Dissertation zur Erlangung des Doktorgrades

der Fakultät für Chemie und Pharmazie

der Ludwig-Maximilians-Universität München

Chemical biology - based investigation of

alternative active DNA demethylation pathway

via direct deformylation and decarboxylation

Ewelina Maria Kamińska

aus

Jelenia Góra, Polen

2023

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Thomas Carell betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde selbstständig und ohne unerlaubte Hilfe erarbeitet.

München, den 4.07.2024 …………………………..

…………………………………………

Ewelina M. Kamińska

Dissertation eingereicht am: 23.08.2023 …………………………………………

1. Gutachter: Prof. Dr. Thomas Carell …………………………………………

2. Gutachter: Dr. Sabine Schneider …………………………………………

Mündliche Prüfung am: 29.11.2023 …………………………………………

"Nothing in life is to be feared,

it is only to be understood.

Now is the time to understand more,

so that we may fear less."

Maria Skłodowska-Curie

Published Work

Parts of this thesis are already published in the scientific journals:

Schön, A.*, Kamińska, E.*, Schelter, F.*, Ponkkonen, E., Korytiaková, E., Schiffers, S., & Carell, T. (2020). Analysis of an Active Deformylation Mechanism of 5-Formyl-deoxycytidine (fdC) in Stem Cells. Angewandte Chemie (International ed. in English), 59(14), 5591–5594. https://doi.org/10.1002/anie.202000414

Korytiaková, E.*, Kamińska, E.*, Müller, M., & Carell, T. (2021). Deformylation of 5-Formylcytidine in Different Cell Types. Angewandte Chemie (International ed. in English), 60(31), 16869–16873. https://doi.org/10.1002/anie.202107089

Kamińska, E.*, Korytiaková, E*., Reichl, A., Müller, M., & Carell, T. (2021). Intragenomic Decarboxylation of 5-Carboxy-2'-deoxycytidine. Angewandte Chemie (International ed. in English), 60(43), 23207–23211. https://doi.org/10.1002/anie.202109995

Other publications:

Becker, T., Cappel, C., Di Matteo, F., Sonsalla, G., Kamińska, E., Spada, F., Cappello, S., Damme, M., & Kielkowski, P. (2021). AMPylation profiling during neuronal differentiation reveals extensive variation on lysosomal proteins. iScience, 24(12), 103521. https://doi.org/10.1016/j.isci.2021.103521

* There authors contributed equally.

Note: All illustrations enclosed in this dissertation were created by Ewelina Kamińska with a BioRender tool and publishing license.

Acknowledgements

My work would have never been so fruitful without the support of many people within and outside the lab. I would like to use this opportunity to thank everyone who made it possible to me to finish my PhD journey and who shared my happiness and sorrows along the way.

I would like to thank *Prof Dr Thomas Carell* for allowing me to join his beautiful, strong and extremely talented research group. Thank you for always raising the expectations bar a bit higher. Many thanks for giving me advice and guidance throughout my studies and for allowing me to work on very interesting epigenetics projects. I very much appreciate the support given to me, especially during my illness which was the most difficult time of my life.

Thanks to the members of AK Carell for scientific discussions, friendships, support, kindness and parties! *Dr Angie Kirchner* for being my older sister, my forever patient mentor who guided me from day one of my PhD experience. Without you I would not be here today.

Thank you *Dr Fabio Spada* for sharing your stem cell biology knowledge as well as giving me the courage to fight for my future. *Dr Markus Müller*, I am greatly thankful for the many hours you have spent on solving complicated scientific problems with me, for shining bright when the last hopes were down. *Eva Korytiaková* thank you a lot for being my project buddy, for your hard work and friendship. *Kerstin Kurz* and *Luis de la Osa de la Rosa*, thank you for your enormous help and for always being so kind. *Dr Franziska Traube*, *Dr Leander Runtsch*, *Dr Antony Crisp*, *Hanife Sahin* and *Dr Mirko Wagner* I am forever thankful for your guidance.

Many thanks to *Dr Pavel Kielkovski* and his research group members for your dedication and passion seen throughout our collaboration. Děkuji!

Dr Sabine Schneider, thank you for sharing your expertise with me, for your help and letting me use your lab space and equipment to perform the experiments.

My words of appreciation go to the members of my PhD examination committee for participating in my PhD defence process. Prof. Thomas Carell, Dr Sabine Schneider, Prof. Kirsten Jung, Prof. Konstantin Karaghiosoff, Prof. Ivana Ivanović-Burmazović and Prof. Oliver Trapp, I am humbled and deeply honored to have such excellent academic experts as part of my doctoral examination committee. Feeling privileged to have the opportunity to engage in a rigorous scholarly dialogue under your esteemed guidance.

Thank you, *Prof Dr Stefanie Kellner*, for always being there for me and teaching me the competencies of 21st century as well as mass spec skills. I feel honoured working with a woman in science who so greatly supports

Everything I achieved as a PhD student, I would like to dedicate to my late grandfather Adam, who was the first one to show me the two worlds of science and philosophy.

TABLE OF CONTENTS

SUMMARY

This is a story of studying an alternative pathway of a mechanism necessary for almost all cellular processes – from fertilisation to cellular differentiation, aging or oncogenesis. The following thesis sheds more light on the alternative active DNA demethylation pathways. DNA methylation mechanism is a basic epigenetic regulation of great importance. Alongside other developmental processes, it allows a single pluripotent stem cell, found within an early embryo, to form any type of tissue.[1] Even though all types of cells found in mammalian organisms contain the same DNA sequence, due to the regulation of DNA methylation, they acquire different functions and specific morphology.[2] Moreover, DNA methylation controls cell cycle via enzymatic activity, response to stress or apoptosis.[3–5] This first level of epigenetic modification is found on – the 5 position of cytosine. This pyrimidine may be reversely methylated as depicted in Figure 1. The interplay between methyl group writing and removal represents a one of the key processes directing gene expression patterns. Finding the exact locations and ways in which genomic 5-methylcytosine turnover occurs will give researchers the tools for epigenetic fine-tuning.

Researchers from Prof Thomas Carell's group found in the past that 2'-fluorinated (F-fdC) nucleosides successfully stop TDG's activity and still undergo some form of deformylation and decarboxylation. This raised questions about the existence of an alternative DNA demethylation pathway that bypasses the TDG-mediated base excision. We proved this hypothesis firstly using 6-aza-formyl-deoxycitidine as a feeding material for in vitro studies. Later, we were able to quantitatively trace the F-fdC to F-dC turnover in non-dividing cells, checking for a TDG-independent active demethylation pathway. This investigation is largely covered in the following thesis.

The first chapter of this thesis consists of the publication that studies the mechanism of action of the proposed but not yet identified deformylase enzyme. I introduced a specifically modified nucleoside 6-aza-fdC into thymine DNA glycosylase knockout (TDG-/-) mouse embryonic stem cells (mESCs).
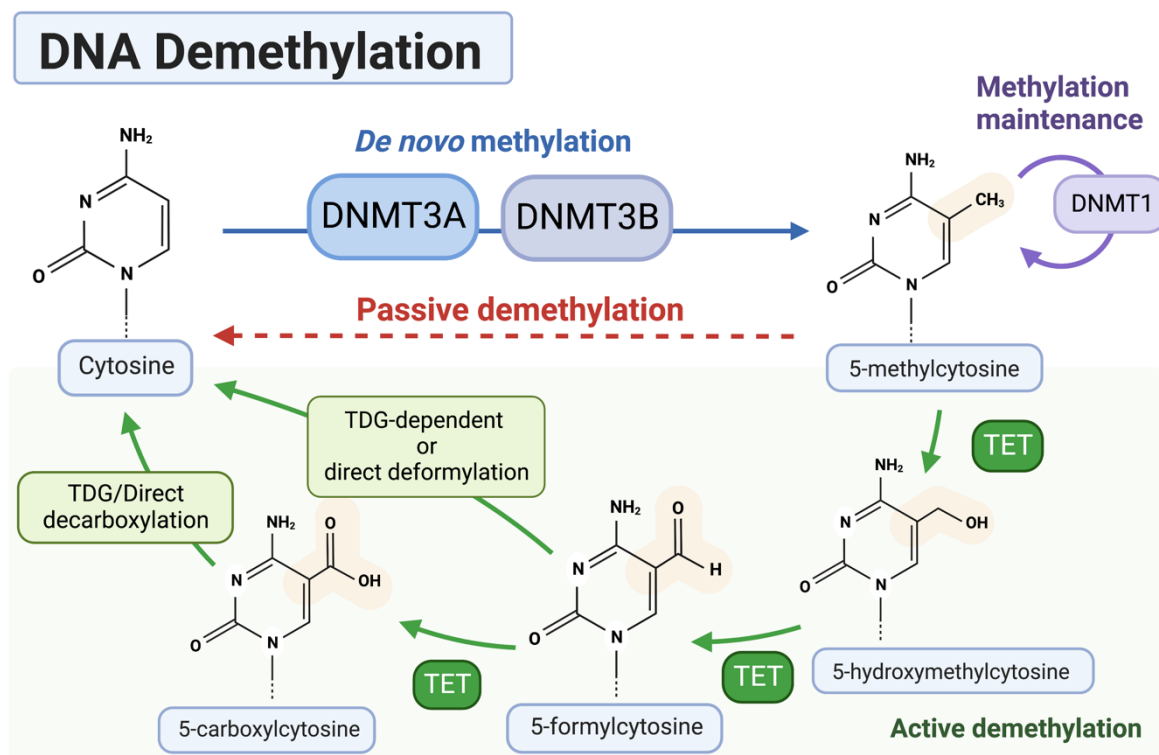
Figure 1. **DNA demethylation pathway.** Cytosine base becomes methylated by DNA methyltransferases (DNMTs) in two circumstances – *de novo* when it is newly introduced via a maintenance mechanism during DNA synthesis. In the absence of DNMT1, 5-methyl-deoxycytidine (5mdC) mark is passively by genome replication. However, it may also undergo the series of oxidations catalysed by Ten-eleven-translocation (TET) Iron^{2+}/α-Ketoglutarate-dependent enzymes. The last two oxidative cytosine derivatives – 5-formyl-deoxycytidine and 5-carboxyl-deoxycytidine can be recognised as lesions and excised by Thymine DNA Glycosylase (TDG) or directly removed via an alternative route using C-C bond cleavage. *Created using BioRender tool.*

This modification featured a nitrogen atom instead of a carbon atom at position C6 of the base. The aim of this design was to block the nucleophilic activation that was previously postulated to be required for the direct deformylation process to occur. The general workflow of all feeding experiments described in this thesis is shown in Figure 2. In conclusion, we discovered that there was no deformylation product formed from 6-aza-fdC after cellular uptake. A comparative double simultaneous feeding showed that deformylation occurred in the case of other compounds, thus the lack of 6-aza-dC was not caused only by the absence of TDG enzyme activity but also because the nucleophilic attack (needed for the initiation of the alternative demethylation pathway) was blocked.

The second publication concerns TDG-independent direct 5fdC deformylation in different cell types and cell lines. We incorporated 2'-fluorinated-5-formylcytosine (F-fdC) as a probe molecule and followed its turnover into the deformylated product: F-dC. Since 2'-fluorinated dC derivatives are recognised by DNMT enzymes, we could additionally see the products of F-dC re-methylation (F-mdC).
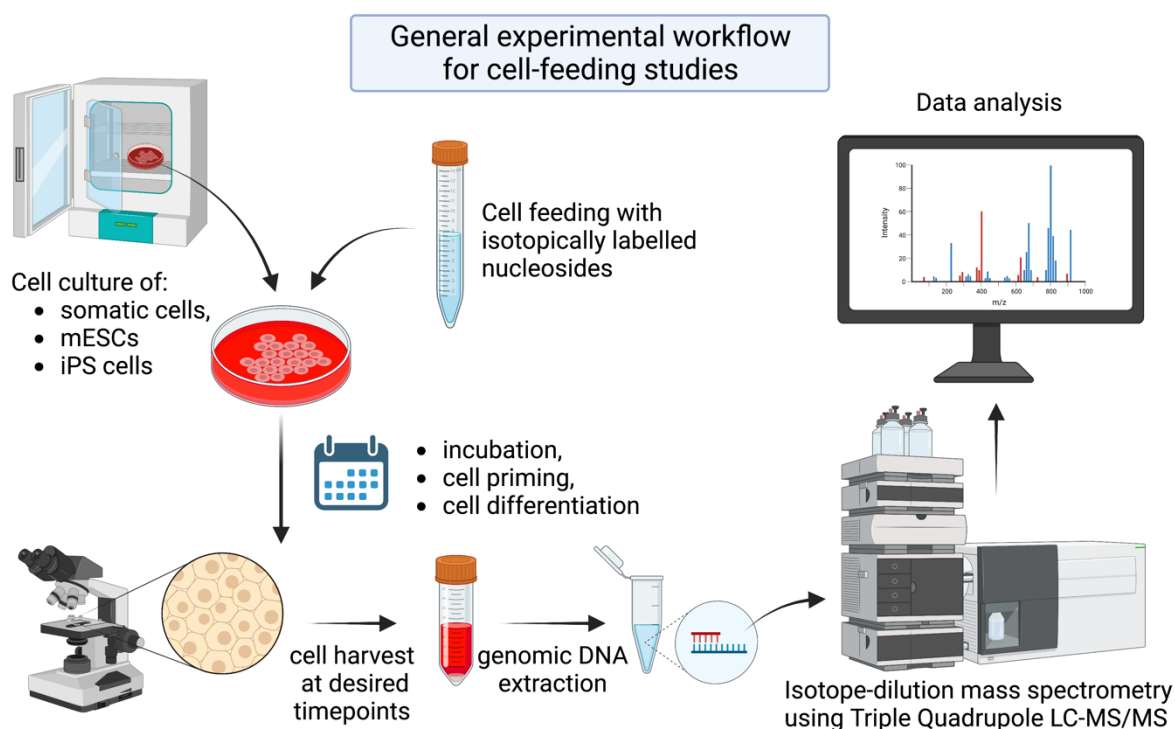
Figure 2. **The general workflow for cell-feeding studies.** The cell culture conditions depended on the specific cell line. Somatic cells were maintained in their differentiated epigenetic states, whereas the stem cells were often primed in order to increase the levels of 5mdC and its oxidative derivatives in the genomic DNA. The iPS cell cultures followed the induction and differentiation protocols. The labelled or modified nucleosides were introduced to the cells at previously optimised concentrations. The viability of the cells and their behaviour upon modified nucleoside feeding were monitored using microscopy techniques or biochemical assays. When the desired cell density was obtained, the cells were harvested and DNA was isolated. In some cases, RNA or the pool of free nucleotides were additionally investigated. The samples were measured by Triple Quadrupole LC-MS/MS. The use of isotopically labelled standards allowed the exact quantification of the measured nucleosides using a dedicated software. *Created using BioRender tool.*

The synthesis route of F-fdC was innovative because it bypassed the F-dC intermediate, allowing us to obtain the pure incorporation material without any F-dC traces. We realised that different cells, depending on their characteristics, rates of division and function,

deformylate with different efficiencies. Stem cells showed low deformylation rates but somatic cell lines of different cancers, eg. HCT-116 (colon cancer line) and MCF-7 (breast cancer line) deformylate up to 30 times more. Such results are supporting at the hypothesis, that direct deformylation may be a tightly regulated and epigenetically important process, especially when it comes to carcinogenesis. Although we could not see a direct association between cellular replication and deformylation, we wanted to test whether the division could be a confounding factor. To test this hypothesis, we differentiated inducible neurogenin-expressing iPS cells, which cease division 48 hours from induction. Even though there was no new DNA synthesised post-replication, the deformylation rates in such cells were still increasing assuring us that the direct deformylation is not a passive process.

The third chapter of this thesis describes studies of a sister mechanism – the TDG-independent direct decarboxylation. It was not an easy task, to find a 5-cadC probe molecule that would firstly, be well incorporated into the genomic DNA but secondly remain detectable and quantifiable via LC-MS/MS system. The solution for such challenge was to synthesise a triphosphate of 2'F-cadC and transfer it through the cellular membrane using a cyclodextrin transporter. As a result, the efficient triphosphate uptake resulted in a swift DNA incorporation of the desired compound. We could see the decarboxylation products slowly formed within the cells. Interestingly, this time higher decarboxylation rates were noticed in the stem cells than somatic cell lines. When quantifying the 5-cadC derivative, we managed to successfully overcome the challenge of signal suppression

INTRODUCTION

When researchers published the results of The Human Genome Project in 2003, many belived that they had uncovered the Holy Grail of medicine – the genetic code. The world anticipated a deeper understanding of the causes of both common and rare diseases. Moreover, it was expected that by knowing the sequences of genes, we could identify the DNA sequences responsible for human traits, such as appearance, intelligence or character. However, the sequencing of the human genome brought more questions than answers in the research community. Despite the unimaginable international scientific efforts to describe the genotype, more information was needed about the regulation of gene expression. It is not the sequence of a gene, but its expression level that results in the desired phenotype. Without this information, geneticists resemble musicians who are equipped with plenty of notes but still do not know when to play each sound. Despite the knowledge of the genome sequence contained within every nucleus-containing cell of our body, there was little understanding of the use of particular genes with respect to their functions. Patterns of genetic expression differentiate cells into distinct lineages, connecting their inherited genotypes with the observable phenotypes. The functional and morphological differences among the cells of a single organism can be attributed to certain modifications of variable transience found within the DNA sequence and the accompanying histones that dictate chromatin organization and alter transcription. The specific transcription-modifying structures found in each organism are named the "epigenome." The word "epigenetics" was coined by Conrad H. Waddington from Cambridge University in 1942. He described it as the task "to discover the causal mechanisms at work and to relate them as far as possible to what experimental embryology has already revealed of the mechanics of development." In his essay, based on observations of *Drosophila melanogaster*, Waddington emphasized that what links the genotype with phenotype is a complex of developmental processes, referred "epigenotype".[6,7]

Eukaryotic inheritance relies on the information encoded within the 2'-deoxyribonucleic acid (DNA) molecule, which exists in the form of a double-stranded helix. This structure was elucidated in 1953 by James Watson and Francis Crick, using Rosalind Franklin's x-ray crystallographic evidence.[8] The DNA molecule is composed of four base building blocks,

paired by hydrogen bonds – 2'-deoxyadenosine (dA) with thymidine (T) and 2'-deoxyguanine (dG) with 2'-deoxycytidine (dC). Over 3 billion of such base pairs form a complete human genome.[9] DNA is an extremely long molecule, reaching 2 m in length when stretched. To fit inside a cell's nucleus, it needs to be tightly packed and well-organised. It wraps around an octamer of histone proteins, specifically heterodimers of H2a, H2b, H3 and H4, to form a nucleosome.[10,11]
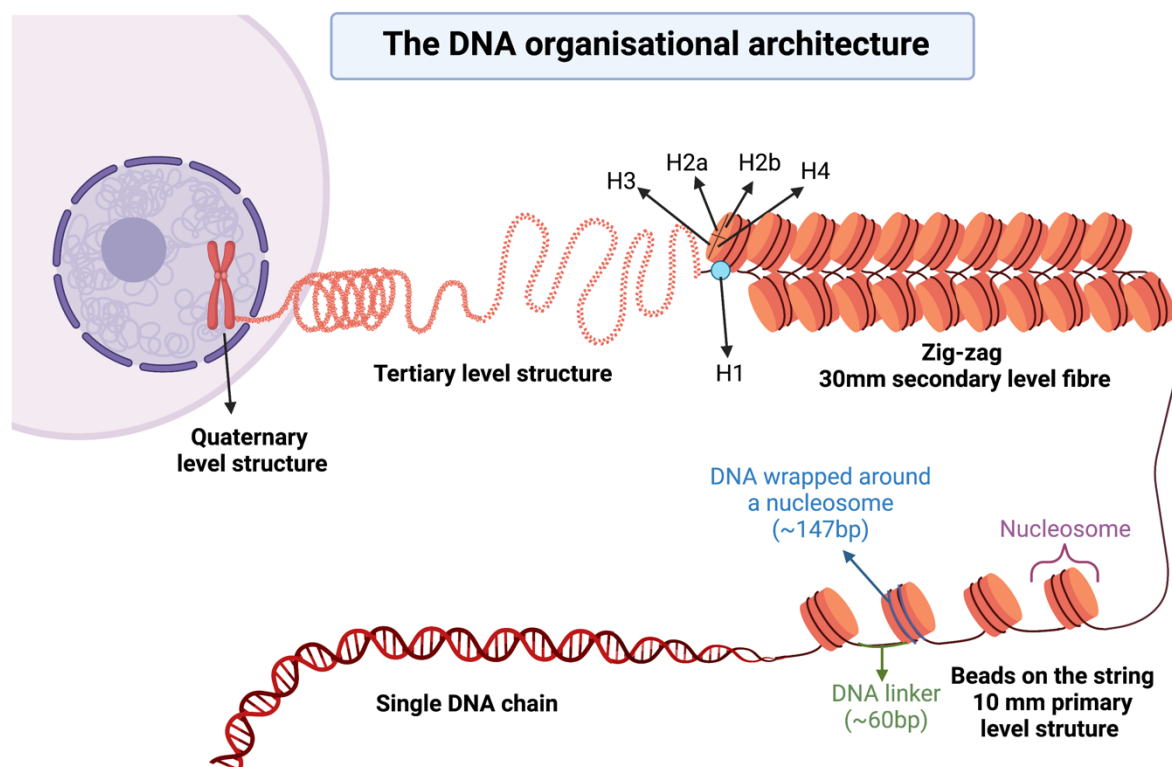
Figure 3. **The nuclear packaging system for long DNA chains.** This figure depicts the process of DNA wrapping around histone protein, along with the involvement of protein chaperones, to form various structural patterns that ultimately result in condensed chromatin, known as a chromosome. This organised structure promotes DNA stability and (most importantly), enables genes to maintain their transcriptional activity or remain silent as required. *Created using BioRender tool.*

The fifth histone, H1, is positioned on the top of the DNA-protein complex, holding its integrity and stability. It is also responsible for the formation of a higher order chromatin structure.[12] Nucleosomes, which serve as the fundamental units of chromatin, typically contain between 145 to 147 base pairs and are connected to each other like beads on a string through DNA linkers (~60 bp), forming a 10 nm primary level structure. When nucleosomes stack on the

top of each other, following a zigzag or solenoid model, they create a 30 nm secondary level fibre. Additionally, the layers of fibre loops (tertiary level) condense to form complete chromosomes, representing quaternary level structures (see Figure 3).[13] The role of chromatin extends beyond DNA packaging and encompasses the regulation of DNA's accessibility to replication and transcription processes.

The current understanding of epigenetics (the *epi-* prefix derived from Ancient Greek, meaning "on top of") emphasises its role in fine-tuning gene expression.[14] A slightly modernised version Robin Holliday's definition, by Wu and Morris states that epigenetics is "the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail change in DNA sequence."[15,16] Understanding epigenetic mechanisms provides valuable information about human development, pathogenesis and aging.[17–19] Epigenetic modifications have become hallmarks of cancer and other diseases, necessitating careful and in-depth basic science research before new clinical implications can be identified.[20]

EPIGENETIC MODIFICATIONS

The epigenetic modifications are found at different levels of chromatin organisation allowing its structure to quickly adapt gene expression patterns in response to endogenous and exogenous signals. These modifications can occur globally across the whole genome or be are targeted to specific areas. Some are found within DNA nucleobases, especially dC and dA.[21–23] Others, in a form of post-translational modifications (PTMs) are found on the N-terminal tails of histone proteins. Collectively, they control factors that lead to chromatin remodelling.[24]

DNA condensation, characterised by tighter DNA coiling and heterochromatisation, is associated with reduced accessibility to the RNA polymerase, resulting in gene silencing. In contrary, a relaxed genomic structure, euchromatin, allows for active transcription. Higher level epigenetic modifications include the actions of non-coding RNA (ncRNA) molecules. For instance, micro RNAs (miRNAs) can act directly on the transcription products, targeting the 3'-untranslated region (3'-UTR) of messenger RNAs (mRNAs). Piwi-interacting RNAs (piRNAs), in contrast, bind and regulate epigenetically significant proteins. The last group, long non-

coding RNAs (lncRNAs) are much longer (>200nt in length) and present a wide range of epigenetic functions.[25] All four levels of epigenetic transcription regulation impact each other and are depicted in figure 4.
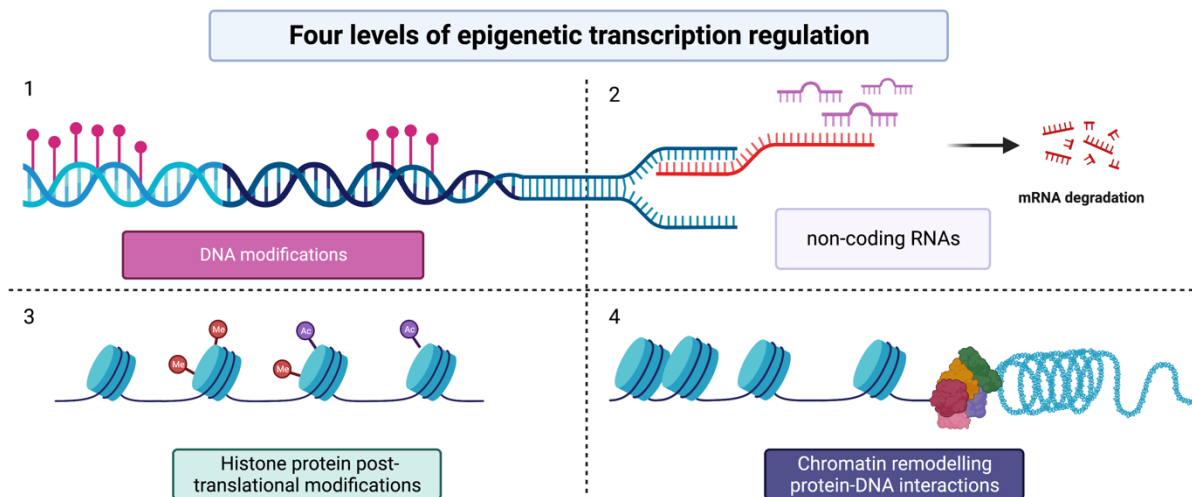

Figure 4. **The four levels of transcriptional regulation.** It consists of DNA modifications (1), histone proteins modifications (3), histone interactions with chromatin remodelling enzymes (4), and the significant role played by various types of non-coding RNAs (2) in posttranscriptional regulation. The interplay between different levels of transcriptional regulation cooperatively influences gene expression. *Created using BioRender tool.*

DNA CYTOSINE MODIFICATIONS

Cytosine may be differentially modified (see figure 5). Currently researchers look for the epigenetic importance of each modification. 5-methylcytidine (5mdC) is the best known noncanonical DNA nucleobase. It is characterised by owning a methyl group at C-5 position of the base. It was first chromatographically identified by Hotchkiss in 1948.[26] 5mdC is generated by a family of DNA methyl transferases (DNMTs). DNA methylation is associated with transcription repression, silencing of transposable elements, genomic imprinting and X-chromosome inactivation. It is also closely related to histone methylation which helps to shape chromatin structure and function. 5mdC presence is associated with specific histone marks such as H3K9me3 and unmethylated H3K4, which are associated with transcriptional silencing.[27] Increased levels of 5mdC are associated with condensed, transcriptionally inactive chromatin. The sites that are frequently methylated in human genome are dC-dG (5' to 3') dinucleotides with the exception of dC-dG rich areas (>550bp long sequences with <50% CG content) of gene promoters. They are so-called CpG islands and usually remain

unmethylated.[28,29] The human genome contains approximately 47 000 CpG islands, frequently located in the transcription regulation sites (such as promoters) of housekeeping genes.[30–33] By default, they are associated with an open chromatin structure. They destabilise DNA-nucleosome binding, thus attract other proteins. They may also contain sequence-specific sites that are consistently nonmethylated at all developmental stages of the embryo.[34,35] In turn, the intragenic CpG islands (counting for a third of all CpG islands) become methylated and are associated with transcriptional activity and chromatin modifications.[33] In differentiated cells, about 6% of the dCpdGs are methylated in a tissue-specific manner.[31,36] Theoretical studies have suggested that the hydrogen bonding within 5mdC–dG pair is slightly stronger than in unmethylated dC-dGs, potentially enhancing DNA stacking interactions and increasing the likelihood of local structural distortions.[37] 5mC is also found in both, coding and non-coding human RNAs and its distribution is not random. Thus, 5mC in RNAs may play a role in transcriptome regulation.[38–40]
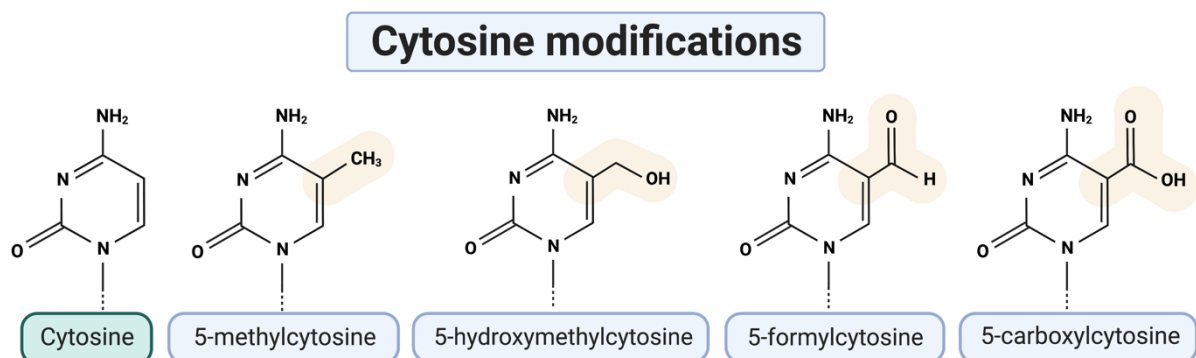
Figure 5. **Cytosine base modifications**. Cytosine undergoes methylation by DNMT1/3a/3b enzymes, resulting in the formation of a modified DNA base called 5-methyl-deoxycytidine (5mdC). This epigenetic mark may be further oxidised by TET1/2/3 oxidases to 5-hydroxymethyl-deoxycytidine (5hmdC), followed by 5-formyl-deoxycytidine (5fdC) and 5-carboyxl-deoxycytidine. *Created using BioRender tool.*

Called "the sixth base of the genome", another important epigenetic modification – 5-hydroxymethylcytosine (5hmdC) induces a gene expression regulation that is opposite to 5mdC.[41,42] It is a stable, slowly-forming modification, that is associated with active transcription and found in gene enhancers.[43–45] 5hmdC, as well as two latter oxidation products – 5fdC and 5cadC are formed by Ten eleven translocation enzymes (TET1-3)-

mediated oxidation.[46] They are 2-oxoglutarate and Fe(II)-dependent enzymes that catalyse conversion of 5mC to 5-hydroxymethylcytosine (5hmdC). They also oxidise 5hmdC to 5fdC, then to 5cadC.[47] The loss of Tet enzymes in mice leads to promoter hypermethylation and impairs proper differentiation of embryonic cells.[48] The detailed description of the TET enzymes can be found in the "Epigenetic tools" section of the introduction.

