6	151	121	827	16399	17502	27.8%	28.5%
Tdg-/-	17.6	63.7	101	93	590	12318	14896	28.1%	
Tdg-/-	18.9	63.7	106	103	585	12308	13102	30.2%	
Tdg-/-	28.8	103	181	149	1002	12743	14075	27.9%	
Tdg+/-	11.6	74.0	54	56	676	11085	11878	15.1%	15.7%
Tdg+/-	11.3	67.2	54	70	698	12722	13544	16.5%	
Tdg+/-	10.7	69.6	52	83	806	13994	14935	15.7%	
Tdg+/-	10.8	72.0	56	66	738	12490	13350	15.4%	
Dnmt TKO	4.0	115.0	10	29	1049	10327	11415	3.7%	5.2%
Dnmt TKO	3.1	61.1	11	17	625	11913	12566	4.4%	
Dnmt TKO	4.1	58.0	14	30	620	12508	13174	6.9%	
Dnmt TKO	2.8	60.6	8	21	600	11484	12119	4.8%	
Dnmt TKO	5.7	96.0	35	28	986	11998	13047	6.2%	

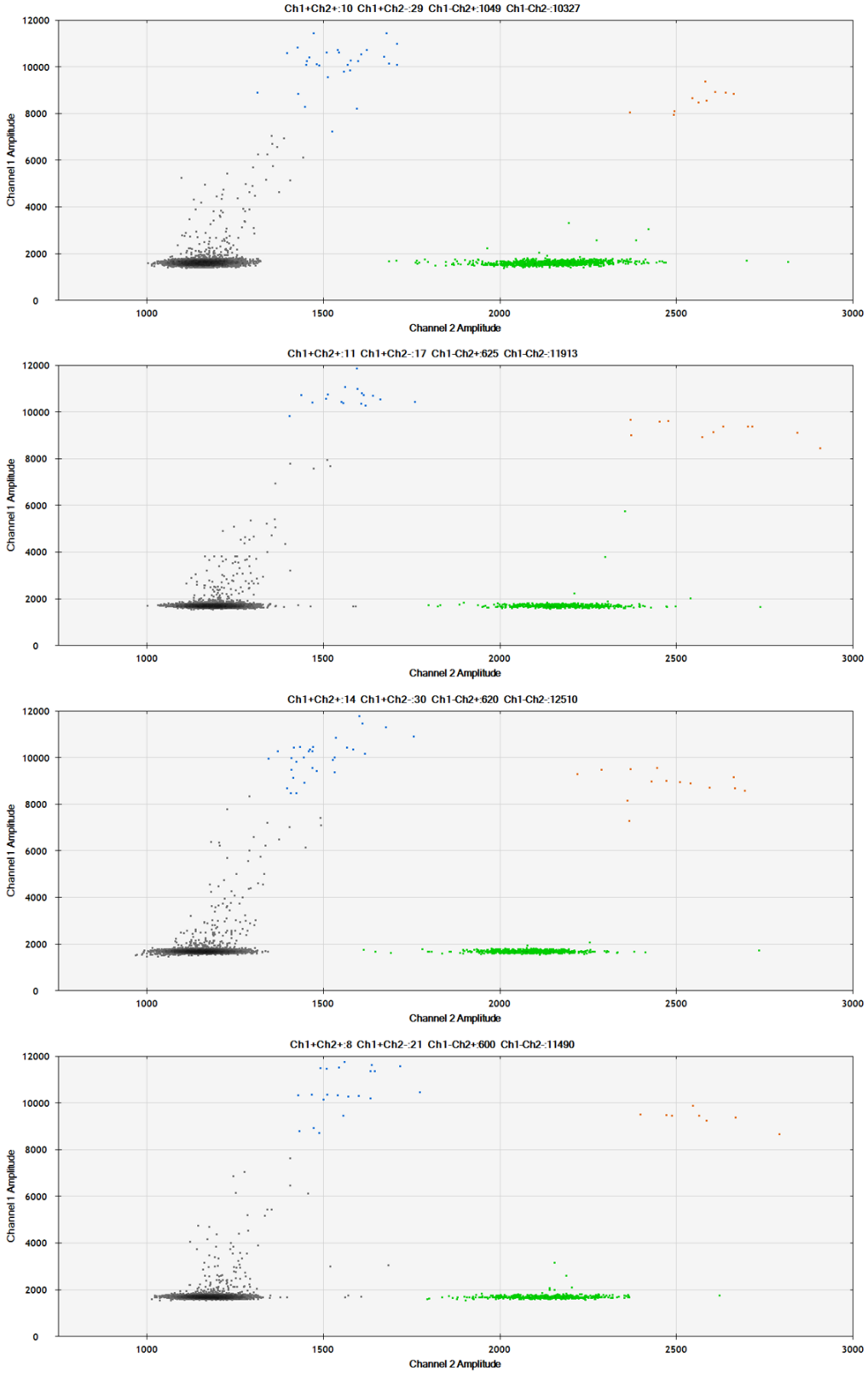
Locus 1 $Tdg^{-/-}$ mES cell sample



Locus 1 $Tdg^{+/-}$ mES cell sample.



Locus 1 Dnmt TKO mES cell sample.



Locus 1 Raw data of fdC detection in wild-type cells during priming.

	Ch1	Ch2	1+2+	1+2-	1-2+	1-2-	AD	η	Average
WT0	5.4	108	20	34	1017	10787	11858	5.21%	6.06%
WT0	6.8	110	30	58	1336	13841	15265	6.44%	
WT0	6.6	112	26	53	1255	12738	14072	6.17%	
WT0	7.8	108	23	58	1241	13060	14382	6.41%	
WT2	9.7	89.2	40	98	1193	15552	16883	11.19%	10.90%