Figure 6. **The different epigenetic states of gene promoters. A)** The 5mdC mark is associated with a condensed (closed) chromatin structure and repressed transcription. The 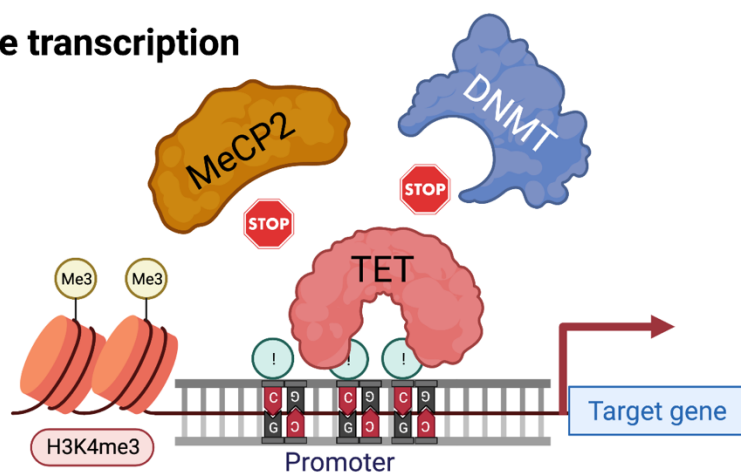DNMT enzymes catalyse the cytosine methylation reaction and are additionally recognised by the epigenetic reader MeCP2 that recruits histone deacetylases which indirectly contribute to histone 3 methylation at lysine 9 and gene silencing. **B)** When the 5mdC marks undergoes the first TET-mediated oxidation, they turn into 5hmdC, which itself is a substrate for the next oxidation. TET enzymes effectively block the docking of DNMTs and MeCP2 to DNA. 5hmdC at the promoters as well as trimethylation of H3K4 is associated with an open chromatin structure and active transcription. *Created using BioRender tool.*

Moreover, 5hmdC has been found upstream of gene start sites within mitochondrial DNA (mtDNA), reflecting a similar 5-hydroxymethylcytosine distribution pattern to genomic DNA.[49] All three 5mdC oxidation products take part in DNA demethylation via passive depletion upon

DNA replication or active reversion to the unmethylated dC. Studies have shown that the presence of TET readers on 5hmdC prevents binding of chromatin-remodelling proteins, such as MeCP2 and methyltransferases, thereby protecting promoters from methylation and what follows – deactivation (see Figure 6).[50] The sequence-specific distribution of 5hmdC has been demonstrated in stem cells.[51] Tissue-specific measurements have revealed high concentrations of 5hmdC in neurons of the Central Nervous System (CNS).[52] Decreased levels of 5hmdC in cancer tissue indicate a higher likelihood of metastasis and overall poor prognosis for patient survival.[53]

The epigenetic role of 5-formylcytosine (5fdC) remained enigmatic for years. For some it was only a by-product of active DNA demethylation or even just a DNA lesion, others believed that it is an epigenetically relevant modification of unknown function.[54–59] In 2015, Bachman et al. showed that 5fdC is indeed a semipermanent base.[54] It was later confirmed by Su et al. in 2016.[60] The fact that neither 5fdC nor 5cadC cause accumulation of harmful DNA repair intermediates in the stem cells is a supporting piece of evidence that both modifications may play an important, epigenetic role in mammalian development.[61] 5fdC global levels are 100-1000 fold lower than 5hmdC. The highest frequency of 5fdC is found in the stem cells and reach up to 20 fdC per 10^6 dCs.[62] Genome-wide 5fdC mapping in mouse embryonic stem cells (mESCs) showed its possible role in epigenetic priming within TDG-controlled regulatory elements such as enhancers.[57,63] At certain genomic locations of pluripotent cells, the levels of 5fdC is comparable to 5mdC and 5hmdC.[64] 5fdC is also known to form DNA-histone protein cross-links with histone primary amines that may have an additional epigenetic function in conjunction with DNA demethylation.[65,66] It is worth mentioning that such crosslinks of 5fdC and histone proteins lower accessibility of 5fdC, which fully stops replication or lowers transcription efficiency.[67,68] Scientists are yet unsure whether DNA-peptide cross-links are indeed transcriptionally error-prone.[69–71] The recent developments in the field of double-stranded DNA (dsDNA) kinetics show that perhaps 5fdC itself (without the accompanied histones), in contrary to 5mdC, promotes dsDNA opening facilitating gene expression.[72,73] Recent CLEVER-seq quantification showed that 5fdCpdG is present specifically at long interspersed nuclear elements (LINEs) and long terminal repeats (LTRs) in the genomes of preimplantation embryos. The authors also postulate differences between 5fdC levels seen

in the maternal and the paternal pronuclei.[74] 5fdC, similarly to 5hmdC and 5cadC is partially recognised as an oxidative stress marker, thus its presence in differentiated cells may reflect cellular reaction to stressors.[75,76] However, the influence of environmental changes on 5fdC formation in stem cells is still unknown.

5-Carboxylcytosine (5cadC) is formed upon TET-mediated oxidation of 5fdC. It is the least abundant base from all TET oxidation products. 5cadC is recognised by TDG as an undesired lesion and primarily repaired by BER pathway.[77–80] Similar to 5fdC, 5cadC is associated with actively transcribed regions of the genome.[57] However, at the base resolution, both modifications are found at low frequency (when compared to other cytosine modifications such as 5mdC) and at distinct locations, implying different epigenetic roles and arguing against their functional redundancy.[81,82] This has been confirmed by studies on human preimplantation embryos.[83] The pairing of 5cadC with guanine stimulates DNA polymerase exonuclease activity and is recognised by DNA mismatch repair (MMR) proteins during DNA replication.[84] Both – 5cadC and 5fdC decrease transcriptional efficacy.[85] The Rpb2 subunit fork region recognizes 5cadC in the major grove of the template strand. Crystal structure analysis reveals that 5cadC interacts with RNA polymerase II elongation complex, leading to a decrease the transcription rates.[86] Using Markov state models, RNA polymerase II has been recently found to pause transcription upon forming a hydrogen bond at epi-DNA recognition loop (fork loop 3 of Rpb2) with 5cadC.[87] Its functional epigenetic role may be confirmed by *in cellulo* studies. Short-term 5cadC accumulation has been observed at promoter regions of genes responsible for lineage differentiation, such as *TBx3* and *HNF-4α* as well as *TTR* and *A1AT,* known for hepatic and neural cell type formation in mammals.[88,89] However, quantifying and mapping of 5cadC poses challenges. The measurements described in this thesis were obtained from tandem mass spectrometry analysis data.[90,91] Other 5cadC detection methods are based on sequencing or PCR technologies.[92,93]

Previously, 5fdC and 5cadC were considered DNA lesions or demethylation intermediates. Nevertheless, novel findings revealed that both modifications may carry important epigenetic functions. Apart from interacting with RNA polymerase and directly influencing transcription rates, they also recruit specific transcription factors and DNA repair proteins.[94,95] In 2013, different excision mechanisms of 5cadC and 5fdC from DNA were proposed. Supposedly, a

"caC" monoanion exhibits poor acidity that promotes its resistance to TDG-mediated excision.[96] Three years later, a more advanced NMR analysis showed that N3 hydrogen bonding between 5cadC/5dfC and guanine is weekended when compared to the canonical dC:dG pairs. This facilitates TDG recognition and what follows - base excision, confirming findings by Dubini et al.[97,98] Other sources also confirm our hypothesis of direct C-C bond cleavage in human cells as an alternative DNA demethylation pathway.[99]

Non-coding RNA

Many aspects of DNA methylation or histone PTMs and in result, transcriptome regulation are linked to different RNA non-coding molecules (ncRNAs). When unravelling epigenetically important processes, it is essential to take into account extremely complex interactions between different layers of gene expression regulators found on DNA, histone, chromatin and RNA levels. Huge non-coding parts of the genome, which for years were regarded as a "genomic black matter" or "junk DNA" become more and more interesting to researchers focused on epigenetics. Long-lasting and surprising ENCODE efforts showed a high utility of regulatory non-coding elements of the genome. The best-known non-coding RNA molecules so far are microRNAs, reaching up to 25nt in length, which modify expression post-transcriptionally via binding to their cognate mRNA. These may include transcripts of epigenetic writers, readers and erasers.[100] miRNA-mediated epigenetic regulation of *DNMT1/3a/3b* genes is known to contribute to a variety of cancers, including pancreatic, cervical, breast, prostate, lung, gliomas and chondrosarcomas.[101]

On the other hand, miRNAs themselves can be regulated by circular RNAs (circRNAs) that act as "microRNA sponges". CircRNAs are single-stranded, stable loop structures that can be to some extent translated into proteins. They are believed to form through exon back-splicing, lacking 5'caps and 3' tails.[102]

Long non-coding RNAs (lnRNAs) are RNA molecules that are longer than 200 nucleotides. They are considered tissue-specific regulators of chromatin and play diverse roles in gene regulation. On one hand lncRNAs can recruit chromatin-modifying enzymes to specific gene promoter site, facilitating or inhibiting their activity. They can act as scaffolds for protein

complexes involved in chromatin remodelling and epigenetic modifications. Additionally, lncRNAs have been shown to modulate gene expression both in transcription-dependent and transcription-independent manners.[103] They can interact with transcription factors, RNA polymerase complexes and other regulatory molecules to influence gene expression patterns. Some studies show that lnRNAs can address site-specific demethylation and maintain "healthy level" of methylation by binding the GADD45A protein that recruits both TETs and TDG to the desired locations within the genome. In cancer, promoters of such lnRNAs are often hypermethylated, leading to their dysregulation and loss of function.[104]

EPIGENETIC TOOLS

Chromatin architecture is modified by a wide variety of enzymes. Some introduce DNA and histone modifications and are recognised as "epigenetic writers". Other enyzmes, so-called "epigenetic readers" are specialised proteins containing modification-matching domains that not only recognise desired chemical tags but also act and mediate biological responses. Since epigenetic changes are often reversible processes, the last group of enzymes "epigenetic erasers" are capable of removing such modifications.[105] The single-cell resolution data revealed that even two morphologically identical cells may differ in gene expression levels and chromatin organisation.[106]

THE WRITERS

DNA METHYLTRANSFERASES

The writers introduce specific epigenetic modifications. One important family of enzymes responsible for cytosine methylation is the DNA methyltransferases (DNMTs). The mechanism of action of DNMTs in establishing methylation at C-5 position of dC DNA base is well understood (see figure 7). The process begins with Cys thiol nucleophilic attack to C-6, that brings a negative charge to the cytosine ring. This allows then for methyl (CH_3) transfer from S-adenosyl-L-methionine (SAM) to C-5, then H5 proton is removed via β-elimination and 5mdC is formed. The withdrawal of DNMT's Cys, reinstalls aromaticity of the cytosine ring.[107]

Among the DNMT family, DNA methyltransferase 1 (DNMT1) is found in nearly all human tissues throughout the entire life.[108] It is especially highly expressed in the mammalian brain, where it influences neuronal plasticity.[109,110] *Dnmt1* KO mice embryos cannot sustain life after major global demethylation ends (about E10.5).[111] DNMT1 is targeted to the replication forks and CpG hemimethylated sites by UHRF1 in two ways – direct or indirect, via H3 ubiquitination.[112] When DNMT1's catalytic domain interacts with a hemimethylated DNA, it maintains dCpdG methylation patterns upon cell division.[113,114] DNMT1 binds to newly synthesised DNA strands and replicates the methylation marks, allowing to keep the same epigenetic state within the cell lineage.
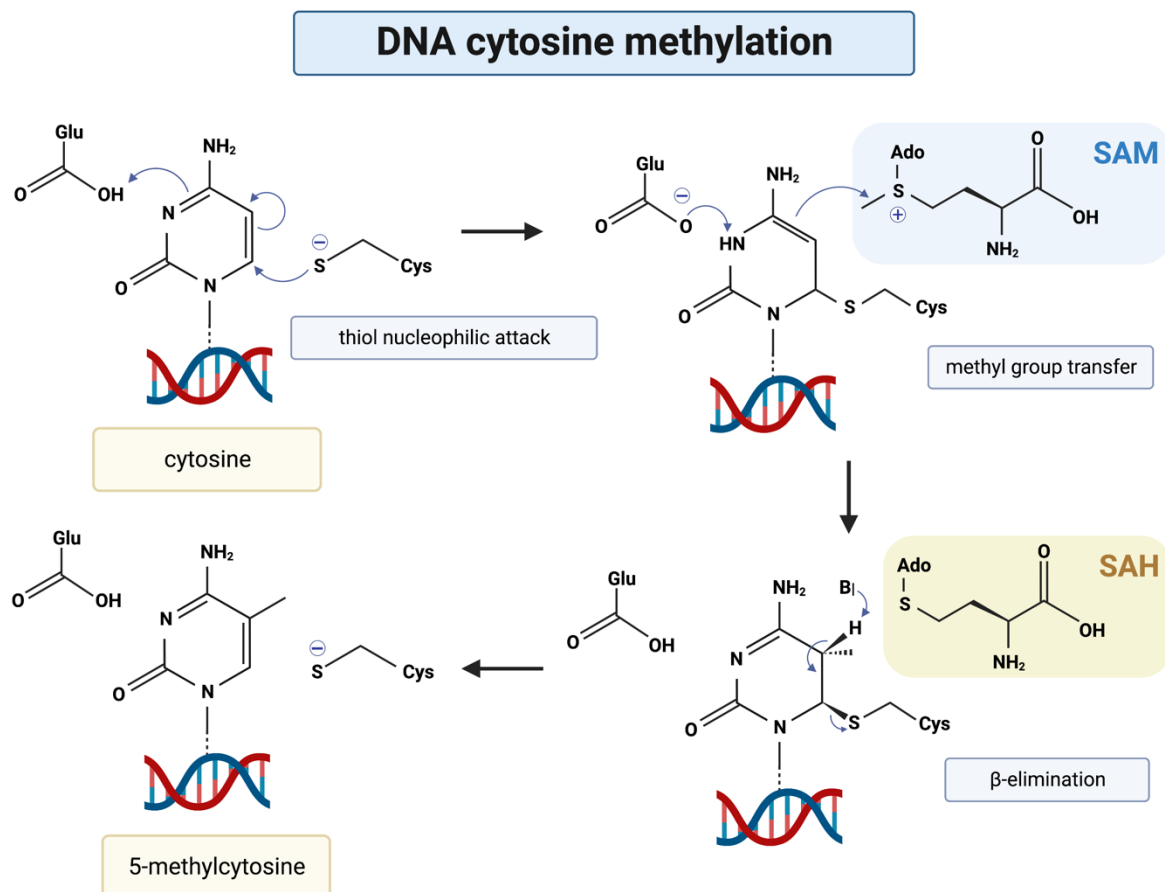
Figure 7. **The cytosine methylation by DNMT enzymes follows SAM-pathway.** This process involves three chemical transitions states (TS). At TS1, the Cys thiol nucleophilic attack brings a negative charge to the cytosine ring (-1). TS2 is characterised by the methyl group transfer from SAM, which is followed by Cys withdrawal (β-elimination) in TS3 that restores the aromaticity of cytosine and results in 5mdC formation. *Created using BioRender tool.*

What is important, is its CXXC domain which, when bound to unmethylated CpGs, exhibits an autoinhibitory mechanism towards DNMT1's MTase catalytic domain, ensuring that only semi-methylated CpGs undergo methylation.[115] The term "maintenance" also has a phenotypic significance. DNMT1 maintains longevity and its insufficiency leads to premature aging and age-related diseases.[116] Moreover, it is the only DNA methyltransferase found to accumulate at DNA damage sites, such as double strand breaks (DSB). As a result, it is also considered a potential component of the DNA mismatch repair pathway, ensuring that the methylation patterns are copied over to the newly repaired strands.[117]

Although DNA methyltransferase 2 (DNMT2) is probably the most conserved among all mammalian DNMTs, it is known to methylate only 38th cytosine within the anticodon loop of tRNA^{Asp} , which enhances the stability of this structure.[118,119] Its methylating activity towards DNA has not yet been identified. Furthermore, genomes of Dnmt2-dependent organisms do not exhibit DNA methylation patterns.[120]

DNA methyltransferases 3A and 3B (DNMT3A/3B) are referred as "*de novo*" methylating enzymes as they establish new 5mdC epigenetic marks on any CpG within the genome.[121] DNMT3B seems to be engaged in the majority of early developmental methylation events and imprinting, while DNMT3A maintains cellular differentiation throughout the lifespan. DNMT3A mutations are found in hematopoietic malignancies.[122] Since 2020, it has been known that both enzymes exhibit distinct substrate recognition patterns.[123] DNMT3A manifests a strong preference towards "CGC" and "CGT" motifs, while DNMT3B has a higher activity at "CGG" and "CGA" motifs. DNMT3B, in contrary to DNMT3A is associated with non-CG dinucleotide methylation.[124] The third member of the DNMT3 family, DNMT3L, is mainly expressed in embryos and lacks the N-terminal part of the PWWP regulatory domain.[125] Although catalytically inactive, depending on the circumstances, it acts as competitor or stimulant of *de novo* methylation.[126,127] In mice, it plays a role in establishing maternal imprints.[128] In adult tissue, it is be found only in germ cells or the thymus.

HISTONE METHYLTRANSFERASES AND ACETYLASES

Histones undergo several PTMs that contribute the histone code. These modifications among others include phosphorylation, parylation, ubiquitination, methylation and acetylation. The

last two modifications are the most prominent. Histone lysine methyltransferases (KTMs) incorporate one, two or three methyl groups on histone lysine residues. Each modification is associated with different chromatin transcriptional activity states, shown in table 1. For instance, trimethylation of H3K9 is found in the promoters of actively transcribed genes or transcriptional start sites (TSS) playing role in transcription initiation.[129] Additionally, H3K4 monomethylation is observed in active enhancers but it is still unclear whether this mark is causative of transcription activation or a consequence of it.[130] Polycomb repressive complex 2 (PRC2) is responsible for maintaining transcriptional repression through di- and trimethylation of H3K27.[105] Interestingly, different KTMs write the epigenetic marks depending on the localisation (heterochromatin or euchromatin) and the stage of methylation (mono-, di- or tri-) within the genome.

Lysine acetylation of histone proteins performed by histone acetyltransferases (HATs) leads to neutralisation of lysine charge and reduction of DNA-histone affinity. These modifications promote increased DNA accessibility to protein complexes, such as RNA polymerase II, thereby enhancing transcription, replication and DNA repair processes. Futhermore, histone acetylation modulates protein-protein interactions and is involved in chromatin modelling, as some chromatin bromodomains specifically recognise acetylated lysines.[131] HATs themselves are regulated by PTMs, primarily through phosphorylation.[132] Conversely, deacetylation of histone proteins usually correlates with gene silencing.[133] A summary of important histone modifications associated with either euchromatin or heterochromatin is depicted in Figure 8.

THE READERS

The Readers play crucial role in recognising and binding epigenetic modifications present on DNA-protein complexes, thereby mediating subsequent cellular responses. They often possess specific recognition domains and are components of chromatin remodelling complexes. DNA methylation readers, known as methyl CpG binding proteins, bind to methylated DNA and recruit chromatin transcription repression factors.

There are three different protein domain families involved in this process – MBD, SRA and zinc finger (ZF). The earliest discovered MBD domain-containing protein was MeCP2. MeCP2 recognises CpG as well as non-CpG methylated cytosines and in the past was believed to

repress transcription. Current understanding of this protein function suggests that it can function as both a transcriptional activator and repressor.[134] Several transcriptional and chromatin regulators have been found to bind 5fdC including forkhead box family (FOX) proteins, which are essential transcription factors for embryonic development.[135]

Histone modifications

Euchromatin, active transcription	
H3K4me1	H3K4me2
H3K4me3	H3K27me1
H3K36me3	H4K20me1
H3K9ac1	H3K9ac2
H3K9ac3	H3S10ph
H3K36me1	H2AXS139ph
H4K20me2	H2AT120ph
H4K8ac	H4K20me1
H3K27ac	H4K5ac

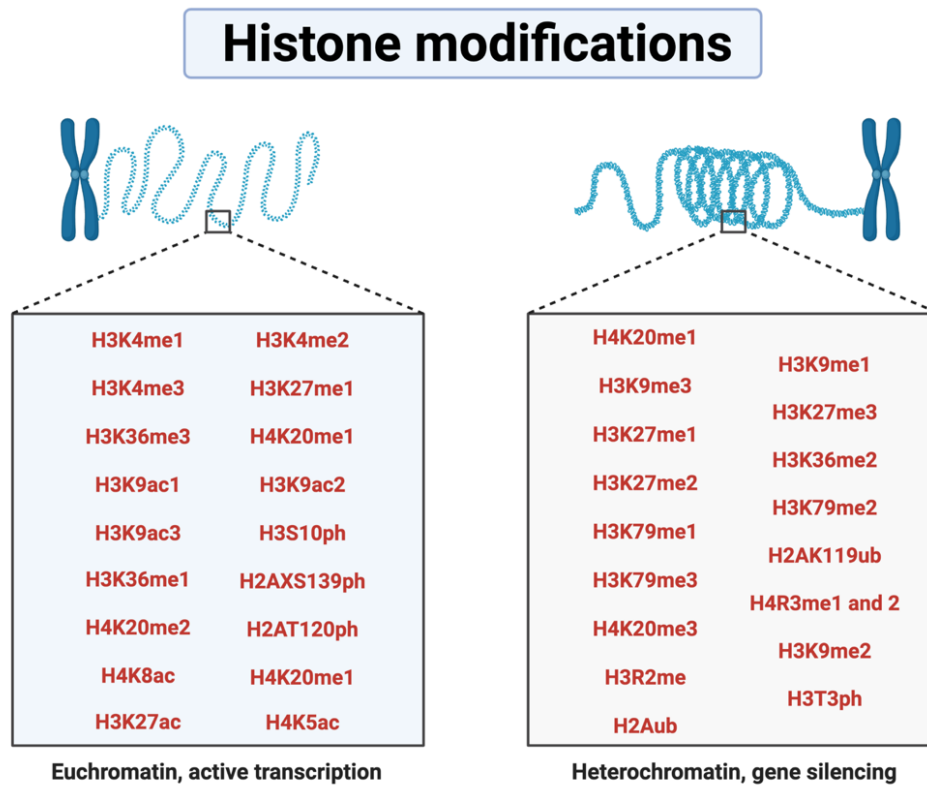| Heterochromatin, gene silencing	
H4K20me1	H3K9me1
H3K9me3	H3K27me3
H3K27me1	H3K36me2
H3K27me2	H3K79me2
H3K79me1	H2AK119ub
H3K79me3	H4R3me1 and 2
H4K20me3	H3K9me2
H3R2me	H3T3ph
H2Aub	

Figure 8. **The most common histone modifications that are associated with active transcription or gene silencing.** Methyl groups on histone residues sometimes may be associated with heterochromatin but in other circumstances, they may activate transcription.[136–138] Often, they recruit DNA modifying enzymes such as DNMT proteins. Sometimes they are damage - induced DNA repair factors or cancer biomarkers.[139–141] Histone acetylation is transcription-dependent, because in RNAPII takes part in HATs recruitment to the gene bodies.[138,142] HATs already localised at the gene promoters are unable to acetylate until transcription starts. Moreover, histone acetylation is increased at the sites where transcription is disrupted and RNAPII stalls.[143] Two histone proteins, H2A and H2B, undergo ubiquitination that plays significant role in signal transduction of transcription initiation or repression but most importantly – its elongation.[144] Increased levels of a freshly discovered histone PTM - citrullination are involved in human carcinogenesis. These repressive modifications are catalysed by peptidyl arginine deiminases and their inhibitors may serve as a future anti-cancer therapy.[145,146] *Adapted from"Euchromatin vs Heterochromatin Callout" template by Biorender.com (2022). Retrieved from https://app.biorender.com/biorender-templates*

Nucleosome remodelling deacetylases, histone methyltransferase EHMT1 as well as histone methyl-lysine binding protein L3MBTL2 have been identified as readers of 5fdC. Specific proteins are responsible for recognising and binding 5hmdC and other oxidative cytosine derivatives, and they typically do not overlap. In embryonic stem cells, these proteins are often members of DNA repair complexes.[147] Interestingly, the CXXC domain of Tet3 enzyme isoforms (Tet3FL, Tet3s and Tet3o) specifically read 5cadC, and they may have regulatory functions in transcription regulation or DNA repair.[148]

In the context of DNA methylation, DNMT3A/3B proteins, apart from their writing capabilities, can be also regarded as H3K36me3 readers.[149] Their PWPP domains recognise this epigenetic mark using it as a promoter DNA *de novo* methylation guide.[150] Neverthelss, Bröhm et al. argue that the H3K36me3 stimulatory effect on DNA linker methylation is independent from PWPP DNMT binding.[151] Similarly, the ADD domain of DNMT3A recognises and binds unmethylated H3K4, what stimulates cytosine methylation. However, binding to modified H3 lysine tail changes DNMT3A's conformation promoting its autoinhibition.[152] Histone methylation and DNA methylation are often found in close proximity within chromatin structures. These processes appear to closely correlate and coordinate their functioning.[153]

Histone methylation readers consist of various protein families. For example, polycomb-like single tudor domain-containing PHD finger protein 1 (PHF1) recognises H4K20me2, while p53 binding protein with tandem tudor domain binds demethylated H4K20. Histone demethylase JMJD2A, interacts with both H3K4me2 and H4K20me3, and UHRF1 recognises H3K9me3. One of the subunits of PRC2 promotes chromatin condensation through its binding to H3K27me3.[105] In recent years, it has been revealed that many effector proteins and signal mediators bind to acetylated lysine residues, directly influencing transcriptional regulation. Acetyl groups are recognised by bromodomains, which is why PCAF binds H3K14ac and H4K8ac, while FALZ has a specificity towards H4K16ac.[154,155] Bromodomains adopt four-helical structures that form a hydrophobic pocket where acetylated lysine inserts and weakly binds. The binding strength increases with the number of acetylated residues. Histone acetylation readers are also essential for DNA repair.[156] For example, a histone chaperone-

containing double pleckstrin homology (PH) domain and catalytic subunit of SWI/SNF associating with acetylated H3, especially H3K56ac, initiate DSB repair mechanisms.[157]

THE ERASERS

The covalent PTMs are transient, their removal is catalysed by epigenetic erasers. This allows the transcriptome to maintain a certain level of flexibility in answering to endogenous and exogenous signals through the whole life of the cell. The process of active DNA demethylation will be described in detail in later chapters, nevertheless it is the perfect moment to introduce a group of epigenetic regulators - TET family proteins.[158] Three members of TET enzymes have been identified in mammals – TET1, 2 and 3. They bind DNA via CXXC domain and oxidise 5mdC stepwise to 5hmdC, 5dfC and 5cadC.[159] In stem cells that give rise to gametes (primordial germ cells, PGCs), TET1 removes imprinted methylation patterns, while TET2 plays important part in haematopoiesis and TET3 is associated with active demethylation of zygote's paternal DNA.[160,161] Pluripotent cells show so-called "focal competition" between DNMT3 and TETs at somatic enhancers.[162] About 30% of somatic enhancers are methylated but actively recruit TET enzymes in order to facilitate cell fate lineage decisions. This kind of regulation allows for rapid changes during the cell cycle that lead to cell specification, exit from pluripotency and formation early germ layers.[163]

For many years, histone modifications were regarded as static and permanent. However, current knowledge appreciates the dynamic processes of writing and erasing of both – methyl and acetyl epigenetic marks. Histone deacetylases (HDACs) play a role in promoting chromatin condensation by catalysing the removal of the ε-amino acetyl group from lysine residues of histone tails. Mammalian HDACs are categorised into four classes of enzymes. Classes I, II and IV consist of Zn^{2+} - dependent enzymes, while class III includes the well-known NAD-dependent Sirtuin enzyme family.[164] Histone lysine demethylases (KDMs) include lysine-specific demethylases (LSD1 and LSD2) which require a protonated nitrogen and can demethylate mono- and demethylated lysines as well as a family of Jumonji C domain (JmjC) proteins catalyse the removal of methyl groups from trimethylated histone tails.[165] Arginine demethylation can be carried out by arginine deaminase, which converts monomethylated

arginine residues into citrulline. Alternatively, methyl marks on arginine can also be removed by JmjC domain-containing proteins.[166]

EPIGENETIC REPROGRAMMING IN MAMMALIAN DEVELOPMENT

Methylation of the 5 position of the cytosine nucleotide base is an ancient characteristic of eukaryotic genomes. However many model organisms such as *S. cerevisiae* or *C. elegans* are known to poses no or extremely low levels of 5mdC.[167–169] One reason why some genomes cannot afford 5mdC formation is that this modification can undergo spontaneous deamination, causing mutagenic C→T transitions.[170] On the other hand, mammalian genomes are equipped with specific DNA repair enzymes that allow them to maintain CpG dinucleotides that are up to 80% methylated, typically outside of CpG islands. DNA methylation is essential for the mammalian development and Dnmt deficiency in mammals leads to lethality at the early embryonic stage and causes various developmental malformations.[111] DNA methylation is associated with transcriptional silencing, imprints and X-chromosome inactivation and plays an important role in transposon repression. The role of 5mdC within the bodies of actively transcribed genes is not yet known.

The epigenetic tools used to methylate DNA have been discussed already; nevertheless it is important to understand that this process is reversible. The process of demethylation occurs during epigenetic reprogramming twice – shortly after the fertilisation at the preimplantation embryos and during germ cell maturation (See figure 9).[171] Studies on mouse models have shown that paternal DNA in mature sperm cells is enclosed is tightly wrapped around protamines.[172] Upon fertilisation, the protamines are replaced by highly acetylated histones, accompanied by a global loss of methylation. Active demethylation occurs when the rate of 5mdC loss is faster than it would be through DNA replication.[173] In this case, the entire paternal DNA is demethylated within 6 hours after fertilisation. This process involves several stages of Tet-mediated oxidations (eg. from 5mdC to 5hmdC, then 5fdC and/or 5cadC) and other enzymatically driven processes. However, there are cases in which the demethylation process is semi-active and semi-passive with the products of Tet-mediated oxidations (for example 5hmdC) undergoing passive dilution.[159] Demethylation must be completed before DNA replication occurs within the paternal pronucleus. The only genomic locations protected

from DNA demethylation are mono-allelic imprinted genes, heterochromatin-associated regions around centromeres and retrotransposons.[174] This is likely to maintain chromosomal stability and preserve imprints. Meanwhile, the oocyte containing maternal genome, previously stalled at metaphase II, completes the meiosis process. Maternal DNA is hypermethylated when compared to the paternal pronucleus and is epigenetically more stable. After fertilisation, it undergoes passive demethylation, gradually diluting 5mdC mark within the pool of newly synthesised DNA.
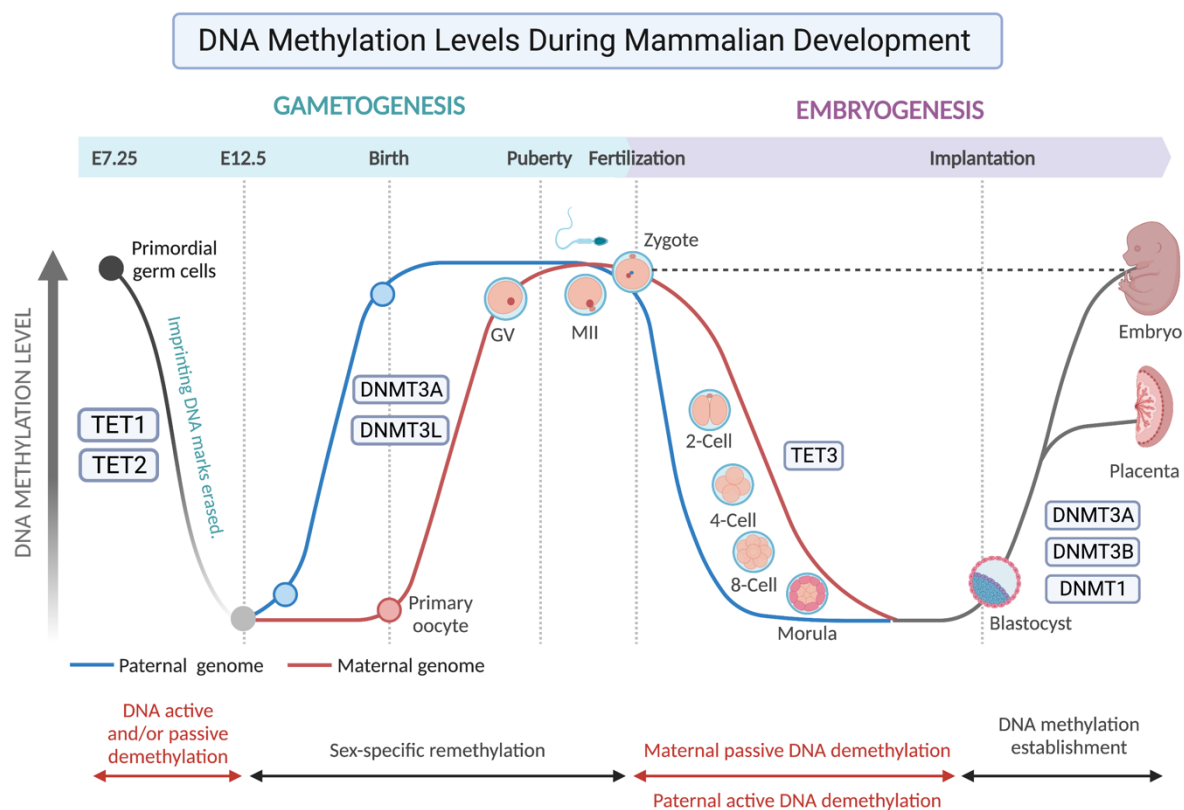
Figure 9. **DNA methylation levels during mammalian development.** There are two waves of DNA demethylation within mammalian development. Once when primordial germ cells differentiate into gametes and secondly, just after fertilisation. *Adapted from "DNA Methylation Levels During Mammalian Development" template by Biorender.com (2022). Retrieved from https://app.biorender.com/biorender-templates*

As a result, the morula, forming three days post fertilisation is characterised by the lowest genomic methylation levels. Subsequently, during blastocyst's implantation, *de novo* methylation occurs due to cell-specific expression of Dnmt3b.[175] This process allows the formation of two basic cell lineages: the inner cell mass (ICM) and trophectoderm (TE). Effectively, the ICM achieves DNA methylation levels close to those of somatic cells, while the

TE, whcih later contributes to the placenta, exhibits relative hypomethylation.[176] This demethylation is a crucial process in epigenetic reprogramming necessary for proper embryonic gene expression, cellular totipotency and lineage comittment.[177]

The second wave of the DNA reprogramming cycle happens after mid-gestation, within primordial germ cells (PGCs) of the embryo. When PGCs are germline stem cells originating from postimplantation epiblast cells that give rise to gametes.[178] When PGCs migrate to the gonads, their imprinted DNA marks are erased (either passively or actively) without detectable activity of Dnmt enzymes. After demethylation is completed, male germ cells undergo mitosis, while oogonia enter meiotic arrest at prophase 1. This state of primary oocytes presists until puberty. During gametogenesis, both germ cells undergo a sex-specific round of *de novo* methylation. This process resets the imprints and removes other undesired epigenetic marks. In general, the genomes of sperm and oocytes are hypermethylated compared to somatic cells. For a complete scheme refer to Figure 9.

DNA ACTIVE DEMETHYLATION

5mdC can be oxidised via Ten-eleven translocation family enzymes.[178] These are α-ketoglutarate and iron-dependent dioxygenases catalyze the conversion of 5mdC to 5hmdC, 5dfC and 5cadC.[46,159,179–181] Interestingly, a recent *in vitro* study suggests that another class of ketoglutarate and iron-dependent dioxygenases, ALKBH2 and ALKH3, regarded as DNA-repair enzymes, may also have the ability to oxidise of 5mdC to 5fdC.[182,183] The levels of 5mdC and its oxidised derivatives in genomic DNA (gDNA) can be directly measured using novel tandem mass spectrometry techniques with isotopically-labelled nucleosides.[62,76,184–187] Both 5fdC and 5cadC are known to be substrates for thymine-DNA glycosylase (TDG)-mediated base excision repair (BER), which restores the canonical cytosine base (see Figure 1).[188–190] TDG is a DNA repair enzyme that excises pyrimidine bases from dG-dT mismatches, and it can also modulate chromatin structure through nonspecific DNA binding.[191] TDG protects CpG islands from hypermethylation and directs active DNA demethylation of genes regulated by developmental or hormonal signals.[192] Active DNA demethylation occurs at least once during mammalian development, just after the successful fertilisation, when the paternal DNA methylation pattern is removed (see figure 9). It is not yet clear whether primordial germ cells lose their imprints through the same "active" mechanism.[193–196] The enzymes known to

mediate the active removal of methyl moieties from the dCpdG dinucleotides are TETs and TDG. Although base excision repair is a multi-step and energetically expensive process, alternative pathways involving direct C-C bond cleavage have also been proposed.[91,99,197–200] However, some (perhaps epigenetically relevant) 5fdC marks evade the BER removal, possibly remaining inaccessible to TDG.[60]

A MECHANISM OF NEURONAL DIFFERENTIATION

Another class of stem cells utilised and described in this project are doxycycline-inducible Neurogenin 1 and 2-expressing human pluripotent stem cells (iNGNs). These cells, when induced, undergo functional and phenotypic differentiation into bipolar neurons within four days. The iNGNs were derived from reprogrammed skin fibroblasts obtained from a healthy male adult and were reprogrammed by the lab of Dr George Church from Harvard Medical School using Yamanaka transcription factors – Oct4, Sox2, Klf4 and c-Myc.[201,202]
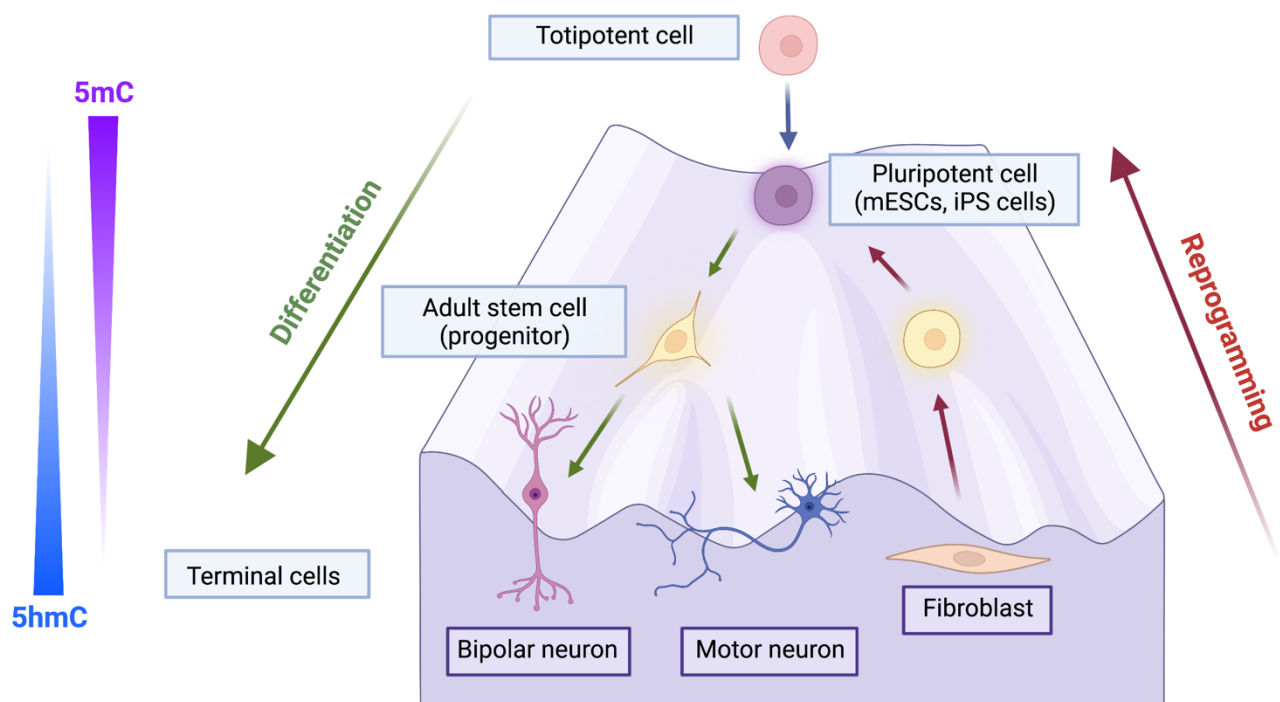

Figure 11. **The process of generating iNGN cells from mature adult fibroblast based on the Waddington's epigenetic landscape theory.** Firstly, the mature fibroblasts undergo reprogramming using Yamanaka factors, which revert them to a pluripotent stem cell state. Subsequently, a viral vector is employed to introduce a doxycycline-inducible promoter into the genome, resulting in the establishment of iPS cel line capable of differentiating into bipolar neurons. Our study demonstrates, that these cells continue to undergo

demethylation even after they have ceased replication. *Adapted from "Waddington landscape" template by Biorender.com (2022). Retrieved from* https://app.biorender.com/biorender-template

In the past, the reprogramming of differentiated human cells, such as fibroblasts, to an embryonic-like state involved transferring the content of their nucleus to the oocyte or fusing them with mESCs. Published in 2006 Yamanaka factors allowed for such programming to be performed directly within the differentiated cell in cell culture conditions.[201] This breakthrough led to the generation of induced pluripotent stem (iPS) cells, which exhibited similar morphology, function and differentiation capabilities as embryonic pluripotent stem cells. The subsequent challenge was to differentiate iPS cells into the neuronal lineage. Existing methods at the time required extensible preparations over months and resulted in heterogenic and widely variable neuron-like cells that were difficult to characterise. A new approach emerged, using neurogenic transcription factors such as Neurogenin 1 and/or Neurogenin 2, which enabled rapid neural transformation.[202–205] In the protocol published by Busskamp et al., lentiviral vector transfections were employed, wherein a doxycycline-inducible Neurogenin expression cassette was inserted into iPS cells to initiate neurogenesis and facilitate the formation of neurons within a short period of four days.[202]

It has been demonstrated that two weeks after induction, iNGNs exhibit electrical activity and transform into functional neurons. When co-cultured with astrocytes, these neurons can survive more than 100 days and express hallmarks of mature neurons.[206] Some studies have successfully achieved the expression of channelrhodopsin-2 within iNGNs, opening doors to study optogenetic characteristics and eye-like organoid formation.[206] Given that neuronal DNA contains relatively high levels of 5mdC and 5hmdC and other oxidation products of 5mdC, differentiating neurons provide an excellent model for studying DNA demethylation mechanisms. Moreover, iNGNs cease mitotic divisions within 48 hours post induction, which eliminates the confounding effects of passive DNA demethylation from the analysis. The processes of differentiation and reprogramming are depicted within the Waddington's landscape scheme in Figure 11.

AIM

5-methylcytosine (5mdC) was first identified within DNA almost a century ago and has since been found to have non-random distribution in the genome, playing crucial roles in cellular differentiation, mammalian development and cancer progression. The study of gene promoter regulation through methylation in the 1990s led to the term "epigenetic switch" to describe the influence of methylation on gene transcription activity. Today, we have the ability to detect, map and trace 5mdC within the genome as well as identify its interaction partners through immuno-proteomics studies. However, despite a century of extensive research, the pathways involved in the removal of this modification from the DNA are still not fully understood. The oxidation products of 5mdC, which are cryptic epigenetic modifications, pose additional challenges in unravelling mechanisms of their turnover during DNA methylation pathways.

The first active removal of DNA methylation happens in early embryogenesis during epigenetic reprogramming. Maternal methylation is passively diluted through newly synthesized DNA, while paternal genetic material requires at least semi-active removal of its epigenetic marks using enzymes. One known demethylation mechanism involves the Base Excision Repair (BER) process, where thymine-DNA glycosylase (TDG) removes cytosine, leaving behind an abasic site. This complex and multistep process of 5mdC turnover that poses risk to DNA double strand stability. This work aims to investigate the alternative DNA demethylation pathway(s) that may not require energetically costly DNA repair. It is possible that the fine-tuning of DNA methylation regulation may necessitate faster and safer turnover of 5mdC to cytosine through TET-mediated pathways, potentially via one or both direct TDG-independent routes: deformylation of 5fdC or decarboxylation of 5cadC.

The studies presented in the following chapters are based on optimising various cell cultures for probe molecule intake. Then the turnover of the molecules through possible DNA demethylation pathways is tracked, and quantification is performed using mass spectrometry-based analysis of isotopically labelled natural cytosine modifications and the probe molecules. 2'-fluorinated dC derivatives, known to safely incorporate into cellular DNA and block TDG activity, were commonly used as probe molecules. The first study (Chapter 1)

aims to elucidate the mechanism of direct deformylation in cellular systems. The proposed C-C bond cleavage mechanism initiated by nucleophilic activation was found to be possible in somatic as well as stem cells. However, the specific deformylase enzyme or protein complex involved in this process remains unknown.

The second study (Chapter 2) provides insights into TDG-independent active deformylation of 5fdC using 2'fluorinated-fdC probe molecule. The goal of this investigation was to explore the on and off switch of the direct deformylation and its biological relevance in various cellular states such as embryonic development, differentiation, cell division, cancer progression using different cell types.

The third published study (Chapter 3) focuses on the direct decarboxylation pathway. The main purpose of this study was to establish a method for direct decarboxylation pathway monitoring via mass spectrometry quantitative analysis. After the visibility of the decarboxylation product was established, there was a need to compare the captured F-cadC to F-dC transition rate and speed with deformylation levels in cellulo, to check which of the two pathways may be more pronounced in different type of cell. The data obtained showed that decarboxylation probably happens faster than deformylation.

MATERIALS AND METHODS

CELL CULTURE

DMEM/FBS somatic cell medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
DMEM	-	-	87
L-Alanyl-L-Glutamine	200 mM	2 mM	1
NEAA	100x	1x	1
Pen-Strep	100x	1x	1
FBS (heat inact.)	-	10%	10

Table 1. DMEM/FBS medium used for somati ccell culture.

DMEM/F12 somatic cell medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
Ham's F12	-	-	43.5
DMEM	-	-	43.5
L-Alanyl-L-Glutamine	200 mM	2 mM	1
NEAA	100x	1x	1
Pen-Strep	100x	1x	1
FBS (heat inact.)	-	10%	10

Table 2. DMEM/F12 medium used for CHO-K1 cells.

McCoy's/FBS somatic cell medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
McCoy's 5A	-	-	87
L-Alanyl-L-Glutamine	200 mM	2 mM	1
NEAA	100x	1x	1
Pen-Strep	100x	1x	1
FBS (heat inact.)	-	10%	10

Table 3. McCoy's/FBS medium used for MCF-7 cells.

RPMI-1640/FBS somatic cell medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
RPMI-1640	-	-	88.9
FBS (heat inact.)	200 mM	2 mM	10
β-ME	50 mM	0.05 mM	0.1
Pen-Strep	100x	1x	1

Table 4. RPMI-1640/FBS medium used for THP-1 cells.

DMEM/FBS mESC basal medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
DMEM	-	-	86.8
L-Alanyl-L-Glutamine	200 mM	2 mM	1
β-ME	50 mM	0.1mM	0.2
NEAA	100x	1x	1
Pen-Strep	100x	1x	1
FBS (ESC grade)	-	10%	10

Table 5. DMEM/FBS mESC basal medium used for mouse embryonic stem cells.

DMEM/FBS a2i naïve state medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
DMEM/FBS mESC basal media	-	-	87
LIF	1×10^6 U/mL	1000 U/mL	0.1
CGP 77675	15 mM	3 µM	0.02
CHIR-99021	15 mM	3 µM	0.02

Table 6. DMEM/FBS a2i medium used for mouse embryonic stem cells.

DMEM/FBS C/R priming medium			
Ingredient	Stock concentration	Final concentration	Volume in 100mL (mL)
DMEM/FBS mESC basal media	-	-	99.78
LIF	1×10^6 U/mL	1000 U/mL	0.1
IWR-1	2.5 mM	3 µM	0.1
CHIR-99021	15 mM	3 µM	0.02

Table 7. DMEM/FBS C/R medium used for mouse embryonic stem cells.

N2 100x stock preparation for N2 medium			
Ingredient	Stock concentration	Final concentration	Volume in 10mL (mL)
DMEM	-	-	7.28

Insulin	10 mg/mL in 25 mM HEPES, pH8.2	2.5 mg/mL	2.5
BSA	200 mM	5 mg/mL	10
Putrescine 2HCl	0.5 M	10 mM	0.2
Apo-Transferrin	100 mg/mL	10 mg/mL	1
Progesterone	2 mM	2 µM	0.01
Sodium selenite	3 mM	3 µM	0.01

Table 8. N2 medium used for neural progenitor cells and .

E7 medium*			
Ingredient	Stock concentration	Final concentration	Volume in 1000mL (mL)
DMEM	-	-	490
Ham's F-12	-	-	490
Pen-Strep	100 x	1x	10
L-Alanyl-L-Glutamine	200 mM	2 mM	10
L-Ascorbic Acid 2-phosphate	-	-	64mg
Nicotinamide	100 mM	0.1 mM	1
Sodium selenite	3 mM	78 nM	25.9 mg
Insulin	10 mg/mL	20 µg/mL	2
Holo-transferrin	10 mg/mL	10 µg/mL	1
NaCl	-	-	640 mg

Table 9. E7 medium used iNGN cells.**E9 medium** is composed of E7 medium + FGF-2 SI 20 µg/mL at 1:1000 and TGF-β1 100 µg/mL at 1:1000 ratio added freshly into the dish.

SOMATIC CELLS

CHO-K1 (ATCC® CCL-61™)

It is an epithelial-like, adherent cell line. A subclone from a parental CHO cell line created from a biopsy of an ovary from a Chinese hamster. Cells underwent passaging every three days at 1:8 ratio. The 3d feeding was performed at density ~100 000 cells per 60mm plate to obtained optimal confluency at the harvest timepoint.

Culture medium: DMEM/F12

Neuro-2a (ATCC® CCL-131™)

Neuroblasts originated from a mouse neuroblastoma. Their morphology resembles neuronal and ameboid stem cells. Unstable karyotype, the number of chromosomes vary between 94 and 98. Known to produce large quantities of microtubular proteins. Cells maintained by passaging in ratio 1:6 every three days. The 3d feeding was performed at density ~400 000 cells per 60mm plate to obtained optimal confluency at the harvest timepoint.