WT2	10.2	88.7	35	105	1249	15948	17337	10.90%	
WT2	10.8	93.0	35	96	1168	14476	15771	10.89%	
WT2	10.2	96.0	54	94	1279	15637	17064	11.10%	
WT2	6.3	56.2	19	65	744	15582	16410	11.01%	
WT2	6.0	57.3	27	54	759	15780	16620	10.31%	
WT4	6.2	68.0	19	30	501	8702	9252	9.42%	8.76%
WT4	5.4	68.4	26	46	811	13942	14821	8.60%	
WT4	4.5	59.0	14	24	477	9449	9964	7.74%	
WT4	5.8	64.5	23	43	688	12579	13333	9.28%	
WT6	11.1	117	42	67	1081	10691	11881	9.71%	8.67%
WT6	8.2	114	28	41	888	9000	9957	7.53%	
WT6	9.8	120	32	41	820	7909	8802	8.57%	
WT6	8.5	114	43	57	1086	11010	12196	8.86%	

Locus 2 Raw data of fdC detection in $Tdg^{-/-}$, $Tdg^{+/-}$cells

Well		Ch1	Ch2	1+2+	1+2-	1-2+	1-2-	AD	η	Average
H04	Tdg-/-	10.5	52	55	92	650	15671	16468	20.9%	19.8%
E06	Tdg-/-	20.8	115	88	133	1084	11285	12590	18.9%	
F06	Tdg-/-	22.5	114	98	173	1226	12805	14302	20.5%	
G06	Tdg-/-	20.1	109	94	149	1178	12927	14348	19.1%	
H06	Tdg-/-	21.8	116	90	159	1179	12141	13569	19.6%	
B09	Tdg+/-	5.5	66	28	49	884	15696	16657	8.4%	9.2%
A10	Tdg+/-	5.9	59	13	29	400	8002	8444	10.2%	
E01	Tdg+/-	8.3	96	43	78	1299	15749	17169	9.0%	
G01	Tdg+/-	8.1	93	39	73	1205	15070	16387	9.0%	