Culture medium: DMEM/FBS

HCT 116 (ATCC® CCL-247™)

Human male colorectal carcinoma cell line with epithelial morphology. The cells were passaged every two days at 1:6 ratio. The 3d feeding was performed at density ~400 000 cells per 60 mm plate to obtained optimal confluency at the harvest timepoint.

Culture medium: McCoy's/FBS

MCF7 (ATCC® HTB-22™)

Human female metastatic adenocarcinoma epithelial cell line originating from a mammary gland. Cells were passaged at 1:8 ratio every three days. The 3d feeding was performed at density ~500 000 cells per 60mm plate to obtained optimal confluency at the harvest timepoint.

Culture medium: DMEM/FBS.

RBL-2H3 (ATCC® CRL-2256™)

They are fibroblast-like basophilic leukaemia cells originating from rat. Characterised by high affinity IgE receptors and histamine expression. Cultured in DMEM/FBS medium, passaged at 1:8 ratio every three days. The 3d feeding was performed at density ~400 000 cells per 60 mm plate to obtained optimal confluency at the harvest timepoint.

Culture medium: DMEM/FBS.

Hep G2 (ATTC® HB-8065™)

They are epithelial-like cells isolated from a hepatocellular carcinoma of a teenage white male suffering from liver cancer. The cells were passaged at the ratio 1:6 every three days.

Culture medium: DMEM/FBS.

THP1 (ATCC® TIB-202™)

Are monocytes isolated from peripheral blood of a patient suffering from the acute monocytic leukemia patient. Non-adherent cells passaged at 1:6 ratio every three days.

Culture medium: RPMI-1640/FBS.

MOUSE EBMRYONIC STEM CELLS

Murine Embryonic Stem Cells (mESCs) were often used when performing experiments described in this dissertation. The cells were cultured at two stages of their pluripotency – naïve and primed, using different culturing conditions. mESCs cells are established from one of the ICM lineages of the developing blastocyst – the epiblast.[207] The preimplantation epiblast is regarded as developmental ground state, meaning it can contribute to any cell lineage of a mouse foetus. Practically speaking, mESCs are immortalised naïve epiblast, which under suitable culture conditions (e.g. the presence of leukaemia inhibitory factor (Lif)) maintain self-renewal and pluripotency. Cellular multipotency is established by the interplay of several transcription factor genes such as NANOG, OCT4, SOX2 and KlF2,4,5, as well as Esrrβ and STELLA.[208,209] Lif growth factor inhibits mESC differentiation activating Signal transducer and activator of transcription 3 (Stat3).[210] Its activation is also needed to maintain cellular self-renewal.[211] Stat3 is known to maintain stable levels of transcription factor Myc by inhibiting GSK3β-mediated phosphorylation and degradation.[212] Removal of Lif from cell culture medium results in self-renewal antagonization and promotes cellular differentiation.[213] Interestingly, the activation of Wnt/β-catenin signalling pathway has a similar, independent from JAK/STAT pathway, effect on GSK3β suppression.[214] Both rely on promoting Myc activity that propagates epigenetic stem cell identity.[215] The lack of Lif is not the only mechanism for stem cell differentiation. The activation of Erk pathway is also

responsible for stem cell lineage specification and blocking it via Gsk3 suppression using small molecule inhibitors (e.g. 2i) allows mESCs to stay fully pluripotent. Replacing Erk inhibition with Src inhibition using so-called alternative 2i (a2i) culture conditions, leads to the alternative ground state of the mESCs.[216] Those cells completely block epithelial-to-mesenchymal transition, even downstream of FGF-ERK pathway.[217] A comparison between different culture conditions is seen in Figure 10.

Different culture conditions of mESCs

Conditions	DMEM/10%FBS/2i	DMEM/10%FBS/a2i	DMEM/10%FBS/C/R
Inhibitors added	GSK3β inhibitor CHIR99021 Mek1/2 inhibitor PD035901 Leukaemia inhibitory factor LIF	GSK3β inhibitor CHIR99021 Src inhibitor CGP77675 Leukaemia inhibitory factor LIF	GSK3β inhibitor CHIR99021 Tankyrase inhibitor IWR-1-endo
DNA methylation	hypomethylated	hypermethylated	hypermethylated
mESCs characteristcis	Cells resemble ground pluripotent state but upon long-term culture become more heterogenous and sometimes able to spontaneously differentiate. They contain abnormal karyotype and erased imprints.	Morphologically homogenous cells kept in pulripotent state with normal developmental potential. Stable chromosomes with preserved imprints.	After 72 hours cells drastically change their morphology but remain pluripotent. They grow and become less round on the edges and grow as a flat monolayer. They completely rely on aerobic glycolysis.

Naive preimplantation epiblast-like cells

Noncommitted post-implantation epiblast - like cells

Figure 10. **Different mouse embryonic stem cell (mESC) culture conditions.** mESCs as in vitro models reflecting stem cells contained within pre- and post-implantation embryos are cultured in different medium conditions. Cells cultured in 2i and a2i resemble naïve epiblast (early and late respectively), while C/R medium primes cells. Importantly, a2i cultured cells have increased levels of Dnmt enzymes and express normal karyotype and in effect - chromosomal stability. Long-term 2i passaging loses imprinted genes has abnormal karyotype and impaired developmental potential.[218] *Created using BioRender tool.*

5mdC oxidative derivatives are difficult to detect in 2i-cultured pluripotent ground cell states due to the previous genome-wide demethylation process, loss of imprints and low expression of DNMTs. Generally, in naïve cells the levels of 5fdC and 5cadC are low, nevertheless in cells cultured in a2i media both modifications are at least detectable. Often also quantifiable using LS-MS/MS systems.[197] Priming converts mESCs from pre-implantation to post-implantation epiblast state. It also increases their genomic DNA methylation levels. This is the first step towards cellular lineage specification that murine embryos acquire between 4th and 5th days post fertilisation.[219] To achieve such epigenetic profile, the cells are transferred from pluripotency-promoting 2i or a2i culture conditions into glycogen synthase kinase 3β inhibitor (CHIR) and tankyrase inhibitor IWR-1-endo (IWR1) containing medium (C/R)Co for 72 hours.[220] CHIR stabilises β-catenin within the cytoplasm and IWR1 stops it from shifting inside the nucleus, protecting cells' pluripotency.[221] As a result, naïve cells, upon priming, become epiblast-like stem cells (EpiSCs).[222] A similar effect can be obtained by culturing cells for 48 hours under FGF2/Activin A conditions, generating cells transcriptionally resembling late-gastrula primitive streak cells, the early embryonic structure that establishes bilateral symmetry of mammals and other amniotes.[223] Alternatively exposing mESCs to FGF2/IWR1 conditions gives posterior-proximal epiblast-resembling cells.[224]

J1 (ATCC® SCRC-1010[TM])

(wildtype and KOs)

Male mouse embryonic stem cell wild-type line originated from inner cell mass. Derived from a male agouti 129S4/SvJae embryo (1991). The cells were passaged at 1:6 ratio every second day. Form spherical colonies. Medium was changed every day. The culture dishes require gelatin coating applied for at least 5 mins prior to seeding (0.2% gelatin in dd H_2O, autoclaved) and stored at in 37°C.

Knock-out cells used: Dnmt3b (-/-), Dnmt3a (-/-), Dnmt triple knock-out (TKO)

Culture media:

- mESC DMEM a2i for maintenance
- mESC DMEM basal medium for passaging

- mESC DMEM C/R medium for priming

Priming: seeding cells at optimal density ~500 000 cells per 35 mm well (6 well plate) in mESC DMEM C/R medium, allow 72 h to grow. After 48h morphological changes become visible.

R1 (ATCC® SCRC-1011™)

(wildtype and KOs)

A pluripotent male mouse embryonic stem cell line (originating from the inner cell mass) established in 1991 via crossing of two strains 129S1/SvlmJ and 129X1/SvJ. Harvested from a 3.5 day blastocyst. Form spherical colonies. The cells were passaged at 1:6 ratio every second day. Medium was changed every day. The culture dishes require gelatin coating applied for at least 5 mins prior to seeding (0.2% gelatin in dd H_2O, autoclaved) and stored at in 37°C.

Knock-out cells used: Sirt1 (-/-), Sirt6 (-/-)

Culture media:

- mESC DMEM a2i for maintenance
- mESC DMEM basal medium for passaging
- mESC DMEM C/R medium for priming

Priming: seeding cells at optimal density ~500 000 cells per 35mm well (6 well plate) in mESC DMEM C/R medium, allow 72h to grow. After 48h morphological changes become visible.

ES-E14TG2A (ATCC® CRL-1821™)

(wildtype and KOs)

A blastocyst-derived, pluripotent mouse embryonic stem cell line - 129/Ola strain (1987). Form spherical colonies. The cells were passaged at 1:6 ratio every second day. Medium was changed every day. The culture dishes require gelatin coating applied for at least 5 mins prior to seeding (0.2% gelatin in dd H2O, autoclaved) stored at 37°C.

Culture media:

- mESC DMEM a2i for maintenance

- mESC DMEM basal medium for passaging

- mESC DMEM C/R medium for priming

Knock-out cells used: TDG(-/-)

MOUSE FGF-2/EGF-DEPENDENT NSCS

Neural Stem Cells established by Dr Fabio Spada (Prof. Carell Research Group) from E10 mESCs.

Totipotent cells forming "lattice" networks on gelatin - coated dishes. Cells were passaged every three days at 1:6 ratio. When stimulated with epidermal growth factor (EGF) or fibroblast growth factor (FGF-2) are capable to produce neuronal precursors, astrocytes or oligodendrocytes.

Culture medium: N2 (the cells were used in pluripotent state only)

IPS CELLS

iNGN CELLS

(obtained from Prof. Church Group, Oxford)

Neurogenin 1 and 2 – induced human iPS cells. For maintenance, passaged at the ratio 1:4 every three days without induction. They need to be plated in presence of ROCK inhibitor - Thiazovivin at 1µM concentration in medium. For maintenance at the pluripotent state E7 medium is required with addition of two growth factors - FGF2 (2 ng/mL) and TGF-β1 (100 ng/mL). Both replenished every 24 hours. The induction process requires passaging the cells and plating them in the new media - E7 medium supplemented by doxycycline at 0,5 µg/mL concentration. The induction process undergoes as follows:

Culture media:

- E7 medium + growth factors: FGF2 (2ng/mL) and TGF-β1 (100ng/mL). for maintenance at the pluripotent state, the growth factors need to be replenished every 24 hours.

- E7 medium supplemented by doxycycline at 0,5 µg/mL concentration for the induction process

Coating medium			
Ingredient	Stock concentration	Final concentration	Volume in 500mL (mL)
DMEM	-	-	248.25
Ham's F-12	-	-	248.25
Pen-Strep	100x	0.45x	2.5
β-ME	1000x	1x	1

Table 10. Coating medium for iNGN cell culture plates.
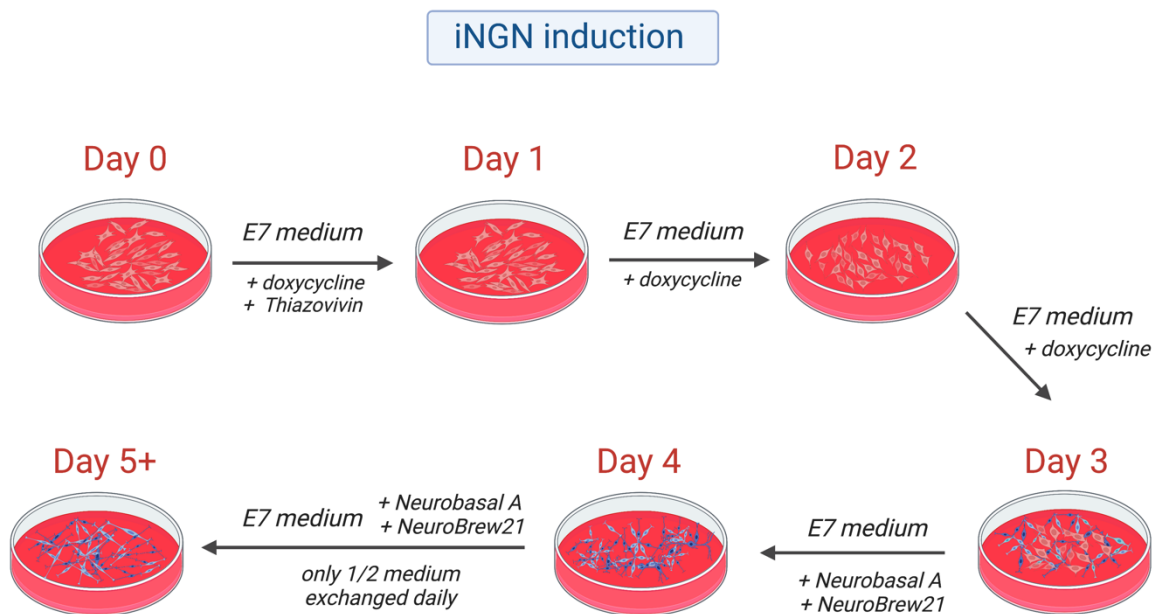
Figure 12. **The process of neural induction of iPS cells.** The stem cell-like cells when exposed to induction and growth factors develop into bipolar neuron-like iNGN cells within 4 days follows a specific protocol developed by Busskamp et al. The cells are induced with doxycycline and the medium content is gradually changed to adapt to neural-like cell specificity.

Plate coating procedure: coating medium can be prepared and stored at 4 °C for up to a month. Geltrex^{TM} (Gibco) should be thawed on ice, when reaches room temperature it gelatinizes instantly. Thawed Geltrex^{TM} should be added to the cold coating medium at a ratio of 1:1000. The medium mixture should be immediately transferred to the petri dishes, which can be then incubated for 90 minutes in 37 °C, 5% CO2. Afterwards, the volume should be topped-up with water as shown in the table below. The plates should be then transferred again into the incubator overnight. After 12 hours, the plates can be used for plating or and stored (covered with parafilm) at 4°C for up to 3 months.

Other laboratory techniques not described in the published studies

gDNA extraction

In order to extract genomic DNA (gDNA), the cells need to be harvested. The medium is removed and the cells are washed twice with PBS. Then, the lysis buffer is applied. The wells/plates are swirled until all cells detach and lyse. The lysate is transferred to Eppendorf tubes for further processing. The lysis buffer consists of RLT buffer (Qiagen) supplemented by the oxidation inhibitors - BHT (400 µM) and Desferal (400 µM) as well as β-ME (0.5 mM). BHT is dissolved first in 1/6 of the volume DMSO, then in two parts of water and vortexed until liquid becomes transparent. The volumes of the buffer: 1.6 mL per 3 cm diameter dish, 3.2 mL per 6 cm dish, 6.4 mL per 10 cm dish.

DNA is sheered using tissue lyser via bead milling process at 30Hz for one minute. The stainless beads are 5mm in diameter and are used one per microcentrifuge tube. Subsequently, the tubes are centrifuged for 10 minutes at 21 000 rpm. The lysate is then transferred to the columns from Quick gDNA extraction kit (Zymo Research). The flow-through was kept if RNA or protein was needed to be extracted from the sample. After the DNA bound column residues, the RNA digest was performed by incubating the RNAse digestion buffer on the columns for 15 mins at room temperature. RNAse digestion buffer was prepared with Genomic Lysis Buffer (Zymo Research), RNAse A at a ratio of 1:250 (Qiagen) and BHT (400 µM). and Desferal (400 µM).

After the digest, the columns were centrifuged for 2 minutes at 10 000 rpm. DNA pre-wash and wash buffers were applied according to the manufacturer's protocol. The samples were eluted with H_2O with BHT (40 µM), minimum volume of 25µL per column. The aim was to obtain at least 250 ng/µL gDNA concentration per sample, which would give enough material for Triple Quadrupole MS triplicate measurements.

RNA extraction was performed using Zymo Research Quick RNA extraction kit protocol.

CELL PROLIFERATION ANALYSIS

CellTrace Violet was used according to the Cell Proliferation Kit protocol to label and trace generations of proliferating cells via Fluorescence-Activated Cell Sorting (FACS). Cell trace reagent was mixed with cell suspension (1 µM final concentration) in PBS and incubated for 20 minutes at 37 °C protected from light in 15 mL falcon. Then, the mixture was centrifuged and the supernatant removed. Cells were washed twice with their dedicated medium and re-plated. The cells were stained for 24, 48 and 72 hours. CellTrace Violet covalently binds to the intracellular amines making fluorescent staining half-intense with every division round of the cells. The collected timepoints were measured with a lot of help and advice from Dr. Lisa Richter from BMC, LMU Department of Biology.

QPCR

All cDNA samples were prepared using iScript cDNA Synthesis Kit:

iScript cDNA Synthesis Kit	
Ingredient	Volume
RNA	Equal to 1 µg
iScript Buffer	4 µL
iScript Enzyme Mix	1 µL
H_2O	up to 20 µL total volume

The reaction was performed in a thermocycler as follows:

Reverse transcription		
Step	Temperature	Time
Priming	25 °C	5 min
Reverse Transcription	46 °C	20 min
Enzyme inactivation	90 °C	1 min
Hold at	4 °C	∞

Primers were resuspended at 100µM concentration. They were then diluted in a 1:4 ratio for the absorbance measurement on Nanodrop in UV-Vis Mode at 260 nm. The real concentrations were calculated using Beer's law:

$$c = \frac{A}{\varepsilon_{260}} \times 1000 \times 4$$

Primers efficiency check: the primers were diluted to 1,25 µM concentration each. Control untreated cDNA was used to prepare 50 ng/µL concentration as a starting point, six serial dilutions were prepared, 1:10 each. Efficiency was checked for all pairs of each gene tested on 96-well plate using qTower3 thermocycler using the following protocol:

Reverse transcription		
Step	Temperature	Time
Heat denaturation	95 °C	2 min
Denaturation	95°C	15 sec
Annealing	55 °C	15 sec
Elongation	72 °C	20 sec

40x

Melting curve was measured between 60 and 96 °C

Blank control: primer pair + H_2O and negative control: RNA + H_2O

The average *Ct* value from a technical triplicate was calculated. SLOPE function in MS Excel was used to calculate the slope of the log (primer concentrations) for each primer pair. The equation for primer efficiency was used:

$$Efficiency = (10^{\frac{-1}{slope}} - 1) \times 100$$

SOLUBLE NUCLEOSIDE POOL ISOLATION AND PURIFICATION

The soluble pool extraction protocol started with a cell harvest using typical cell passaging trypsinisation technique using TrypLE reagent, Gibco. The cells were counted with trypan blue via Countess II FL Automated Cell Counter, Life Technologies (ThermoFisher Scienitfic) and pelleted via centrifugation at 260 g for 3min. The pellet was resuspended with the ice-cold mixture of acetonitrile and water (1:1 ratio), 1mL per $1x10^6$ cells and gently vortexed. The cell suspension was incubated on ice for 15min and centrifuged at 10 000 g for 10min at 4°C. The supernatant containing the cytosolic fraction was collected and lyophilised, while the cell pellet was treated with the lysis buffer for subsequent gDNA isolation.

PURIFICATION WITH SUPEL-SELECT HBL 60MG COLUMNS:

A) Extraction of nucleosides:

1) Dissolve the lyophilized samples in H_2O pH = 4. 100 µL per 400 000 cells.

2) Equilibrate the columns with 1.5 mL MeOH, then 3 mL H_2O pH = 4

3) Run sample through the column

4) Wash the column with 4 mL H_2O pH = 4

5) Collect the washed aqueous phase containing mainly phosphates and lyophilize it for dephosphorylation

6) Dry the columns under vacuum in Schlenk flasks

7) Elute nucleosides with 1.5 mL MeOH/MeCN = 1:1

8) Lyophilise the samples

9) Elute each sample with 43 µL H_2O for MS measurements

B) Dephosphorylation using Antarctic Phosphatase

1) Dissolve the lyophilized aqueous phase using 16 µL H_2O

2) Add 2 µL Antarctic Phosphatase reaction buffer and mix gently

3) Add 2 µL Antarctic Phosphatase

4) Incubate at 37 $^{\circ}$C for 30 min

5) Inactivate the enzyme by heat shock at 80 $^{\circ}$C for 2 min

6) Cool down the sample on ice and purify using Super-Select HBL 60mg columns (using extraction of nucleosides - A)

Chemicals	Company
3,5-Di-tert-4-butylhydroxytoluene	Sigma Aldrich
6% Hydroxyethyl starch (Volulyte)	Sigma Aldrich
Agarose	Biozym
Apo-Transferrin	Sigma Aldrich
Benzonase	Sigma Aldrich
Bovine serum albumin	New England Biolabs
BSA	Sigma Aldrich
CellTrace TM Violet Cell Proliferation Kit	Thermo Fisher
CGP 77675	Sigma Aldrich
CHIR-99021	Sigma Aldrich
Deferoxamine mesylate	Sigma Aldrich
Degradase Plus	Zymo Research
Dimethylsulfoxide (DMSO)	Merck
Doxycycline	Sigma Aldrich
Dulbecco's Modified Eagle's Medium (DMEM)	Sigma Aldrich
Dulbecco's Phosphate Buffered Saline (PBS)	Sigma Aldrich
Ethanol absolut	Merck
Ethylenediaminetetraacetic acid (EDTA)	Sigma Aldrich
Fetal bovine serum heat-inactivated (FBS HI)	Gibco
Fetal bovine serum mESC-tested (FBS)	VWR
FGF-2 SI	MACS
Ham's F12	Sigma Aldrich
HEPES	Roth
Holo-transferrin human	Sigma Aldrich
Insulin human recombinant	Sigma Aldrich
iScript cDNA synthesis kit	Bio-Rad
Isopropanol	Sigma Aldrich
iTAQ universal SYBR® Green kit	Bio-Rad
L -Ascorbic acid 2-phosphate	Sigma Aldrich
L-Alanyl-L-Glutamine	Sigma Aldrich
Leukemia Inhibitory Factor (LIF)	Sigma Aldrich
Lipofectamine 3000	Invitrogen
McCoy's 5a	Sigma Aldrich
NAD+	Merck
NEB digestion enzyme mix	New England Biorad
NeuroBrew-21	MACS
Non-essential amino acids (NEAA)	Lonza
OptiMEM	Gibco
PD-0325901	Sigma Aldrich
Penicillin Streptomycin (Pen-Strep)	Gibco
peqGREEN	VWR
Progesterone	Sigma Aldrich
Putrescine 2HCl	Sigma Aldrich
Quick-gDNA extraction kit	Zymo Research
RLT-Buffer	Qiagen
RNaseA	Qiagen
Sodium chloride	Roth
Sodium selenite	Sigma Aldrich
TGF-Beta	MACS
Thiazovivin	MACS
TrypLETM Express Enzyme	Gibco
Trypsin	Invitrogen
β-Mercaptoethanol (β-ME)	Sigma Aldrich

sıRNA

esiRNA targeting mouse Sirt1   Sigma Aldrich

esiRNA targeting EGFP   Sigma Aldrich

PCR PRIMERS

Primer name	Sequence 5'-3'
mouse_Dnmt1_FW	CGTTCAGAGCAGATCGGAGA
mouse_Dnmt1_RV	CACTGGTCAACATCTGGGGT
mouse_Dnmt3a_FW	GAGCCTGAGTATGAGGATGGC
mouse_Dnmt3a_RV	CCACCAAGACACAATTCGGC
mouse_Dnmt3b_FW	TACACGCAGGACATGACAGG
mouse_Dnmt3b_RV	GTCCCATTGCTATGTCGGGT
mouse_Sirtuin1_FW	TACCTTGGAGCAGGTTGCAG
mouse_Sirtuin1_RV	GCACCGAGGAACTACCTGAT
mouse_Sirtuin2_FW	GTCCACTGGCCTCTATGCAA
mouse_Sirtuin2_RV	GATGGTTGGCTTGAACTGCC
mouse_Sirtuin6_RV	CTGGCGGTCATGTTTTGTGG
mouse_TDG_RV	GCCCATTTGAAGCAATGCCA
mouse_TDG_FW	GCGCAAGAGGACCGAAAGAA
mouse_GAPDH_FW	TGGAGAAACCTGCCAAGTATGA
mouse_GAPDH_RV	GTTGCTGTTGAAGTCGCAGG

CHAPTER 1. INVESTIGATION OF THE C-C BOND CLEAVAGE BY A DIRECT DEFORMYLATION MECHANISM

"*Analysis of an Active Deformylation Mechanism of 5-Formyl-deoxycytidine (fdC) in Stem Cells*" M. Sc. Alexander Schön*, <u>Ewelina Kaminska*</u>, M. Sc. Florian Schelter*, M. Sc. Eveliina Ponkkonen, M. Sc. Eva Korytiaková, Dr. Sarah Schiffers, Prof. Dr. Thomas Carell

*These authors equally contributed to this work

SUMMARY

Previous studies by Schröder et al. from 2016 already proved that 2'-fluorinated derivatives F-mdC F-hmdC, F-fdC, and F-cadC are resistant to TGD-mediated BER and are suitable substrates for DNA polymerases and TET enzymes.[225,226] Based on these findings, these compounds became good candidates for tracking DNA active demethylation. Indeed, Iwan et al. only a year later suggested that deformylation from fdC to dC occurs via C-C bond cleavage, as demonstrated in in vitro (in cellulo) studies using 2'-fluorinated fdC genomic incorporation.[197] The robustness of this experimental setup allowed for the establishment of isotope-dilution mass spectrometry protocol for the precise quantification of noncanonical DNA nucleosides.[186] Considering the previous findings indicating that 5fdC may be an epigenetically important DNA mark, we examined the mechanism of direct enzymatic deformylation. The first step of the hypothesised C-C bond cleavage involves a nucleophilic attack on the C-6 position of the base, which we blocked by replacing it with a nitrogen atom (see Figure 13). [198] We observed that the 6-aza-fdC compound effectively crosses the cellular membrane and becomes incorporated into genomic DNA.
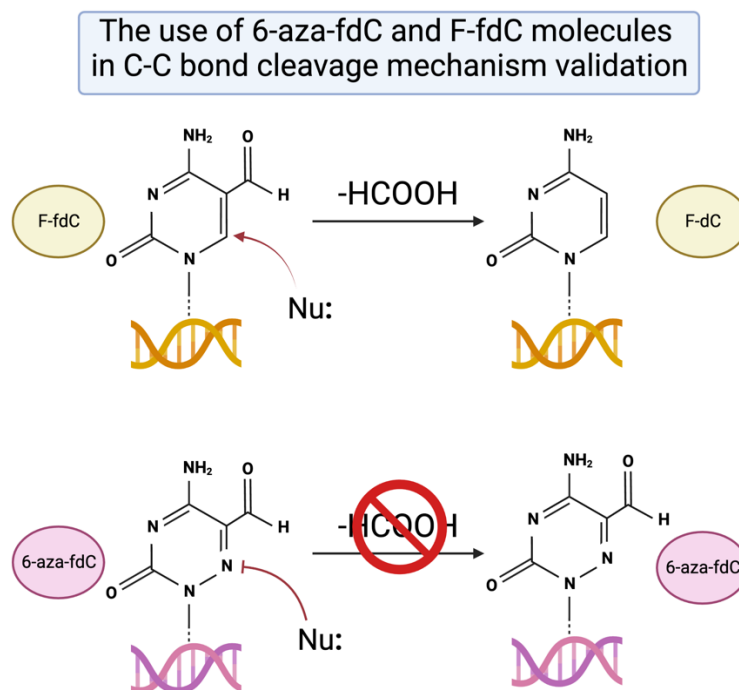
Figure 13. **The use of 6-aza-fdC and F-fdC molecules in C-C bond cleavage mechanism validation.** The aim of this study was to investigate the mechanism of C-C bond cleave by blocking the first step towards this process - the nucleophilic attack at the C-6 position. 6-aza-fdC appeared to be a suitable compound that is safe to use *in cellulo* on sensitive TDG-KO mESCs. In order to test the efficiency of the deformylation we performed a co-feeding of 6-azafdC and F-fdC nucleosides, from which only the latter has been found to deformylate. This finding is an convincing piece of evidence towards the idea of the existence of the deformylase enzyme, that provides the system with a helper nucleophile. *Created using BioRender tool.*

AUTHOR'S CONTRIBUTION

I cultured all the cell lines and performed the experiments involving cell feeding and material extraction for the analysis. I also optimised the cell-feeding protocols for all synthesised compounds (6-azadC oxidative derivatives) and tested their uptake/tolerance by the cells. I primed mESCs, harvested them and extracted their gDNA. I prepared samples for Triple Quadrupole LC-MS/MS measurements. I participated in manuscript writing and draft review process. The 6-aza-fdC compound was provided by Alexander Schön from the Carell research group.

LICENSE

Analysis of an Active Deformylation Mechanism of 5-Formyl-deoxycytidine (fdC) in Stem Cells

Alexander Schön[+], Ewelina Kaminska[+], Florian Schelter[+], Eveliina Ponkkonen,
Eva Korytiaková, Sarah Schiffers, and Thomas Carell*

Dedicated to Dr. Klaus Römer on the occasion of his 80th birthday

Abstract: *The removal of 5-methyl-deoxycytidine (mdC) from promoter elements is associated with reactivation of the silenced corresponding genes. It takes place through an active demethylation process involving the oxidation of mdC to 5-hydroxymethyl-deoxycytidine (hmdC) and further on to 5-formyl-deoxycytidine (fdC) and 5-carboxy-deoxycytidine (cadC) with the help of α-ketoglutarate-dependent Tet oxygenases. The next step can occur through the action of a glycosylase (TDG), which cleaves fdC out of the genome for replacement by dC. A second pathway is proposed to involve C−C bond cleavage that converts fdC directly into dC. A 6-aza-5-formyl-deoxycytidine (a-fdC) probe molecule was synthesized and fed to various somatic cell lines and induced mouse embryonic stem cells, together with a 2′-fluorinated fdC analogue (F-fdC). While deformylation of F-fdC was clearly observed in vivo, it did not occur with a-fdC, thus suggesting that the C−C bond-cleaving deformylation is initiated by nucleophilic activation.*

$\mathbf{T}$he nucleobase modification 5-formyl-deoxycytidine (fdC, **1**) is found in stem cells during early development and in the brain.[1–5] These tissues are particularly rich in 5-hydroxymethyl-deoxycytidine (hmdC) from which fdC (**1**) is produced.[6,7] The formation of hmdC and fdC requires oxidation reactions that are performed by α-ketoglutarate-dependent Tet enzymes, with 5-methyl-deoxycytidine (mdC) being the initial starting molecule.[8–10] This cascade of oxidation reactions is a part of an active demethylation process, in which mdC as a silencer of transcription is replaced by unmodified dC.[11] The central molecule that is removed seems to be

fdC.[12,13] It can be cleaved out of the genome by a dedicated DNA glycosylase, which creates an abasic site that is further processed, leading to the insertion of an unmodified dC.[14] Because abasic sites are harmful DNA-repair intermediates that can cause genome instability, it was suggested early on that fdC might be directly deformylated to dC by C−C bond cleavage.[15,16] Evidence for the existence of such a direct deformylation process was recently reported.[17] Model studies showed that direct deformylation of fdC and potentially also decarboxylation of 5-carboxy-deoxycytidine (cadC) are indeed possible.[15] Nevertheless, it requires activation of the nucleobase by a nucleophilic addition to the C6 position. For fdC, an additional hydrate formation on the formyl group seems to be necessary, as depicted in Figure 1A. Although activation with a helper nucleophile is well known as the central mechanistic process during the methylation of dC to mdC by DNA methyltransferases (Dnmts),[18,19] it remains to be confirmed whether such activation occurs in vivo as well.

[*] M. Sc. A. Schön,[+] E. Kaminska,[+] M. Sc. F. Schelter,[+]
M. Sc. E. Ponkkonen, M. Sc. E. Korytiaková, Dr. S. Schiffers,
Prof. Dr. T. Carell
Department of Chemistry, Ludwig-Maximilians Universität München
Butenandtstr. 5–13, 81377 München (Germany)
E-mail: thomas.carell@lmu.de
Homepage: https://www.carellgroup.de

[+] These authors contributed equally to this work.

🖥 Supporting information and the ORCID identification number(s) for
ⓘD the author(s) of this article can be found under:
https://doi.org/10.1002/anie.202000414.

Figure 1. A) The mdC removal pathways that involve oxidation to hmdC, fdC, and cadC followed by either base-excision repair (magenta) or C−C bond cleavage (blue). B) Structures of fdC (**1**) and the two probe molecules **2** and **3** used for this study.

In this work, we investigated this hypothesis with two probe molecules, 2′-fluorinated-fdC (F-fdC, **2**) and 6-aza-fdC (a-fdC, **3**). The two compounds were simultaneously fed to different cell types, including primed stem cells. This led to random incorporation of these bases at the "C" sites in the respective genomes. Furthermore, it led to the presence of F-fdC and a-fdC not only at CpG sites. Ultrasensitive UHPLC-QQQ-MS[2] was subsequently used to interrogate the chemical processes that occur at F-fdC and a-fdC in the genomes. The data show that while F-fdC is efficiently deformylated, this does not occur for a-fdC. The only difference between the two nucleobases is the presence of an in-ring nitrogen atom (6-aza atom), which features a lone pair that prohibits nucleophilic addition. These results thus provide strong evidence that nucleophilic activation is the central governing mechanistic event that is required for C−C bond cleavage in vivo.

The fluorinated nucleoside F-fdC (**2**) was recently introduced by us as a deformylation probe.[17] Compound **2** is an antimetabolite that is effectively incorporated into the genomes of growing cells. The 2′-fluoro group is required to block all types of glycosylases, so that base-excision repair is efficiently inhibited. This ensures high levels of F-fdC (**2**) in the genome, as required to observe potential deformylation processes.

The synthesis of the novel nucleoside a-fdC (**3**) is depicted in Scheme 1. The synthesis was started with bromo pyruvic acid (**4**), which we first converted into the semicarbazone **5**, followed by conversion into the acid chloride, subsequent cyclization, and hydrolysis to give hydroxymethylated 6-aza-uracil (**6**).[20] Vorbrüggen nucleosidation with Hoffers' chlorosugar subsequently provided the nucleoside **7** as a mixture of the α- and β-anomers, which could be separated by recrystal-

lization. Next, we acetyl-protected the hydroxymethyl group to give **8**, and then used a standard procedure to convert the U base **8** into the C-derived base **9** by amination of the 4-triazole intermediate with ammonium hydroxide. This led to the concomitant cleavage of the acetyl protecting group. Dess–Martin oxidation of **9** to **10** and final removal of the toluoyl groups furnished the 6-aza-5-formyl-deoxycytidine nucleoside (a-fdC) **3** in a good total yield of 22% with respect to **6** (Supporting Information).

Compound **3** features a nitrogen atom instead of a carbon atom at the 6-position, which possesses a lone pair that blocks any nucleophilic addition to this position. Compound **3** is consequently a perfect model system to investigate whether such a nucleophilic activation is required for the deformylation, as mechanistically postulated (Figure 1 A).

The nucleosides **2** and **3** were subsequently added at a concentration of 350 μM to the media of Neuro-2a, RBL-2H3, CHO-K1 cells for 72 hours (see the Supporting Information). During this time, the nucleosides are converted in vivo into the corresponding triphosphates and then incorporated into the genome of the dividing cells. Initial studies in which we fed the nucleosides individually allowed us to determine that neither compound decreases cell viability up to a concentration of 400 μM, thus the experiments were conducted below the toxicity level. In addition, we tested **2** and **3** at 350 μM on E14 TDG +/− and −/− mouse embryonic stem cells (mESCs) under a three-day priming process with C/R media. This system allowed us to exclude the BER pathway, leading to a detectable and quantifiable accumulation of natural fdC (see the Supporting Information). After three days, the cells were harvested and lysed, and the genomic DNA was extracted using an optimised protocol (see the Supporting Information). This was followed by an enzymatic digestion of gDNA to single nucleosides and analysed according to a method that we reported recently in detail.[21] The obtained nucleoside mixture containing mostly the canonical nucleosides dA, dC, dG, and dT, plus the non-canonical nucleosides mdC, hmdC, and fdC, as well as the incorporated molecules F-fdC and a-fdC and their potential downstream products (F-dC, F-mdC, a-dC, a-mdC). Nucleosides were separated by ultra-HPL chromatography and characterized by coupling of the UHPLC system to a triple-quadrupole mass spectrometer. For exact quantification of the nucleosides by isotope dilution, isotopically labelled standards of F-fdC and of the product F-dC were spiked into the analysis mixture as internal standards (see the Supporting Information). To enable exact quantification, calibration curves using these standards were determined (see the Supporting Information). Quantification was performed in the linear region.

During the analysis, we noted that an unusually low amount of a-fdC (**3**) was detected because it showed a broad elution profile with very low intensity (Figure 2B). All attempts to sharpen the elution profile in order to gain sensitivity failed. NMR analysis of compound **3** showed the reason for broad elution profile (see the Supporting Information). Due to the additional electron-withdrawing in-ring nitrogen atom, compound **3** exists partially as its hydrate in aqueous solution (20%, see the Supporting Information).
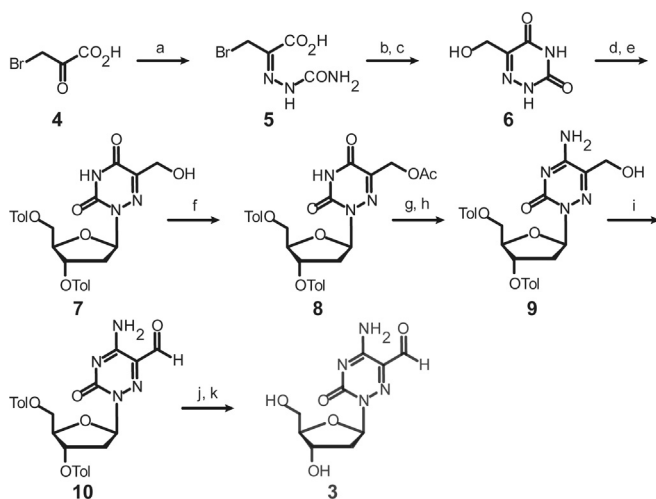
Scheme 1. Synthesis of the probe molecule a-fdC (**3**). a) semicarbazide·HCl, NaOAc, HOAc, H_2O, 0°C to r.t., 2.5 h, 49%. b) pyridine, SOCl$_2$, 80°C, 75 min. c) H_2O, 110°C, 17 h, 74% over 2 steps. d) TMSCl, HMDS, 135°C, 75 min, *then* e) Hoffer's chlorosugar, CHCl$_3$, r.t., 17 h, 56% over 2 steps. f) Ac$_2$O, pyridine, r.t., 22 h, 96%. g) 1,2,4-triazole, POCl$_3$, NEt$_3$, MeCN, 0°C to r.t., 18 h, then h) NH$_4$OH, 1,4-dioxane, 40°C, 5 h, 84%. i) Dess–Martin periodinane, CH$_2$Cl$_2$, −15°C to r.t., 1 h, 89%. j) NaOMe, MeOH, benzene, r.t., 1.5 h, then k) reversed-phase HPLC, 54%.
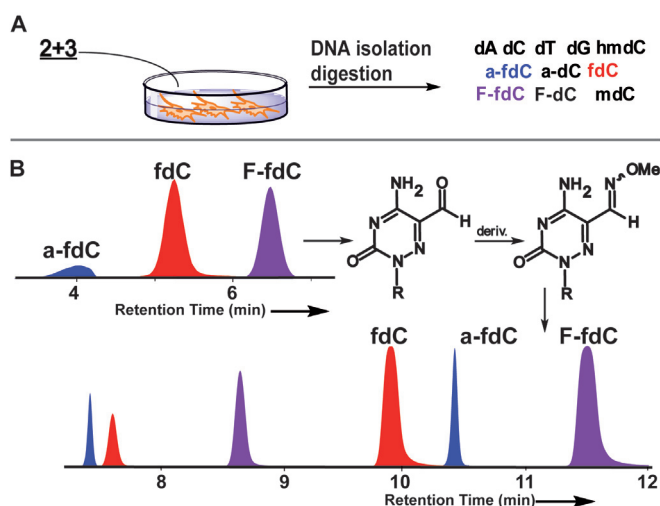
Figure 2. A) Overview of the experimental steps with the feeding and analysis. B) Analysis scheme and the reaction of a-fdC with methoxyamine to block hydrate formation and of a typical UHPL-chromatogram before (C-8 column) and after derivatization (C-18 column) for exact quantification. Peak splitting is due isomerization (blue peaks: a-fdC, red peaks: fdC, and purple peaks: F-fdC).

Figure 3. A) Deformylation data for F-fdC in different cell types, showing that F-fdC is deformylated in very different cells. Deformylation rate was calculated by the F-dC + F-mdC/dN per F-fdC/dN, then the values were normalized to the cell line with the lowest deformylation level (CHO-K1 = 1). B) The deformylation of F-fdC/dG and a-fdC/dG, showing the induced differences due to C6-carbon-to-nitrogen exchange. C) The bisulfite data show that the deformylation of fdC and F-fdC is comparable, thus showing that the 2′-F substitution has only a small accelerating effect, whereas the reaction of a-fdC could not be detected.

Although the ease of hydrate formation may foster deformylation, the hydrate/carbonyl equilibrium makes efficient detection of compound **3** basically impossible. In order to circumvent the problem, we started to derivatize a-fdC (**3**) before analysis with methoxyamine. Addition of CH_3ONH_2 (150 mM) to the digestion solution indeed provided the methoxyoxime of a-fdC in quantitative yield after just 15 min at 25°C and pH 10. The naturally present fdC (**1**) and the probe molecule F-fdC (**2**), however, react as well, but unfortunately not quantitatively. To reduce impurities during MS measurements, we decided against using a catalyst for oxime formation. We therefore decided to analyse the digested DNA in two batches. The first one contained the digested untreated DNA to quantify all bases other than a-fdC. In the second batch, we treated the digested DNA with methoxyamine for a-fdC quantification. For quantification of the derivatized a-fdC, we constructed an external calibration curve (see the Supporting Information).

With this method in hand, we next quantified all nucleosides present in the genome of the cells treated with a mixture of **2** + **3**.

Figure 3 A shows that we indeed detected the fluorinated F-dC (**2**), thus confirming very efficient deformylation activity. We tested different cell types and found different levels of deformylation activity. But in all cases, the conversion of F-fdC into F-dC was clearly detectable. Most interesting is that we observed the highest deformylation activity in cells associated with neuronal properties. This is in line with neurons featuring the highest levels of hmdC and fdC. In contrast, Figure 3 B shows that for a-fdC (**3**), we were unable to detect any formation of the deformylated compound a-dC despite the high propensity of **3** to exist in the hydrated form, which is one prerequisite for efficient C−C bond cleavage. This result suggests that the ability to react

with a nucleophile at the 6-position is also required in vivo for efficient deformylation.

In order to substantiate this result, we next performed in vitro studies with bisulfite. Bisulfite is a strong nucleophile that has been reported to cause deformylation of fdC by first attacking the C6 position, followed by conversion of the C5,C6-saturated fdC adduct into the bisulfite adduct, which then undergoes deformylation.[22] The deformylated product dC is then further converted into dU by the well-known bisulfite-induced deamination reaction of dC (see the Supporting Information). Indeed, when we reacted fdC with bisulfite, we observed efficient deformylation and deamination to dU. We then studied to what extent the reaction is influenced by the 2′-F atom present in F-fdC, in order to estimate whether the in vivo deformylation could be just the result of the 2′-F atom. Treatment of F-fdC with bisulfite also led to deformylation and deamination to F-dU, and indeed the reaction is a little faster compared to fdC (see Figure 3 C). Although the difference is measurable, it is in total rather small. With these data in hand, we can conclude that we may overestimate the amount of deformylation that can occur with fdC lacking the 2′-F atom. We can certainly exclude that deformylation in vivo occurs only with F-fdC. It is unfortunate that we are unable to measure the direct deformylation of fdC because of the presence of efficient BER processes. A TDG −/− cell line showed a huge increase in fdC compared to the TDG +/−, whereas a-fdC and F-fdC stay constant, thus showing that these compounds are indeed not repaired by the

TDG protein (see the Supporting Information). The bisulfite studies, however, show that the F-fdC compound is not a perfect but sufficient reporter of this C−C bond cleavage. Treatment of a-fdC (**3**) with bisulfite did not provide the deformylated product a-dC under any circumstances, showing that the inability to react with a nucleophile at the 6-position totally blocks the C−C bond cleavage. We can therefore conclude that the deformylation of fdC during active demethylation requires oxidation of mdC to fdC. fdC can undergo a direct C−C bond cleavage to dC, but this reaction requires a helper nucleophile to attack the C6-position, which is blocked in the case of a-fdC by the lone pair introduced by the C6-carbon-to-nitrogen exchange. While the chemistry that allows the transformation of fdC into dC is now elucidated, we next need to find the nucleophiles that perform the reaction in vivo.

Conflict of interest

The authors declare no conflict of interest.

[1] 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.

[2] S. Liu, J. Wang, Y. Su, C. Guerrero, Y. Zeng, D. Mitra, P. J. Brooks, D. E. Fisher, H. Song, Y. Wang, *Nucleic Acids Res.* **2013**, *41*, 6421–6429.

[3] C. X. Song, K. 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.

[4] T. Pfaffeneder, B. Hackner, M. Truss, M. Münzel, M. Müller, C. A. Deiml, C. Hagemeier, T. Carell, *Angew. Chem. Int. Ed.* **2011**, *50*, 7008–7012; *Angew. Chem.* **2011**, *123*, 7146–7150.

[5] M. Bachman, S. Uribe-Lewis, X. Yang, H. E. Burgess, M. Iurlaro, W. Reik, A. Murrell, S. Balasubramanian, *Nat. Chem. Biol.* **2015**, *11*, 555–557.

[6] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–930.

[7] D. Q. Shi, I. Ali, J. Tang, W. C. Yang, *Front. Genet.* **2017**, *8*, 100.

[8] M. Münzel, D. Globisch, T. Carell, *Angew. Chem. Int. Ed.* **2011**, *50*, 6460–6468; *Angew. Chem.* **2011**, *123*, 6588–6596.

[9] C. G. Spruijt, F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. T. C. Jansen, C. Bauer, M. Münzel, M. Wagner, M. Müller, F. Khan, H. C. Eberl, A. Mensinga, A. B. Brinkman, K. Lephikov, U. Müller, J. Walter, R. Boelens, H. Van Ingen, H. Leonhardt, T. Carell, M. Vermeulen, *Cell* **2013**, *152*, 1146–1159.

[10] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, *Science* **2011**, *333*, 1300–1303.

[11] F. Neri, D. Incarnato, A. Krepelova, S. Rapelli, F. Anselmi, C. Parlato, C. Medana, F. DalBello, S. Oliviero, *Cell Rep.* **2015**, *10*, 674–683.

[12] M. Su, A. Kirchner, S. Stazzoni, M. Müller, M. Wagner, A. Schröder, T. Carell, *Angew. Chem. Int. Ed.* **2016**, *55*, 11797–11800; *Angew. Chem.* **2016**, *128*, 11974–11978.

[13] T. Fu, L. Liu, Q. L. Yang, Y. Wang, P. Xu, L. Zhang, S. Liu, Q. Dai, Q. Ji, G. L. Xu, C. He, C. Luo, L. Zhang, *Chem. Sci.* **2019**, *10*, 7407–7417.

[14] R. Rahimoff, O. Kosmatchev, A. Kirchner, T. Pfaffeneder, F. Spada, V. Brantl, M. Müller, T. Carell, *J. Am. Chem. Soc.* **2017**, *139*, 10359–10364.

[15] S. Schiesser, T. Pfaffeneder, K. Sadeghian, B. Hackner, B. Steigenberger, A. S. Schröder, J. Steinbacher, G. Kashiwazaki, G. Höfner, K. T. Wanner, C. Ochsenfeld, T. Carell, *J. Am. Chem. Soc.* **2013**, *135*, 14593–14599.

[16] 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.

[17] K. Iwan, R. Rahimoff, A. Kirchner, F. Spada, A. S. Schröder, O. Kosmatchev, S. Ferizaj, J. Steinbacher, E. Parsa, M. Müller, T. Carell, *Nat. Chem. Biol.* **2018**, *14*, 72–78.

[18] A. Jeltsch, *ChemBioChem* **2002**, *3*, 274–293.

[19] Q. Du, Z. Wang, V. L. Schramm, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 2916–2921.

[20] I. V. Alekseeva, A. S. Shalamai, V. S. Shalamai, V. P. Chernetski, *Ukr. Khim. Zh. (Ukr. Ed.)* **1976**, *42*, 398–401.

[21] F. R. Traube, S. Schiffers, K. Iwan, S. Kellner, F. Spada, M. Müller, T. Carell, *Nat. Protoc.* **2019**, *14*, 283–312.

[22] E. Kriukienė, Z. Liutkevičiūtėa, S. Klimašauskas, *Chem. Soc. Rev.* **2012**, *41*, 6916–6930.

CHAPTER 2. THE INVESTIGATION OF THE DEFORMYLATION RATES AND CIRCUMSTANCES IN WHICH IT OCCURS WITHIN DIFFERENT CELL LINES AND CELL TYPES

"Deformylation of 5-Formylcytidine in Different Cell Types" M. Sc. Eva Korytiaková*, <u>Ewelina Kamińska</u>*, Dr. Markus Müller, Prof. Dr. Thomas Carell.

*These authors contributed equally to this work.

SUMMARY

The initial mechanistic studies on stem cells confirmed the necessity of C-C bond cleave for deformylation to occur.[200] In this study, we aimed to investigate deformylation in different cell types – stem cells, somatic cells (including cancer cells) and human iPS cells, across a wide range of cell lines. The goal was to identify potential regulatory mechanisms or enzymes involved in deformylation and determine if there are any differences in deformylation and determine if there are any differences in deformylation rates among different cell lines. We used the well-established and safe probe molecule, 2'-fluorinated fdC, for our experiments. Interestingly, we observed varying deformylation rates in different cell lines, with somatic and cancer cells exhibiting higher deformylation activity compared to stem cells. Importantly, this deformylation activity was independent of DNA replication. Additionally, our findings suggest that both TDG-mediated and alternative pathways (as compared in Figure 14) may be utilised under different circumstances, depending on genomic sequence-specific localisation and cellular energy expenditure during different stages of the cell cycle. It is also possible that ceratin epigenetically important 5-fdCs, such as those forming Schiff-bases, may not be accessible to TDG and other BER-associated protein complexes.[65,75]
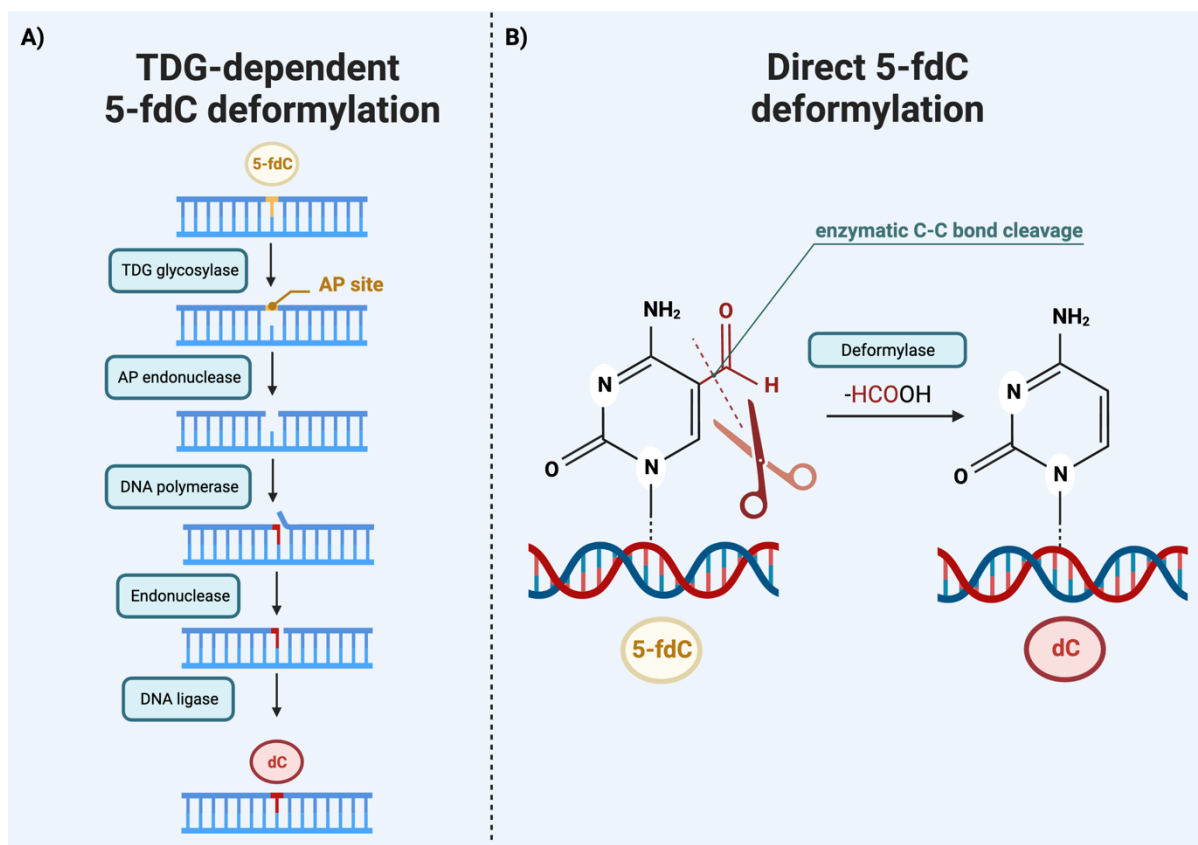
Figure 14. **The comparison between two DNA active demethylation pathways.** Both at the start require 5-mdC and 5-hmdC oxidations by TET family dioxygenases. However, they end by different deformylation mechanisms. **A)** represents the well-known TDG-mediated DNA repair process, in which 5-fdC is recognised as DNA lesion by the DNA glycosylase, which excises the base leaving the abasic site, later removed by AP endonuclease. Later, the DNA polymerase replaces the missing nucleobase that is sealed by the cooperative actions of the endonuclease and DNA ligase. This process requires a lot of energy needed for catalysis of at least 5 different enzymes. Moreover, the Base Excision Repair is associated with a risk of double strand breaks that are not only mutation-prone but also endanger the DNA stability. **B)** represents the direct enzymatic C-C bond cleavage deformylation, without the need for excision of the whole base. The proposed mechanism is simpler and less energetically costly. The unknown deformylase enzyme or complex, based on previous publication described in chapter 1 should provide a helper nucleophile and catalyse the conversion from 5-fdC to 5-dC. *Created using BioRender tool.*

AUTHOR'S CONTRIBUTION

I cultured all cell lines included in this study and preformed the experiments involving cells. I optimised cell feeding protocols for pulse-chase study. I primed mESCs and induced neuronal differentiation of iPS cells. I extracted gDNA, RNA and soluble nucleoside pool samples and prepared them to be used on Triple Quadrupole LC-MS/MS. Together with Eva Korytiaková we analysed the MS data. I looked for correlations between cell divisions and deformylation rates. The final conclusions were drawn together with Prof Thomas Carell. Jointly

with Eva Korytiaková we made the figures for the following manuscript. I participated in manuscript writing and draft review processes.

LICENSE

Nucleosides

Deformylation of 5-Formylcytidine in Different Cell Types

Eva Korytiaková[+], Ewelina Kamińska[+], Markus Müller, and Thomas Carell*

Abstract: Epigenetic programming of cells requires methyl-ation of deoxycytidines (dC) to 5-methyl-dC (mdC) followed by oxidation to 5-hydroxymethyl-dC (hmdC), 5-formyl-dC (fdC), and 5-carboxy-dC (cadC). Subsequent transformation of fdC and cadC back to dC by various pathways establishes a chemical intra-genetic control circle. One of the discussed pathways involves the Tdg-independent deformylation of fdC directly to dC. Here we report the synthesis of a fluorinated fdC feeding probe (F-fdC) to study direct deformylation to F-dC. The synthesis was performed along a novel pathway that circumvents any F-dC as a reaction intermediate to avoid contamination interference. Feeding of F-fdC and observation of F-dC formation in vivo allowed us to gain insights into the Tdg-independent removal process. While deformylation was shown to occur in stem cells, we here provide data that prove deformylation also in different somatic cell types. We also investigated active demethylation in a non-dividing neuro-genin-inducible system of iPS cells that differentiate into bipolar neurons.

Methylation of dC to 5-methyl-dC (mdC) by Dnmt enzymes in higher vertebrates is a means to control the transcriptional activity of genes.[1–4] Reversing the methylation requires oxidation of mdC to hmdC, fdC, and cadC.[5–9] These oxidized bases either block maintenance methylation (hmdC), which leads to global replication-dependent demethylation, or they are removed as nucleosides by the action of the thymine DNA glycosylase (Tdg) repair enzyme (fdC and cadC).[10–16] Alternatively, a direct deformylation and decarboxylation of fdC and cadC to dC may occur.[17–19] The advantage of a direct mechanism, in contrast to the Tdg-repair process, is that it proceeds without the formation of DNA strand breaks.[20,21] These are known to cause genome instability.[22] We could recently show that fdC is deformylated to dC in stem cells.[17] We have also shown that the

nucleophilic attack at the C6 position of fdC is necessary to cleave the C−C bond in a deformylation reaction.[18] Today, it is however still unclear to what extent the deformylation reaction occurs in various cell types. It is also unknown whether this process is influenced by replication. In order to investigate these questions, we performed a series of metabolic feeding studies using a 2′-fluorinated fdC derivative obtained by a synthetic pathway that circumvents F-dC as an intermediate. This probe molecule was used to label genomes in different somatic cell types originating from healthy and cancer tissue. For further investigation of active demethylation upon cell differentiation and its dependence on cell replication, additional studies in an iPS cell line system were established.

In order to trace the incorporated fdC probe as well as its deformylation product and to recognize it in front of the background of the abundant genomic dC, a chemical modification of the probe molecule is needed. This can be achieved by incorporating stable isotope labels (^{13}C, ^{15}N, ^{2}H) into the fdC ribose or the fdC heterocycle (not the cleaved-off formyl group). Another elegant way to achieve the labeling of the nucleoside is to tag the ribose with a 2′-F atom, which has the additional advantage that the base of this modified nucleoside cannot be removed by base excision repair.[23,24] In addition, the 2′-fluorinated dC-derivatives are recognized by Dnmt and Tet-enzymes, which means that they can be epigenetically modified (methylated, oxidized, and demethylated). Thus, a feeding probe such as 2′-F-fdC (**1**) is an ideal tool molecule to study active demethylation of mdC. We proved this principle in an in vivo cell system using Tdg knockout mESCs and the respective wildtype line. While the natural genomic fdC levels increase dramatically in the absence of Tdg, the 2′-fluorinated nucleoside values stay the same in wt and KO (Figure SI-6) cells, showing that Tdg is unable to remove 2′-F-fdC as expected. We also investigated the digestion efficiency of DNA strands containing 2′-fluorinated nucleosides and found only a small difference compared to non-fluorinated nucleosides under our conditions.[23,24]

Since 2′-fluorinated nucleosides do not undergo Tdg-mediated repair, feeding of 2′-F-fdC (**1**) to different cell lines gives higher levels of incorporation into the respective genomes (Figure 1A). Upon deformylation of 2′-F-fdC (**1**), 2′-F-dC (**2**) is formed. This compound can be easily distinguished from genomic dC using an extremely sensitive, non-high resolution UHPLC-MS/MS system (Figure 1B) because of the different retention times, the difference in molecular weight (+18 amu), and a specific fragmentation pattern.

For feeding experiments, it is essential to synthesize 2′-F-fdC (**1**) via a pathway that avoids 2′-F-dC (**2**) as a synthesis intermediate. Even the smallest contamination of 2′-F-fdC (**1**) with 2′-F-dC (**2**) can lead to false-positive results. 2′-F-dC (**2**) is taken up by cells as well and incorporated into the genome

[*] M. Sc. E. Korytiaková,[+] E. Kamińska,[+] Dr. M. Müller,
Prof. Dr. T. Carell
Department of Chemistry
Ludwig-Maximilians-Universität München
Butenandtstrasse 5–13, 81377 Munich (Germany)
E-mail: Thomas.Carell@lmu.de
Homepage: http://www.carellgroup.de

[+] These authors contributed equally to this work.

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under:
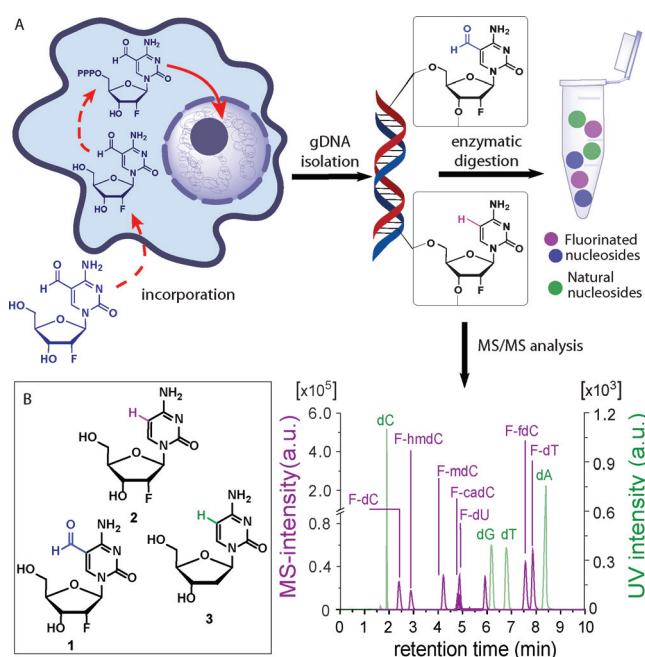https://doi.org/10.1002/anie.202107089.

Figure 1. A) Experimental setup and intra-genomic 2′-F-fdC deformylation reaction. The nucleoside is fed to the cells, where it undergoes subsequent phosphorylation in the cytosol, followed by incorporation into gDNA. The extracted gDNA contains the incorporated 2′-F-fdC as well as the possible deformylation product 2′-F-dC. MS/MS analysis was performed after enzymatic digestion of the gDNA. UHPLC-MS/MS chromatogram of all nucleosides present in gDNA. Green peaks show the UV intensity of the compounds, while purple signals represent MS-intensity values. B) The structures of the probe molecules: 2′-F-fdC (**1**), 2′-F-dC (**2**), and natural dC (**3**).

far more efficiently compared to 2′-F-fdC (**1**).[17] This leads to a strong enrichment of any 2′-F-dC (**2**) impurity and hence potentially to its accumulation during cell division. Therefore, we developed a novel synthesis pathway towards 2′-F-fdC (**1**) exclusively via 2′-F-dU intermediates thereby excluding undesired cytidine-based contaminations (Scheme 1). The synthesis starts with commercially available 2′-F-dU (**4**), which was first TBS-protected (TBS-Cl, imidazole), followed by regioselective iodination at C5 to **5** with iodine and $(NH_4)_2[Ce(NO_3)_6]$ as the oxidant. A subsequent Pd-catalyzed Stille-type formylation (CO, HSnBu₃) furnished the 2′-F-fdU compound **6**. Acetal protection of the formyl group (propane-diol, TSA) was followed by amination at C4 to generate the acetal-protected 2′-F-fdC derivative **7**. Full deprotection of **7** with 1 M HCl followed by treatment with HF/pyridine to cleave the TBS protection groups furnished the 2′-F-dC-free target compound **1**. Subsequent HPLC purification provided 2′-F-fdC (**1**) in a highly pure form. In order to substantiate that the prepared 2′-F-fdC (**1**) is 2′-F-dC (**2**)-free, we performed a highly sensitive UHPLC-MS/MS experiment (Scheme 1, inset).

The analysis of the synthetized 2′-F-fdC (**1**) was performed using the highest-sensitive triple quadrupole MS detector. Under conditions previously developed in our group, it is able to detect **2** in the low fmol range (Table SI-1).[17] The analysis of the synthesized 2′-F-fdC (**1**) provided no signal between 2.1 and 2.8 min (retention time of potential 2′-
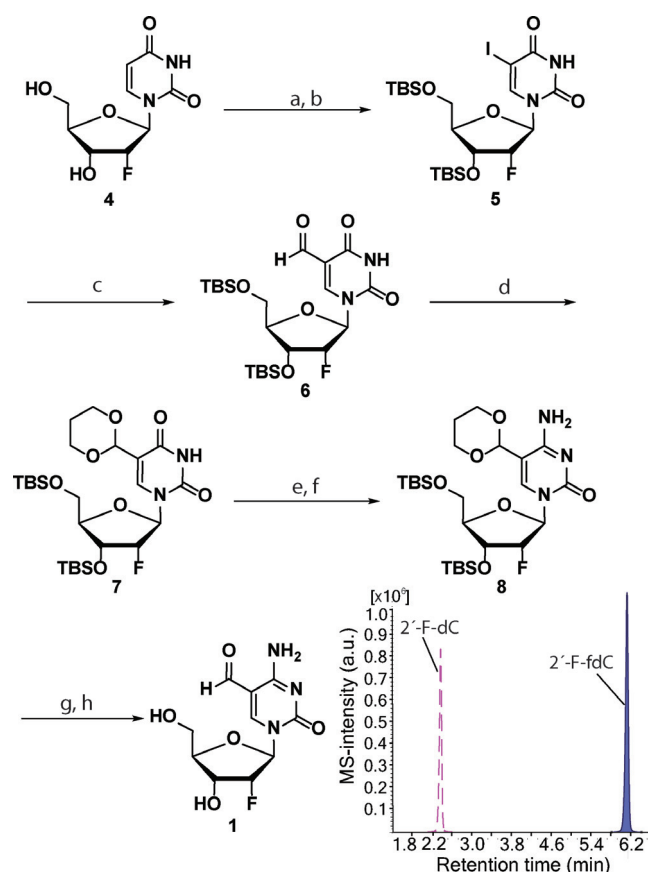
Scheme 1. Synthesis of 2′-F-fdC (**1**) probe molecule and UHPLC experimental data showing the high purity of **1**. a) TBSCl, imidazole, pyridine, rt, 16 h, 91%. b) I₂, $(NH_4)_2[Ce(NO_3)_6]$, MeCN, 60 °C, 1 h, 81%. c) HSnBu₃, Pd₂(dba)₃·CHCl₃, PPh₃, 3.5 bar CO, 60 °C, 18 h, 85%. d) 1,3-propanediol, CH(OEt)₃, p-TSA, CH₂Cl₂, rt, 60%. e) NaH, 2,4,6-triisopropylbenzolsulfonylchloride, THF, 0 °C→rt, 16 h, 60%. f) NH₄OH, 1,4-dioxane, rt, 16 h, 68%. g) 1 M HCl, H₂O/MeCN, rt, 18 h. h) HF/pyridine, EtOAc, rt, 16 h, reversed-phase HPLC 70%.

F-dC 2.3 min), which confirmed the high purity of the synthesized compound. Most importantly, this new synthesis route fully excludes any contamination of **1** with 2′-FdC (**2**).

To investigate the deformylation rates in various somatic cell lines and to compare them to mouse embryonic stem cells (mESCs) that we had studied before, respective cultures were grown in the presence of 2′-F-fdC (**1**) at a final concentration of 350 μM (Figure 2). The treatment continued for 3 days in order to allow all cell lines, having different doubling rates, to incorporate the fed compound. The cells were subsequently harvested, DNA was isolated and fully digested using an optimized protocol for the chemically modified nucleosides (see SI).[25] The obtained nucleoside mixture was subsequently analyzed by UHPLC-MS/MS. Quantification of the 2′-F-fdC (**1**) and 2′-F-dC (**2**) levels as well as of the remethylated species 2′-F-mdC (**9**) was performed using the previously published isotope dilution mass spectrometry technique (Figure SI-9).[17,26] First, we noted that the 2′-F-fdC probe molecule was incorporated into all tested cells and detectable at levels between 2.1×10^{-7} and 5.2×10^{-7} molecules per dN (Figure SI-1).
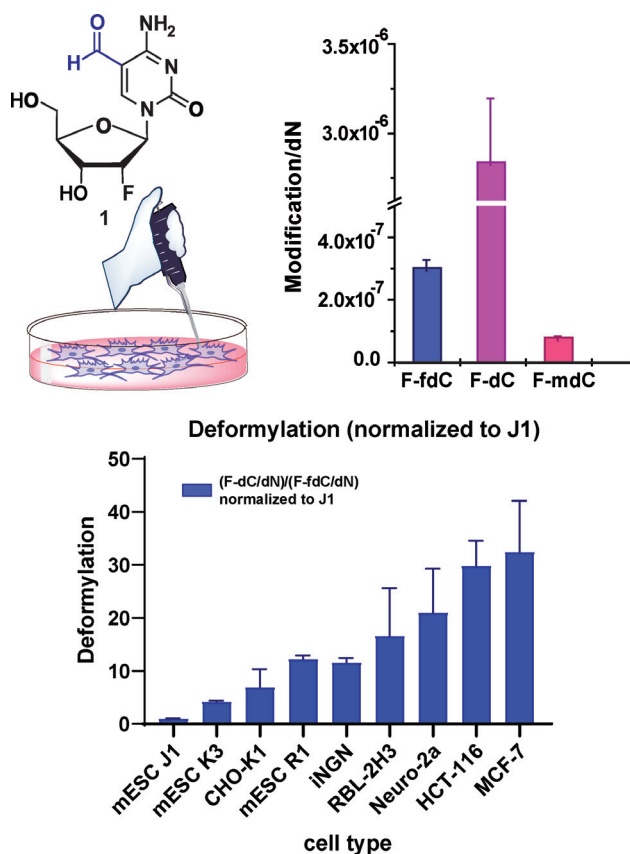
Figure 2. Feeding data of 2′-F-fdC (**1**). The right upper part shows the graph with the amounts of fluorinated nucleosides in gDNA after feeding 2′-F-fdC (**1**) to the J1 cell line. The bottom graph shows the different deformylation efficiencies in different cell lines normalized to the lowest deformylating cell line J1. The error bars represent propagated error based on the standard deviation of all biological triplicates.

As noted earlier, beside the fed compound **1** we also found 2′-F-dC (**2**) and, importantly, 2′-F-mdC (**9**) in all cells (Figure 2).[17] We conclude that the deformylation occurred in all investigated cells and that the newly formed 2′-F-dC (**2**) was subsequently re-methylated by the action of Dnmt enzymes, forming 2′-F-mdC (**9**, Figure 2). This indicates the presence of the chemical genomic cycle dC→mdC→oxdC→ dC→mdC in all investigated cell types. Unfortunately, we are not yet able to show the methylation of **2** in the CpG context, as the incorporation of **1** is assumably random. However, the level of methylation of 2′-F-dC (**2**) to 2′-F-mdC (**9**) is with 2–5 % within the expected range (Figure SI-8B).[27]

Investigation of the deformylation reaction (turnover of 2′-F-fdC (**1**) to 2′-F-dC (**2**)) in different cell lines showed a major variation of this activity (Figure 2). Most interesting is the observation that the previously studied stem cells J1, despite their high proliferation rate, showed the lowest deformylation activity. In order to analyze all the data, we therefore normalized the 2′-F-fdC (**1**) deformylation level of each cell line to J1 mES cells. When analyzing the rates of deformylation, we noticed that they inversely correlate ($R^2 = 0.83$) with the rates of cell division (Figure SI-2A). In general, the cells that divide the fastest are the mESCs. In contrast, R1, J1, and K3 (between 17 and 19 h doubling time) show the

lowest deformylation rates, while the cancer-originating somatic cell lines (with doubling times between 25 and 45 hours) show more efficient deformylation activity. This observation is interesting, as this inverse correlation with cell division could indicate that deformylation is an active process that occurs to a larger extent within resting cells.

We could not correlate the deformylation levels with the activity of Tet enzymes ($R^2 = 0.0005$, based on the levels of genomic natural hmdC, Figure SI-2B). In conclusion, the combined data show that deformylation of 2′-F-fdC (**1**) is abundant in a broad spectrum of cell types. The rate of this process strongly depends on the type of the cell. The observed deformylation process is less efficient in stem cells, while it occurs to high extent in human cancer cell lines that in this study are derived from colon and breast cancer tissue.

We performed a control experiment to evaluate the stability of the fed 2′-F-fdC (**1**) material in the cytosolic nucleoside fraction of the cell and hence under conditions as natural as possible. To this end, the probe **1** was added to the cell culture to achieve its uptake into the cells. Next, after 72 h, we isolated the cytosol containing the soluble nucleosides, extracted the nucleosides and analyzed them by UHPLC-MS/MS to look for the presence of **1** and the deformylated putative product **2** (Figure SI-3A). This experiment allowed us to estimate the stability of **1** and to exclude the possibility that the deformylation reaction occurs in the cytosol. If **2** had formed in the cytosol, it would have been able to incorporate into the DNA, obscuring all discussion of intragenomic deformylation. To our delight, we were unable to detect any **2** in the cytosol, which indicates that **1** is stable and that deformylation does not happen outside the genome. This is true within the limit of detection for 2′-F-dC (**2**), which is, however, as low as 3.0 fmol (Table SI-2). The data together, therefore, again suggest that 2′-F-fdC (**1**) is deformylated after its incorporation into the genome.

Because we detected basic 2′-F-fdC (**1**) levels of 2–5 × 10^{-7}/dN in all investigated cellular systems, we sought to find out if these levels can be reduced to zero when replication ceases. To achieve this aim, we conducted an experiment with iPS cells, which stop replication during differentiation towards neurons. This system allows us to study deformylation independently from replication. For the experiment we used the small molecule-inducible Neurogenin iPS cell line (iNGN cells), which is able to differentiate from the pluripotent state in specific culture conditions (see SI) to cells showing a bipolar neuron-like morphology after only 4 days. Like other neurons, they stop DNA replication and cell division upon maturation, which enables the study of a dynamic process like deformylation without interference from replication-dependent incorporation of fresh material. First, the stability of **1** was again examined analogously to the soluble pool extraction experiment with J1 mESCs. Similar to the previous data we found **1** but no traces of **2** after 72 h in the soluble pool of iNGNs, again showing that **1** is not deformylated under our conditions in the active cell extracts (Figure SI-3B).

To monitor cell division, the cells were fed with a final concentration of 0.5 µм of isotopically labeled dT ($^{13}C_{10}$-$^{15}N_2$-dT, **10**). The incorporation of **10** was then traced over time.

Communications

After initial decrease of the amount of **10** through cell division, we detected stable levels of **10** in the genome of these cells from 48 h post-induction. This result allows us to conclude that the cells terminate their replication under our conditions after two days from the start of differentiation (Figure SI-4).

We next performed a double feeding experiment with 2′-F-fdC (**1**) and simultaneously with $^{13}C_{10}$-$^{15}N_2$-dT (**10**) as depicted in Figure 3 in order to monitor cell division-dependent wash-out of **10**. This experiment allows us to check how deformylation progresses upon cessation of cell replication. Quantification of genome-integrated modified nucleosides and their metabolites within DNA was performed by UHPLC-MS/MS using the isotope dilution technique (Table SI-3). The deformylation (a rate of change of 2′-F-fdC (**1**) to 2′-F-dC (**2**) relative to dN) was monitored by pulse feeding the cells at 350 μM final concentration of **1** up to 48 h

post-induction (start of differentiation) and chasing of the values every 4 hours, from 48 h onwards. The data show that the decrease of 2′-F-fdC (**1**) content within the gDNA (Figure SI-5) continues to a small extent even beyond termination of cell division. We can see that while the levels of **10** in the genome stay constant for the duration of the experiment, the ratios between **2** and **1** increase. Because the total levels of **2** in comparison to **1** are at least 10-fold higher due to the efficiency of turnover, post 48 hours it is not possible to see a clear rise in 2′-F-dC/dN levels. With the knowledge that the deformylation reaction is efficient and occurs already after one hour after incorporation of **1** into the genome, we can assume that the majority of incorporated 2′-F-fdC (**1**) by the 48th hour has already turned over to **2** and **9** (Figure SI-5).[17] Also, the levels of 2′-F-fdC/dN stay constant after 52 h, which further supports the idea that some portion of fdC resides as a (semi)permanent base in the genome. This 2′-F-fdC (**1**) stays hidden in the genome. It seems not accessible for enzymatic processes, potentially because it is situated in condensed chromatin structures.[10,28–33] Following these findings, we decided to perform an in vitro experiment to investigate if histones, with their many amine side chains, could be triggering the deformylation process of fdC.[28] This study, however, was negative, which indicates that more is needed than just the presence of lysine side chains to achieve C—C bond cleavage (Figure SI-7). There is certainly a need for active nucleophiles that are able to attack the C6-position of 2′-F-fdC (**1**) to promote the reaction.[18]

The herein reported data show that the synthesis pathway towards the probe molecule of 2′-F-fdC (**1**) using only 2′-F-dU intermediates ensures the best possible purity of the compound for biological feeding experiments. Feeding of the newly synthesized compound shows that deformylation of 2′-F-fdC (**1**) to 2′-F-dC (**2**) occurs inside the genome and that it is a common process in a variety of cell lines that is inversely correlated with the speed of replication. Within differentiating neurons, we were able to show that the vast majority of deformylation occurs before the 48 hour timepoint, which is the moment when the cells stop replicating. Post-replication, we see stable levels of 2′-F-fdC (**1**) that may be epigenetically relevant, for example, by helping chromatin condensation via Schiff-base formation with the histones.[10,28]

Acknowledgements
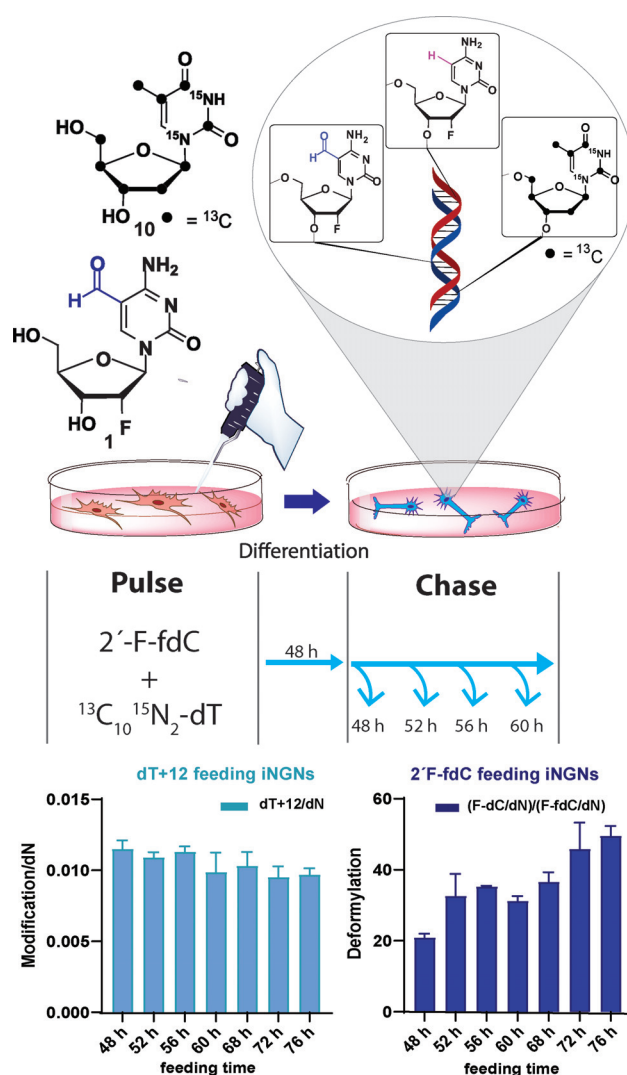
Figure 3. Double feeding experiment of **1** and **10** in iNGN cells, which allow the investigation of the deformylation process depending on cell division. Deformylation data and amounts of incorporated **10** at different time points are shown normalized per dN. The error bars represent propagated error based on standard deviation of each measured technical triplicate.

Conflict of Interest

The authors declare no conflict of interest.

[1] C. C. Chen, K. Y. Wang, C. K. Shen, *J. Biol. Chem.* **2013**, *288*, 9084–9091.

[2] M. Okano, D. W. Bell, D. A. Haber, E. Li, *Cell* **1999**, *99*, 247–257.

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

[4] Z. D. Smith, M. M. Chan, T. S. Mikkelsen, H. Gu, A. Gnirke, A. Regev, A. Meissner, *Nature* **2012**, *484*, 339–344.

[5] S. Liu, J. Wang, Y. Su, C. Guerrero, Y. Zeng, D. Mitra, P. J. Brooks, D. E. Fisher, H. Song, Y. Wang, *Nucleic Acids Res.* **2013**, *41*, 6421–6429.

[6] C. B. Mulholland, F. R. Traube, E. Ugur, E. Parsa, E. M. Eckl, M. Schönung, M. Modic, M. D. Bartoschek, P. Stolz, J. Ryan, T. Carell, H. Leonhardt, S. Bultmann, *Sci. Rep.* **2020**, *10*, 12066.

[7] M. Wagner, J. Steinbacher, T. F. 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] T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, C. A. Deiml, C. Hagemeier, T. Carell, *Angew. Chem. Int. Ed.* **2011**, *50*, 7008–7012; *Angew. Chem.* **2011**, *123*, 7146–7150.

[9] S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, Y. Zhang, *Science* **2011**, *333*, 1300–1303.

[10] M. Bachman, S. Uribe-Lewis, X. Yang, H. E. Burgess, M. Iurlaro, W. Reik, A. Murrell, S. Balasubramanian, *Nat. Chem. Biol.* **2015**, *11*, 555–557.

[11] V. Valinluck, L. C. Sowers, *Cancer Res.* **2007**, *67*, 946–950.

[12] D.-Q. Shi, I. Ali, J. Tang, W.-C. Yang, *Front. Genet.* **2017**, *8*, 100.

[13] M. Münzel, D. Globisch, T. Carell, *Angew. Chem. Int. Ed.* **2011**, *50*, 6460–6468; *Angew. Chem.* **2011**, *123*, 6588–6596.

[14] C. G. Spruijt, F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. Jansen, C. Bauer, M. Münzel, M. Wagner, M. Müller, F. Khan, H. C. Eberl, A. Mensinga, A. B. Brinkman, K. Lephikov, U. Müller, J. Walter, R. Boelens, H. van Ingen, H. Leonhardt, T. Carell, M. Vermeulen, *Cell* **2013**, *152*, 1146–1159.

[15] R. Rahimoff, O. Kosmatchev, A. Kirchner, T. Pfaffeneder, F. Spada, V. Brantl, M. Müller, T. Carell, *J. Am. Chem. Soc.* **2017**, *139*, 10359–10364.

[16] T. Fu, L. Liu, Q. L. Yang, Y. Wang, P. Xu, L. Zhang, S. Liu, Q. Dai, Q. Ji, G. L. Xu, C. He, C. Luo, L. Zhang, *Chem. Sci.* **2019**, *10*, 7407–7417.

[17] K. Iwan, R. Rahimoff, A. Kirchner, F. Spada, A. S. Schröder, O. Kosmatchev, S. Ferizaj, J. Steinbacher, E. Parsa, M. Müller, T. Carell, *Nat. Chem. Biol.* **2018**, *14*, 72–78.

[18] A. Schön, E. Kamińska, F. Schelter, E. Ponkkonen, E. Korytiaková, S. Schiffers, T. Carell, *Angew. Chem. Int. Ed.* **2020**, *59*, 5591–5594; *Angew. Chem.* **2020**, *132*, 5639–5643.

[19] S. Schiesser, T. Pfaffeneder, K. Sadeghian, B. Hackner, B. Steigenberger, A. S. Schröder, J. Steinbacher, G. Kashiwazaki, G. Höfner, K. T. Wanner, C. Ochsenfeld, T. Carell, *J. Am. Chem. Soc.* **2013**, *135*, 14593–14599.

[20] C. Kunz, F. Focke, Y. Saito, D. Schuermann, T. Lettieri, J. Selfridge, P. Schär, *PLoS Biol.* **2009**, *7*, e1000091.

[21] S. C. Wu, Y. Zhang, *Nat. Rev. Mol. Cell. Biol.* **2010**, *11*, 607–620.

[22] K. W. Caldecott, *Nat. Rev. Genet.* **2008**, *9*, 619–631.

[23] A. S. Schröder, E. Parsa, K. Iwan, M. Wallner, S. Serdjukow, T. Carell, *Chem. Commun.* **2016**, *52*, 14361–14364.

[24] A. S. Schröder, O. Kotljarova, E. Parsa, K. Iwan, N. Raddaoui, T. Carell, *Org. Lett.* **2016**, *18*, 4368–4371.

[25] F. Yuan, Y. Bi, J.-Y. Zhang, Y.-L. Zhou, X.-X. Zhang, C.-X. Song, *RSC Adv.* **2019**, *9*, 29010–29014.

[26] F. R. Traube, S. Schiffers, K. Iwan, S. Kellner, F. Spada, M. Müller, T. Carell, *Nat. Protoc.* **2019**, *14*, 283–312.

[27] A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, E. S. Lander, *Nature* **2008**, *454*, 766–770.

[28] E. A. Raiber, G. Portella, S. Martínez Cuesta, R. Hardisty, P. Murat, Z. Li, M. Iurlaro, W. Dean, J. Spindel, D. Beraldi, Z. Liu, M. A. Dawson, W. Reik, S. Balasubramanian, *Nat. Chem.* **2018**, *10*, 1258–1266.

[29] M. Su, A. Kirchner, S. Stazzoni, M. Müller, M. Wagner, A. Schröder, T. Carell, *Angew. Chem. Int. Ed.* **2016**, *55*, 11797–11800; *Angew. Chem.* **2016**, *128*, 11974–11978.

[30] C. X. Song, K. 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.

[31] F. Neri, D. Incarnato, A. Krepelova, S. Rapelli, F. Anselmi, C. Parlato, C. Medana, F. Dal Bello, S. Oliviero, *Cell Rep.* **2015**, *10*, 674–683.

[32] E.-A. Raiber, P. Murat, D. Y. Chirgadze, D. Beraldi, B. F. Luisi, S. Balasubramanian, *Nat. Struct. Mol. Biol.* **2015**, *22*, 44–49.

[33] M. W. Kellinger, C. X. Song, J. Chong, X. Y. Lu, C. He, D. Wang, *Nat. Struct. Mol. Biol.* **2012**, *19*, 831–833.

Chapter 3. The mechanism of direct decarboxylation

"Intragenomic Decarboxylation of 5-Carboxy-2'-deoxycytidine" <u>Ewelina Kamińska*,</u> M. Sc. Eva Korytiaková*, M. Sc. Andreas Reichl, Dr. Markus Müller, Prof. Dr. Thomas Carell.

*These authors equally contributed to this work.

Summary

After investigating of the deformylation process, we wanted to gain better understanding of the sister process – the direct decarboxylation pathway. We initially employed the established experimental methods, nevertheless they proved to be found inapplicable when it came to the chemically demanding carboxyl group. For this study, we created a new pulse-feeding protocol. The first feeding was done with a 2'fluorinated-cadC triphosphates enclosed within a vector that successfully transported the compound through cellular membrane. There were many challenges regarding ion suppression when quantifying the exact amount of the incorporated material using isotopically labelled standards on Triple Quadrupole LC-MS/MS system that were finally solved.[186,227] The general experimental workflow is depicted in Figure 15. The triphosphates are more readily (than previously used F-fdC) incorporated into the genomic DNA, since they do not require any phosphorylation steps. The decarboxylation was not observed in the cytosol, thus we are convinced that the measured high levels of the direct decarboxylation happen intragenomically. Inversely to the previously studied deformylation, the decarboxylation is much more pronounced in stem cells than in somatic lines. The results presented in this manuscript show that decarboxylation may be a more favourable process than deformylation what stays in line with the data from previous group members such as Schiesser et al.[198] Our findings were confirmed by the study of Feng et al. (2021), who also demonstrated the existence of a direct decarboxylation.[186,228]
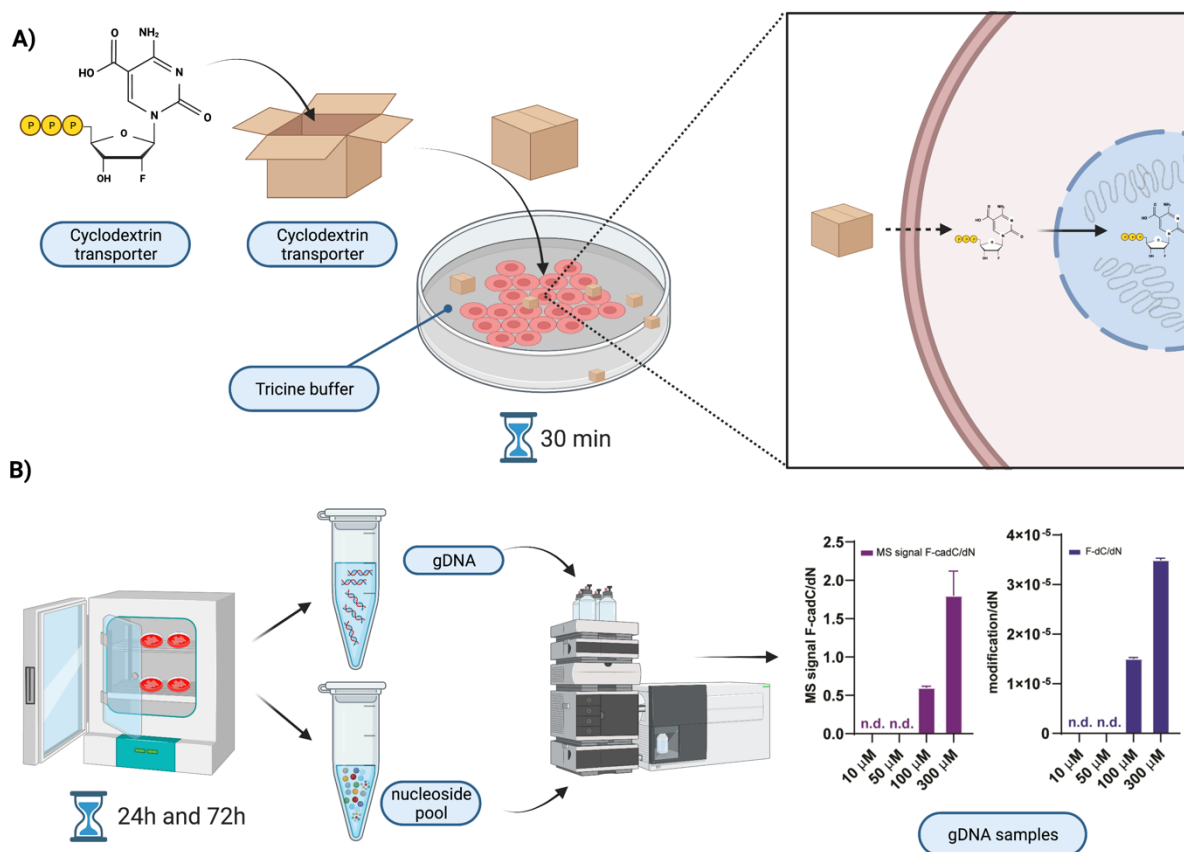
Figure 15. **The general workflow of the F-cadCTP cellular decarboxylation experiments.** The cells are incubated with F-cadCTP and cyclodextrin transporter in tricine buffer for 30 min. The cells were harvested at two timepoints – 24 h and 72 h. Later, the gDNA and nucleoside pool are extracted. The samples are prepared to be measured on the Triple Quadrupole Mass Spectrometry. *Created using BioRender tool.*

AUTHOR'S CONTRIBUTION

I cultured the cells used in this study and performed the cell culture experiments. I also extracted nucleoside soluble pool and gDNA from the harvested cells as well as prepared the samples for LC-MS/MS measurements. The final conclusions were drawn during the discussion with all authors. Next, I participated in the data analysis and writing process of the manuscript. Together with Eva Korytiakovà, I created the figures and reviewed the draft versions of this manuscript. The triphosphate was provided by Eva Korytiaková from the Carell research group.

LICENSE

Nucleosides

Intragenomic Decarboxylation of 5-Carboxy-2′-deoxycytidine

Ewelina Kamińska[+], Eva Korytiaková[+], Andreas Reichl, Markus Müller, and Thomas Carell*

Abstract: *Cellular DNA is composed of four canonical nucleosides (dA, dC, dG and T), which form two Watson–Crick base pairs. In addition, 5-methylcytosine (mdC) may be present. The methylation of dC to mdC is known to regulate transcriptional activity. Next to these five nucleosides, the genome, particularly of stem cells, contains three additional dC derivatives, which are formed by stepwise oxidation of the methyl group of mdC with the help of Tet enzymes. These are 5-hydroxymethyl-dC (hmdC), 5-formyl-dC (fdC), and 5-carboxy-dC (cadC). It is believed that fdC and cadC are converted back into dC, which establishes an epigenetic control cycle that starts with methylation of dC to mdC, followed by oxidation and removal of fdC and cadC. While fdC was shown to undergo intragenomic deformylation to give dC directly, a similar decarboxylation of cadC was postulated but not yet observed on the genomic level. By using metabolic labelling, we show here that cadC decarboxylates in several cell types, which confirms that both fdC and cadC are nucleosides that are directly converted back to dC within the genome by C−C bond cleavage.*

5-Formyl-dC (fdC) and 5-carboxy-dC (cadC) are nucleosides that are found in significant amounts in neurons and stem cells.[1] They are formed by oxidation of 5-methyl-dC (mdC) by the action of Tet enzymes via 5-hydroxymethyl-dC (hmdC).[2,3] hmdC is found in these genomes in large quantities. The initial methylation of dC to mdC is performed by the dedicated methyltransferases Dnmt1, -3a and -3b.[4–7] The higher oxidized mdC derivatives, fdC and cadC are known to be removed by the repair glycosylase Tdg, which cleaves the glycosidic bond between the sugar and the corresponding base.[8,9] The result is the formation of abasic sites that are further processed by AP endonuclease, and finally replaced by an unmodified dC (Figure 1a).[10]

Figure 1. a) Active demethylation pathways via Tdg-mediated excision or direct deformylation and decarboxylation. b) Metabolically fed nucleoside 2′-F-cadC (**1**) and product nucleoside 2-′F-dC (**2**) formed after decarboxylation. MS reference compound [15N]₂-2′F-cadC (**3**) and [15N]₂-2′F-dC (**4**).[17]

[*] E. Kamińska,[+] M. Sc. E. Korytiaková,[+] M. Sc. A. Reichl, Dr. M. Müller, Prof. Dr. T. Carell
Department of Chemistry
Ludwig-Maximilians-Universität München
Butenandtstrasse 5–13, 81377 Munich (Germany)
E-mail: Thomas.Carell@lmu.de
Homepage: http://www.carellgroup.de

[+] These authors contributed equally to this work.

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under:
https://doi.org/10.1002/anie.202109995.

Since the discovery that methylation of dC is followed by oxidation chemistry, it was postulated that fdC and cadC might directly deformylate or decarboxylate to give dC. Chemically, these C−C bond cleavage reactions have the advantage that potentially harmful abasic site intermediates formed during Tdg-mediated active demethylation will not be generated.[11] While for fdC, deformylation was shown to occur in vivo, for cadC just a putative decarboxylation mechanism was postulated in vitro so far.[12,13] It is still unknown whether decarboxylation of cadC occurs in stem cells.[13–16] Here we use our previously described metabolic labelling approach to prove that cadC, if present in the genome of stem and somatic cells, does decarboxylate.[17,18]

Such a direct demethylation reaction in DNA by C−C bond cleavage will give a product that is identical with natural dC. Therefore, it is important to incorporate a reporter nucleoside (cadC*) into the genome of the cells, which generates a decarboxylated dC*. This product must be detectable with high accuracy in the presence of an overwhelming amount of natural dC. In the past we successfully

experimented with 2′-fluoro-labelled nucleosides. Thus, for this study we decided to use 2′F-cadC (**1**) as cadC* (Figure 1 b).[17] The synthesis of **1** and its triphosphate was reported by us previously.[19,20] The 2′F-atom ensures that the formed decarboxylated product 2′F-dC (**2**) is readily detectable by UHPLC–MS due to the $m/z = +19$ Da mass shift relative to dC. The specific shift in retention time for fluorinated compounds also allows to distinguish them from canonical dC. This reduces an overlap of 2′F-dC with dC, which avoids ion suppression that makes quantification of even very small quantities of product possible. The 2′F-atom served also a second purpose. Natural cadC is barely detectable in wildtype cells.[8] The levels, however, increase by almost two orders of magnitude, when the base excision repair (BER) pathway is interrupted by knocking out the *TDG* gene (Figure SI-6).[21] The 2′F substitution has the same effect. It blocks the BER process, which leads to higher amounts of detectable, incorporated 2′F-cadC (**1**).[19]

As a reference compound for exact quantification of **1**, we synthesized the isotopologue [^{15}N]$_2$-2′F-cadC (**3**) as an internal standard. Quantification of **2** required the standard [^{15}N]$_2$-2′F-dC (**4**, Figure 1 b).

A disadvantage of using 2′F-cadC for studying natural decarboxylation is of course the unnatural character of the nucleoside. In order to investigate if the 2′F-atom influences the decarboxylation reaction, we saturated the C5−C6 bond of the cadC- and 2′F-cadC-methylesters and studied the spontaneous decarboxylation behavior after ester cleavage.[13] To our delight we found that both compounds decarboxylate in a similar manner, which lets us conclude that the effect of the 2′F-atom is small (Figure SI-8). When we started to experiment with 2′F-cadC **1** in cellulo, we learned that it is problematic to label the cellular genome with this compound.[22] The negatively charged 2′F-cadC was only taken up by cells to a small extent and in addition, its intracellular conversion to the triphosphate, as needed for incorporation into the genome, happened to be inefficient as well. It is known that the phosphorylation of cadC by kinases is inefficient.[23]

Although we detected 2′F-cadC (**1**) in the cytosol, we were unable to detect incorporated **1** in the genome. We had to continuously feed the cells for 3 days with **1** to reach detectable, but unquantifiable levels of **1**. After intensive experimentation with different 2′F-cadC delivery methods, we finally succeeded with the help of a cyclodextrin transporter which was modified with a cell-penetrating peptide derivative (Figure 2).[22] This transporter encapsulates nucleoside triphosphates and allows them to be transferred across the cell membrane. Application of this transporter was indeed successful. It allowed us to deliver the **1**-TP with only one 30 min feeding pulse in tricine buffer. After the feeding, the cells were washed, and fresh medium was applied (Figure 2).

In the first experiment, we fed Neuro-2a cells for 30 min with the **1**-TP loaded transporter. The cells were harvested after 24 hours. We then isolated the free nucleotide pool and investigated it regarding the presence of 2′F-cadC (**1**) and 2′F-dC (**2**). In the cytosol, we detected to our delight the presence of delivered **1**. In addition, we saw no **2** (Figure 3 a). Together the data show that **1** is a stable compound that, based on our
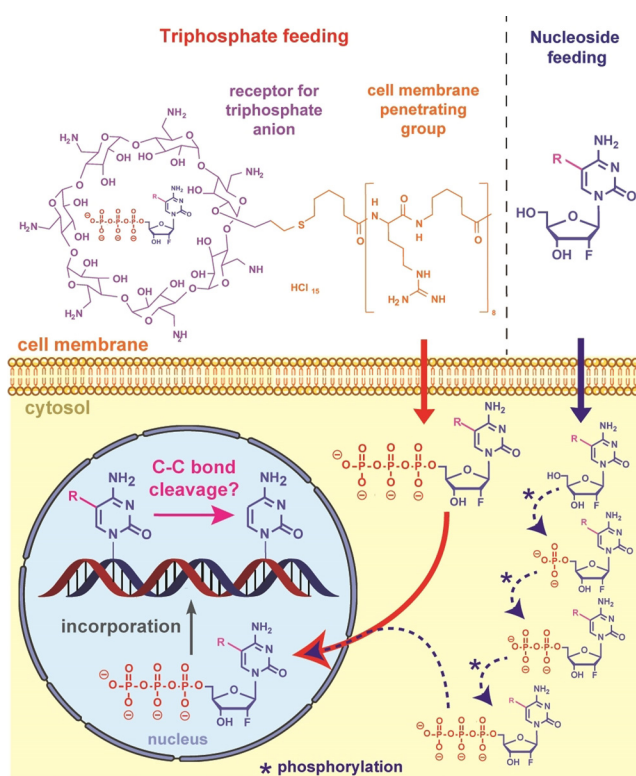
Figure 2. Experimental setup for the inefficient delivery of **1** (blue path) and the accelerated transfer of **1**-TP into cells with the help of the calixarene transporter (red path).

data, does not spontaneously decarboxylate during delivery or under physiological conditions in cells.

Next, we harvested the cells using RLT buffer (Qiagen) supplemented with 400 μM of 2,6-di-*tert*-butyl-4-methylphenol (BHT) and desferoxamine mesylate (DM) as well as β-mercaptoethanol (1:100). The genomic DNA was isolated using a spin column kit (Zymo Research) and digested according to our established method.[18] In brief, the isolated DNA was incubated for 5 min at 95°C, cooled down on ice and incubated with the Degradase digestion mixture (Zymo Research).

After addition of the isotope-labelled internal standards, the mixture was incubated for 4 h at 37°C. Next, the samples were diluted with 450 μL of water and were extracted with chloroform. After lyophilization of the aqueous phase, the digested samples were resuspended in water, filtered and analyzed by UHPLC–MS/MS.[18] In order to prove that this digestion method is efficient for 2′F-cadC-containing DNA, we digested a short ssDNA containing a synthetically embedded **1** (Figure SI-1). Best results were obtained with the Degradase mix (Figure SI-1).[24] We next tested the digestion using a 147-base long dsDNA (Widom 601)[25] containing **1**. Using LC–MS, we detected all canonical DNA bases (dA, dC, dG, T) and in addition, **1** at the expected level, confirming that efficient digestion is possible using the Degradase method. In contrast to this, however, we noticed that when we added the same amount of a 2′F-cadC-containing DNA strand to normal genomic DNA, we obtained an astonishingly small signal for 2′F-cadC (**1**). Indeed, only 10% of the expected signal was

Figure 3. Metabolic feeding of 2′F-cadCTP to Neuro-2a cells. Investigation of the decarboxylation reaction a) in the cytosol and b) in genomic DNA. Dose–response data showing increased MS signal of levels of c) **1** and d) **2** after 24 h with increasing feeding concentration.

2′F-dC (**2**, Figure 3 d). We also see an increase in the MS signal intensity of 2′F-cadC (**1**, Figure 3 c), as expected.

Then, we performed a time-course experiment to investigate at which timepoint decarboxylation becomes detectable (Figure SI-5). We discovered that **2** is detectable only after about 8 h at the earliest, which shows that the C−C bond cleavage has a late onset (Figure SI-5a). As a control experiment, to further exclude that we detect the incorporation of an impurity, we co-fed a 100 μM solution of **1**-TP with a 1 % impurity of 2′F-dCTP (**2**-TP, Figure SI-5b). In this experiment, we see again genome-incorporated **1** and **2**, but now, **2** appeared already after 30 min.[17] Together, the data show that the detected 2′F-dC (**2**) is formed from 2′F-cadC (**1**) by C−C bond cleavage.

In the next experiment, we analyzed deformylation of 2′F-fdC (**5**) and decarboxylation of 2′F-cadC (**1**) side by side (Figure 4 a). We fed Neuro-2a cells with 2′F-fdC-TP in one experiment and with 2′F-cadC-TP in the second. We used the same amount (400 μM) of both materials and of the transporter. We then exchanged the medium and allowed the cells to recover for 24 h. Afterwards, the cells were harvested, the genomic DNA was isolated, digested and analyzed by UHPLC–MS/MS using the synthetic internal standards **4** as well as [15N]₂-2′F-fdC (**6**) for quantification (Figure 4 c).

We detected clearly the fed starting material 2′F-fdC (**5**) and again only just traces (due to ion suppression and/or low incorporation) of 2′F-cadC (**1**). Despite this, the amount of the C−C bond cleavage product 2′F-dC (**2**) was higher when we fed **1**-TP compared to **5**-TP (Figure 4 a). The exact quantification of **2** in both experiments, using the isotopically labelled material **4**, clearly showed more **2** derived from decarboxylation than deformylation. This interesting result

detected, arguing that the detectability of **1** is strongly reduced in a complex environment (Figure SI-2). Due to this strong signal suppression, we therefore abstained from exact quantification of **1**, using internal standard **3** in all further experiments.

To investigate the decarboxylation process, we isolated and digested genomic DNA from Neuro-2a cells fed with **1**-TP. In the above-described method, indeed we clearly detected a signal for **1** in the genome of cells fed with **1**-TP based on its retention time, which was identical with the reference compound **3** and its fragmentation pattern, which was indistinguishable from **3** (Figure 3 b). To our delight, we also detected the decarboxylated product **2**, which was not seen in the cytosolic fraction. Compound **2** was detected at levels of 2.4×10^{-5} per dN, 24 h post feeding. The result shows that while the decarboxylated product **2** does not form in the cytosol by spontaneous decarboxylation, it is present in the genome, arguing that decarboxylation takes place when 2′F-cadC (**1**) is incorporated into genomic DNA.

Next, we investigated if the decarboxylation signals form in a dose-dependent manner (Figure 3 c). Feeding Neuro-2a cells with final concentrations of 100 μM and 300 μM of the **1**-TP-loaded transporter gives indeed an increasing signal for

Figure 4. a) Comparison of deformylation and decarboxylation by feeding of 1-triphosphate and 5-triphosphate as well as investigation of the remethylation of the product **2**. b) Decarboxylation in different cells (J1, R1, E14). c) 2′F-fdC (**5**), 2′F-mdC (**7**), as well as the reference compounds [15N₂]-2′F-fdC (**6**) and [D₃]-2′F-mdC (**8**) needed for exact quantification.[17]

shows that we either dramatically underestimate the genomic incorporation of **1** due to the large ion suppression, or that **1** decarboxylates to **2** more efficiently than **5** deformylates. Based on earlier chemical studies and our current knowledge of both processes, we speculate that faster decarboxylation is the factor contributing more.[26]

We next asked the question if the product nucleoside 2′F-dC (**2**) is remethylated. To this end, we quantified the levels of 2′F-mdC (**7**) in both experiments using the internal standard [D$_3$] 2′F-mdC (**8**). To our delight, we clearly detected **7** in both experiments (Figure 4a). Again, higher amounts were detected from 2′F-cadCTP feeding. This result confirms that after deformylation or decarboxylation, the formed product **2** is remethylated, which suggests the presence of a putative regulatory chemical cycle that starts with methylation of dC by Dnmt enzymes to mdC, followed by oxidation of mdC to hmdC, fdC, and cadC and then deformylation and decarboxylation of fdC and cadC to dC, which could start a new circle.[27] Interestingly, when we measured the levels of **7** formed by methylation of **2** 24 h post **1**-TP feeding, we found only 0.2% methylation. When the same measurement was performed after 72 h, we measured 2.1%, very close to the natural methylation levels of Neuro-2a cells of about 2.5% (Figure SI-7). This result shows that while an early harvest allows us to detect higher amounts of 2′F-dC, the levels of 2′F-mdC are underestimated, potentially because the methyltransferases do not have sufficient time to achieve remethylation.

Finally, we studied how much decarboxylation is taking place in different cells (Figure 4b). We investigated the process in somatic (Neuro-2a, CHO-K1) and in mouse embryonic stem cells (J1, E14, R1). The stem cells were investigated at the pre-implantation and post-implantation stages, obtained by culturing the cells in either a2i or CR media as described in the SI. The data depicted in Figure 4b show that upon feeding **1**-TP at 100 μM over 24 hours, **2** is clearly detectable in all cases but the levels vary between the different cell types. The highest levels of **2** were observed in stem cells cultured under CR conditions. These are the cells representing post-implantation embryos that naturally have the highest methylation levels due to epigenetic reprogramming during cell lineage differentiation.[28] In general, stem cells show an about ten times higher decarboxylation activity than somatic cells, which underpins the potential epigenetic importance of the process.

In summary, the presented data show that next to deformylation of fdC, we also need to consider decarboxylation of cadC as a mechanism for active demethylation. All further efforts now need to be concentrated at finding the cellular entities or circumstances that enable these C—C bond cleavage reactions. Although the here reported data clearly point to the existence of decarboxylation we need to emphasize that cells are complicated entities and we feed an unnatural compound. Without clear identification of the biological entity responsible for the process, we cannot completely rule out that unknown processes other than intragenomic decarboxylation are responsible for the measured data. During the review process of this manuscript Feng and co-workers showed an incorporation of the F-carboxycytosine as a nucleoside and interestingly managed to detect the decarboxylation of cadC to dC much earlier on.[29]

Conflict of Interest

The authors declare no conflict of interest.

Keywords: active demethylation · carboxycytidine · C—C bond cleavage · decarboxylation · epigenetics

[1] 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.

[2] S. Kriaucionis, N. Heintz, *Science* **2009**, *324*, 929–930.

[3] 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] Z. Li, H. Dai, S. N. Martos, B. Xu, Y. Gao, T. Li, G. Zhu, D. E. Schones, Z. Wang, *Genome Biol.* **2015**, *16*, https://doi.org/10.1186/s13059-015-0685-2.

[5] C.-C. Chen, K.-Y. Wang, C.-K. J. Shen, *J. Biol. Chem.* **2013**, *288*, 9084–9091.

[6] C. Luo, P. Hajkova, J. R. Ecker, *Science* **2018**, *361*, 1336–1340.

[7] M. Okano, D. W. Bell, D. A. Haber, E. Li, *Cell* **1999**, *99*, 247–257.

[8] 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.

[9] L. S. Pidugu, Q. Dai, S. S. Malik, E. Pozharski, A. C. Drohat, *J. Am. Chem. Soc.* **2019**, *141*, 18851–18861.

[10] H. E. Krokan, M. Bjoras, *Cold Spring Harbor Perspect. Biol.* **2013**, *5*, a012583.

[11] M. Wossidlo, V. Sebastiano, K. Lepikhov, M. Boiani, R. Reinhardt, H. Schöler, J. Walter, *EMBO J.* **2010**, *29*, 1877–1888.

[12] 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.

[13] S. Schiesser, T. Pfaffeneder, K. Sadeghian, B. Hackner, B. Steigenberger, A. S. Schröder, J. Steinbacher, G. Kashiwazaki, G. Höfner, K. T. Wanner, C. Ochsenfeld, T. Carell, *J. Am. Chem. Soc.* **2013**, *135*, 14593–14599.

[14] G. Song, G. Wang, X. Luo, Y. Cheng, Q. Song, J. Wan, C. Moore, H. Song, P. Jin, J. Qian, J. Hu, *Nat. Commun.* **2021**, *12*, 795.

[15] Z. Liutkeviciute, E. Kriukiene, J. Licyte, M. Rudyte, G. Urbanaviciute, S. Klimasauskas, *J. Am. Chem. Soc.* **2014**, *136*, 5884–5887.

[16] Y. Feng, N.-B. Xie, W.-B. Tao, J.-H. Ding, X.-J. You, C.-J. Ma, X. Zhang, C. Yi, X. Zhou, B.-F. Yuan, Y.-Q. Feng, *CCS Chem.* **2021**, *3*, 994–1008.

[17] K. Iwan, R. Rahimoff, A. Kirchner, F. Spada, A. S. Schröder, O. Kosmatchev, S. Ferizaj, J. Steinbacher, E. Parsa, M. Müller, T. Carell, *Nat. Chem. Biol.* **2018**, *14*, 72–78.

[18] F. R. Traube, S. Schiffers, K. Iwan, S. Kellner, F. Spada, M. Müller, T. Carell, *Nat. Protoc.* **2019**, *14*, 283–312.

[19] A. S. Schröder, O. Kotljarova, E. Parsa, K. Iwan, N. Raddaoui, T. Carell, *Org. Lett.* **2016**, *18*, 4368–4371.

[20] A. S. Schröder, E. Parsa, K. Iwan, M. Wallner, S. Serdjukow, T. Carell, *Chem. Commun.* **2016**, *52*, 14361–14364.

[21] R. Rahimoff, O. Kosmatchev, A. Kirchner, T. Pfaffeneder, F. Spada, V. Brantl, M. Müller, T. Carell, *J. Am. Chem. Soc.* **2017**, *139*, 10359–10364.

[22] Z. Zawada, A. Tatar, P. Mocilac, M. Buděšínský, T. Kraus, *Angew. Chem. Int. Ed.* **2018**, *57*, 9891–9895; *Angew. Chem.* **2018**, *130*, 10039–10043.

[23] M. Zauri, G. Berridge, M.-L. Thézénas, K. M. Pugh, R. Goldin, B. M. Kessler, S. Kriaucionis, *Nature* **2015**, *524*, 114–118.

[24] F. Yuan, Y. Bi, J.-Y. Zhang, Y.-L. Zhou, X.-X. Zhang, C.-X. Song, *RSC Adv.* **2019**, *9*, 29010–29014.

[25] W. J. Lowary, *J. Mol. Biol.* **1998**, *276*, 19–42.

[26] E. Korytiaková, E. Kamińska, M. Müller, T. Carell, *Angew. Chem. Int. Ed.* **2021**, *60*, 16869–16873; *Angew. Chem.* **2021**, *133*, 17005–17010.

[27] A. Parry, S. Rulands, W. Reik, *Nat. Rev. Genet.* **2021**, *22*, 59–66.

[28] S. Takahashi, S. Kobayashi, I. Hiratani, *Cell. Mol. Life Sci.* **2018**, *75*, 1191–1203.

[29] Y. Feng, J.-J. Chen, N.-B. Xie, J.-H. Ding, X.-J. You, W.-B. Tao, X. Zhang, C. Yi, X. Zhou, B.-F. Yuan, Y.-Q. Feng, *Chem. Sci.* **2021**, *12*, 11322–11329.

CHAPTER 4. ADDITIONAL PUBLISHED WORK

"AMPylation profiling during neuronal differentiation reveals extensive variation on lysosomal proteins" Tobias Becker, Cedric Cappel, Francesco Di Matteo, Giovanna Sonsalla, <u>Ewelina Kaminska</u>, Fabio Spada, Silvia Cappello, Markus Damme, Pavel Kielkowski.[229]

AUTHOR'S CONTRIBUTION

I helped with iNGN cell culture system establishment and optimisation for the experiments required in this study as well as revised the manuscript.

SUMMARY

AMPylation is a posttranslational protein modification believed to play a significant role in foetal development of a nervous system. This study reveals its localisation on lysosomal proteins that are associated with neuronal differentiation in iNGN iPS cell system. The chemical-proteomics data combined with gel-based separation techniques showed the correlation between the significant increase of lysosomal exonuclease PLD3 and protein AMPylation levels during maturation of the neurons. This information may help to shed more light on how lysosomal storage disorders may lead to neurodegeneration diseases.

LICENSE

Article

AMPylation profiling during neuronal differentiation reveals extensive variation on lysosomal proteins

Probing protein AMPylation
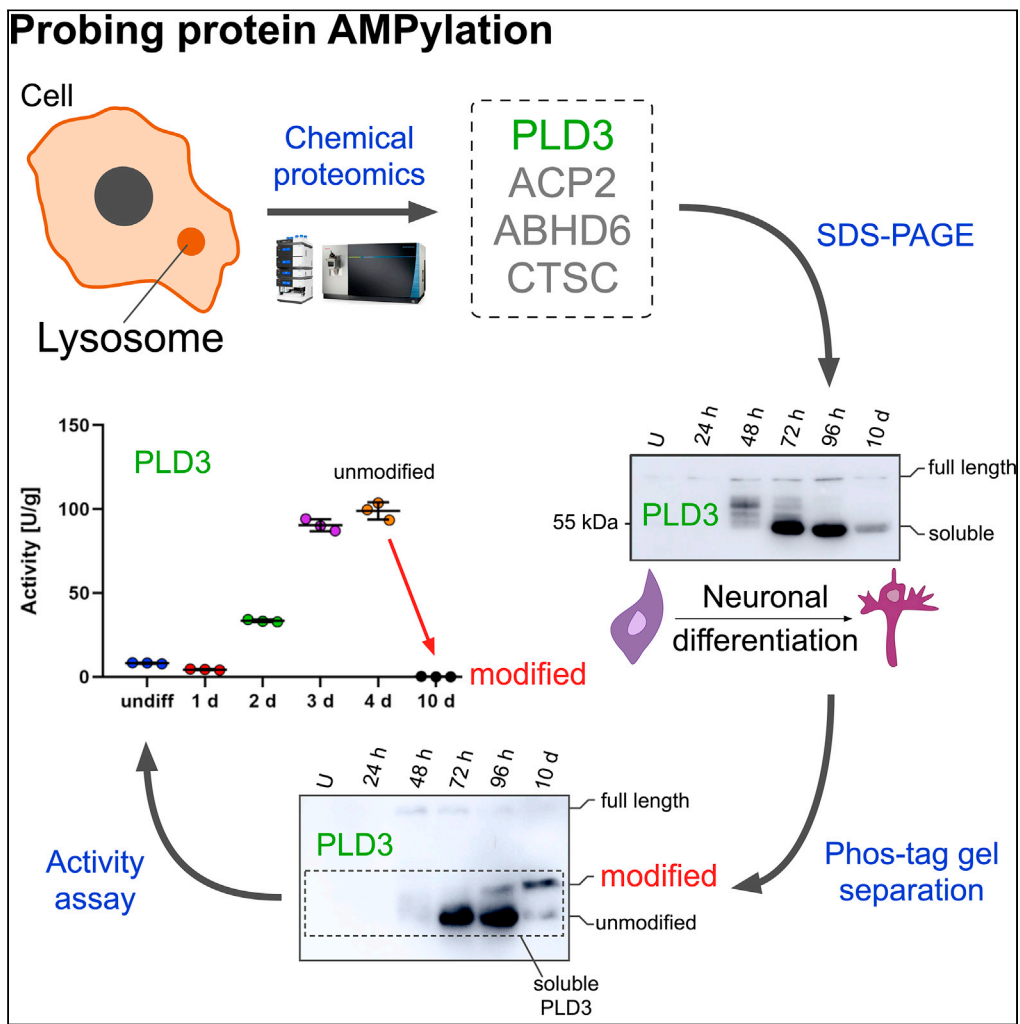
Tobias Becker, Cedric Cappel, Francesco Di Matteo, ..., Silvia Cappello, Markus Damme, Pavel Kielkowski

pavel.kielkowski@cup.lmu.de

Highlights

Profiling of AMPylation during neuronal differentiation

AMPylation is a potential PTM of luminal lysosomal proteins

Phos-tag gel enables the separation of non-AMPylated and AMPylated proteins

The modified lysosomal soluble form of PLD3 increases during neuronal maturation

Article

AMPylation profiling during neuronal differentiation reveals extensive variation on lysosomal proteins

Tobias Becker,[1] Cedric Cappel,[2] Francesco Di Matteo,[3,4] Giovanna Sonsalla,[5,6,7] Ewelina Kaminska,[1] Fabio Spada,[1] Silvia Cappello,[3] Markus Damme,[2] and Pavel Kielkowski[1,8,*]

SUMMARY

Protein AMPylation is a posttranslational modification with an emerging role in neurodevelopment. In metazoans two highly conserved protein AMP-transferases together with a diverse group of AMPylated proteins have been identified using chemical proteomics and biochemical techniques. However, the function of AMPylation remains largely unknown. Particularly problematic is the localization of thus far identified AMPylated proteins and putative AMP-transferases. We show that protein AMPylation is likely a posttranslational modification of luminal lysosomal proteins characteristic in differentiating neurons. Through a combination of chemical proteomics, gel-based separation of modified and unmodified proteins, and an activity assay, we determine that the modified, lysosomal soluble form of exonuclease PLD3 increases dramatically during neuronal maturation and that AMPylation correlates with its catalytic activity. Together, our findings indicate that AMPylation is a so far unknown lysosomal posttranslational modification connected to neuronal differentiation and it may provide a molecular rationale behind lysosomal storage diseases and neurodegeneration.

INTRODUCTION

Protein posttranslational modifications (PTMs) provide the cell with mechanisms to swiftly react on internal and external clues to maintain the cellular homeostasis. Disruption of the protein homeostasis (proteostasis) is a hallmark of many neurodegenerative disorders (Hipp et al., 2019). Regulation of protein function by PTMs includes modulation of protein's catalytic activity, localization or protein-protein interactions (Aebersold et al., 2018). Protein AMPylation comprises the attachment of adenosine 5′-*O*-monophosphate (AMP) onto serine, threonine, and tyrosine amino acid side chains (Figure 1A) (Sieber et al., 2020). So far two AMP transferases are known in metazoans to catalyze protein AMPylation, protein adenylyltransferase FICD (FICD), and SelO (SELENOO) (Casey and Orth, 2017; Sreelatha et al., 2018). FICD is characterized by its endoplasmic reticulum (ER) localization and dual catalytic activity of AMPylation and deAMPylation (Preissler et al., 2017; Sengupta et al., 2019). FICD catalyzes an AMP transfer from its substrate ATP and reverses the modification by the hydrolysis of the AMP-protein ester. FICD's catalytic activity is regulated by its α-helix inhibition loop through the interaction of Glu234 positioned in the inhibition loop and Arg374, which is necessary for the complexation of ATP in the active site (Engel et al., 2012). Initial biochemical studies identified the ER localized heat shock protein HSPA5 (also called BiP or GRP78) as a cognate substrate of FICD (Ham et al., 2014; Sanyal et al., 2015). AMPylation of HSPA5 inhibits its chaperon activity and the downstream unfolded protein response cascade (Preissler et al., 2015). Furthermore, FICD's activity was recently shown to accelerate the neuronal differentiation of progenitor cells in human cerebral organoids, a tissue model of the human cerebral cortex (Kielkowski et al., 2020a). Surprisingly, apart from the increased number of neurons in tissue overexpressing FICD, some neurons displayed migratory defects. The neuronal role of the FIC domain was as well demonstrated in *Drosophila*, in which it is required for the visual neurotransmission in glial capitate projections (Rahman et al., 2012). In addition, FICD was shown to AMPylate α-synuclein *in vitro* (Sanyal et al., 2019). In *Caenorhabditis elegans* FICD has been reported to regulate the aggregation of amyloid-β and α-synuclein (Truttmann et al., 2018). In contrast, there are only scarce data about the function of SelO and its protein substrates (Sreelatha et al., 2018).

Several complementary strategies were introduced to analyze protein AMPylation, including isotopically labeled adenosine probes (Pieles et al., 2014), radioactively labeled adenosine nucleotides, antibodies (Kingdon et al., 1967; Yarbrough et al., 2009), microarrays (Yu and LaBaer, 2015), N^6-biotin-modified ATP

Figure 1. Human neuroblastoma cells (SH-SY5Y) display an enrichment of pro-N6pA probe-labeled proteins in lysosomes

(A) Schematic representation of protein AMPylation.

(B) Structure of the pro-N6pA probe for *in situ* labeling of potentially AMPylated proteins.

(C) Click chemistry staining of pro-N6pA with TAMRA-N_3 (red) and immunocytochemical staining in SH-SY5Y and colocalizations with markers for lysosome (LAMP2), ER (KDEL), and mitochondria (COX-IV).

(D) Volcano plot showing the significantly enriched AMPylated proteins from SH-SY5Y cells. Proteins localized to lysosome are depicted in green. Significantly enriched protein hits with other subcellular localization are in red (cutoff lines at 2-fold enrichment and *p* value of 0.05).

(E) GO terms analysis of significantly enriched proteins in SH-SY5Y cells.

(Sreelatha et al., 2018), and N^6-propargyl adenosine 5'-O-triphosphate (N6pATP) or phosphoramidate (Broncel et al., 2012; Grammel et al., 2011; Kielkowski et al., 2020a, 2020b). We have recently developed a chemical proteomic approach that allows high-throughput screening of protein AMPylation and comparison of the AMPylation levels between different conditions. This strategy utilizes the N^6-propargyl or N^6-ethylazide adenosine phosphoramidate (pro-N6pA and pro-N6azA, respectively) probes, which are metabolically activated to the corresponding N^6-modified adenosine triphosphate upon uptake into cells and used by endogenous AMP-transferases or by bacterial effectors in infection studies for protein AMPylation (Figure 1B) (Kielkowski et al., 2020a, 2020b; Rauh et al., 2020). Of note, the resulting N^6-modified ATP is inherently in competition with endogenous ATP and thus cannot report on the exact stoichiometry of protein AMPylation. The application of this chemical proteomic strategy in various cell types has revealed that protein AMPylation is more prevalent than previously assumed. Interestingly, it has led to the identification of a large group of protein substrates from different subcellular compartments that are not restricted to the ER, including cytosolic (PFKP, SQSTM1), nucleolar (PPME1), cytoskeletal (TUBB, MAP2), and lysosomal (CTSB, PLD3, and ACP2) proteins (Broncel et al., 2016; Kielkowski et al., 2020a). In particular, in the lysosome, there is only limited evidence on PTMs of luminal lysosomal proteins other than N-glycosylation amid numerous diseases associated with lysosomal proteins such as the lysosomal acid phosphatase

ACP2 in lysosomal storage diseases (LSDs) or the 5′-3′ exonuclease PLD3 in Alzheimer's disease and auto-inflammatory diseases (Cappel et al., 2021; Cruchaga et al., 2014; Gavin et al., 2021; Gonzalez et al., 2018; Marques and Saftig, 2019; Schultz et al., 2011; Stadlmann et al., 2017; Stoka et al., 2016). Therefore, the discovery of lysosomal protein PTMs might shed new light on the regulation of their function, localization, and protein-protein interactions and hence putatively uncover unknown pathophysiologic mechanisms.

Here, we characterize protein AMPylation during neuronal differentiation and maturation by a combination of chemical proteomics and a gel-based approach. Thereby, we provide supporting evidence that the two lysosomal proteins PLD3 and ACP2 are increasingly AMPylated during neural differentiation and that the AMPylation of PLD3 correlates with the inhibition of its catalytic activity.

RESULTS

AMPylated proteins localize to lysosomes in SH-SY5Y neuroblastoma cells

We have developed a chemical proteomic strategy in previous studies to analyze the protein AMPylation in living cells using the cell-permeable pro-N6pA probe (Kielkowski et al., 2020a, 2020b). In this study, we have started with the completion of this dataset using fluorescence imaging of the SH-SY5Y cells treated with the pro-N6pA probe to visualize the subcellular localization of potentially AMPylated proteins. Interestingly, probe-labeled proteins exhibit a vesicular distribution pattern overlapping with lysosomes as confirmed by co-staining with the lysosomal-specific marker LAMP2 (Figure 1C). Colocalization of the majority of pro-N6pA probe signal with the ER and mitochondria was excluded by co-staining of the pro-N6pA probe with antibodies against KDEL and COX-IV, respectively. These findings point to an enrichment of AMPylated proteins in lysosomes and raise questions about the localization and origin of their AMPylation. GO analysis of chemical proteomics data from SH-SY5Y cells indicated an overrepresentation of lysosomal proteins among those enriched for AMPylation by 54% of all significantly enriched proteins. These include CTSB, PLD3, GAA, GLB1, TPP1, HEXB, DPP2, and GUSB, with PLD3 and CTSB showing the strongest enrichment (Figures 1D and 1E). Next, we asked whether AMPylation of lysosomal proteins changes during neuronal differentiation and maturation and how the AMPylation status correlates with non-neuronal cell types. Although the chemical proteomic analysis of protein AMPylation variation during neuronal differentiation was previously performed, it has focused solely on two differentiation stages, neuronal progenitors (NPCs) and mature neurons. To investigate the AMPylation of lysosomal proteins in differentiating neurons with higher temporal resolution, we searched for a suitable cellular system that would allow collecting sufficient amounts of total proteins for chemical proteomic analysis in shorter periods of time and thus overcoming the bottlenecks of the standard human induced pluripotent stem cells (iPSCs) differentiation protocols (Boyer et al., 2012).

Chemical proteomic analysis of protein AMPylation in differentiating iNGN cells

To map the changes in protein AMPylation during the neuronal differentiation in more detail, we took advantage of the human induced pluripotent stem cells with inducible overexpression of a pair of transcription factors, Neurogenin-1 and Neurogenin-2, leading to their rapid differentiation into a homogeneous population of functional bipolar neurons within 4 days (iNGN) (Busskamp et al., 2014). The changes in AMPylation were followed at six time points during the iNGN differentiation and maturation using the previously described chemical proteomics strategy utilizing a metabolically activated pro-N6pA probe (Figure 2A). In brief, cells were treated with the pro-N6pA probe 16 h prior to their harvest to allow metabolic activation to the corresponding N^6-propargyl ATP, which is used as a substrate by endogenous AMP-transferases to label proteins. Subsequently, the cells were lysed and the alkyne modified proteins were further coupled with biotin-PEG-azide by Cu(I) catalyzed click chemistry. The pro-N6pA labeled and biotinylated proteins were then enriched on avidin-coated agarose beads and on beads trypsinized to yield complex peptide mixtures, which were resolved by LC-MS/MS measurements. The resulting label-free quantification (LFQ) of proteins from four replicates and their comparison with DMSO-treated cells prepared in the same manner provided the quantitative differences in protein AMPylations between undifferentiated and differentiated iNGN cells (Figures 2B and S1, Tables S1, S2, S3, S4, S5, S6, and S8). Of note, owing to the metabolic activation of the pro-N6pA probe, a minor incorporation, for example, as ADP-ribosylation, cannot be completely excluded. The background from unspecific protein binding to the avidin-agarose beads in both vehicle control (DMSO) and probe-treated cells partially reflects the total protein expression level. To correct for this contribution, we have carried out the whole proteome analysis of iNGNs during the course of differentiation (Figure S2). In addition, the whole proteome analysis confirmed the identity of the cells and the progress of neuronal differentiation and maturation (Figure S3 and Tables
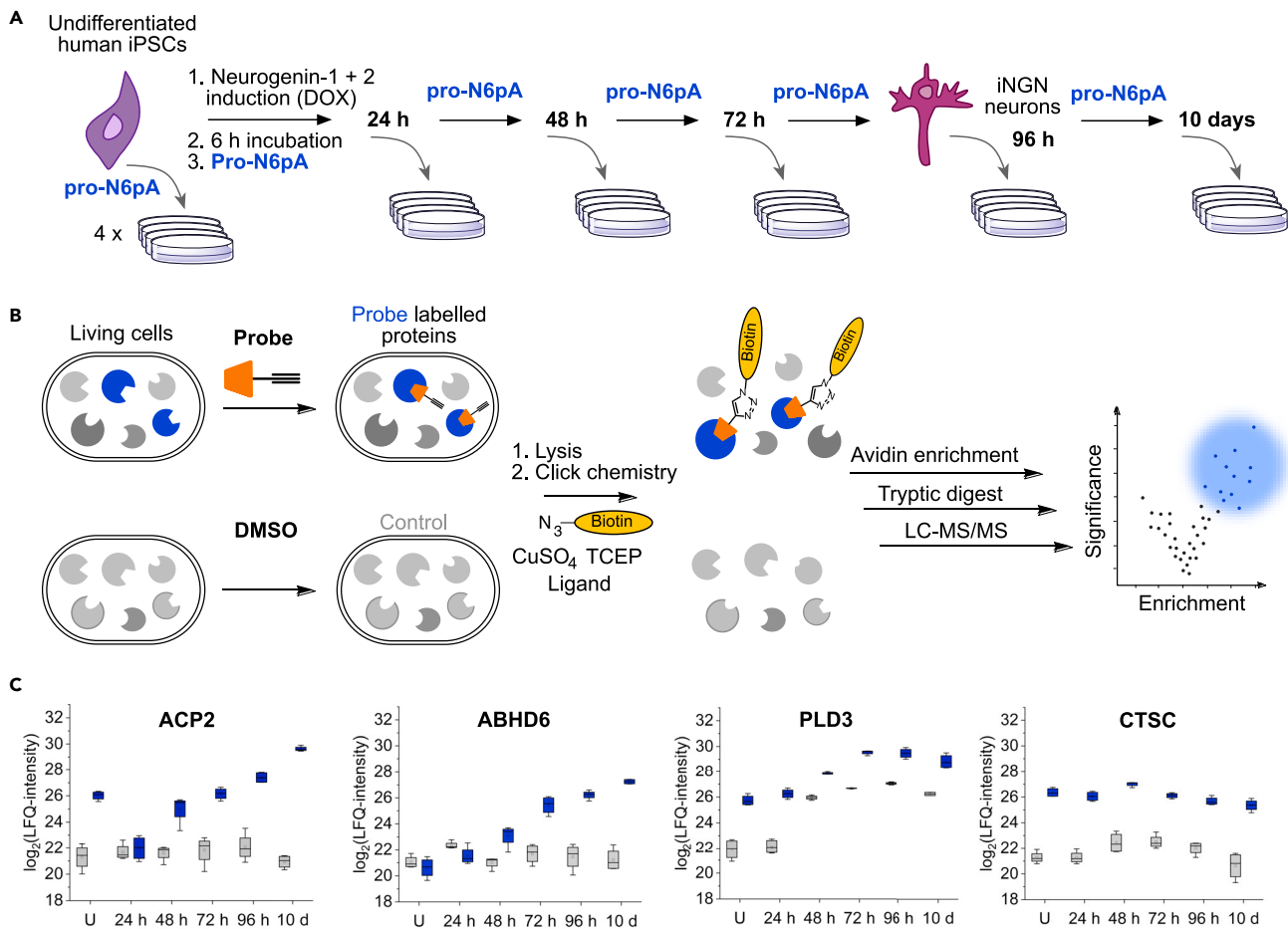
Figure 2. Chemical proteomics of neuronal differentiation and maturation shows specific patterns of protein AMPylation

(A) Schematic iNGNs differentiation procedure and pro-N6pA probe treatment.

(B) Chemical proteomic protocol comparing the probe and DMSO control-treated cells.

(C) Profile plots of significantly enriched proteins ACP2, ABHD6, PLD3, and CTSC from pro-N6pA probe-treated iNGNs. Blue boxes represent the LFQ-intensities from pro-N6pA probe-treated cells after enrichment. Gray boxes represent LFQ-intensities from control (DMSO-treated) cells, showing unspecific binding to avidin-coated agarose beads ($n = 4$, the box is defined by 25th and 75th percentile, whiskers show outliers, line is a median, and circle is a mean).

S7 and S8). The examination of the profile plots of the enriched proteins shows a distinct pattern of AMPylation dynamics during neuronal maturation (Figure 2C). The most distinct composition was observed for the two lysosomal proteins ACP2 and ABHD6, with a linear increase of probe incorporation during the course of differentiation, while the total expression level of both proteins remains stable during the process (Figures 2C and S2). On the other hand, other lysosomal significantly enriched proteins show a stable AMPylation level, for example, CTSC. A similar increase during iNGNs differentiation was observed for the cytosolic protein ATP-dependent 6-phosphofructokinase PFKP, a gatekeeper of glycolysis (Figure S2). Next, we focused on the lysosomal protein PLD3, which was recently associated with Alzheimer's disease, but its physiological function in neurons and regulation have been so far controversial (Arranz et al., 2017; Cruchaga et al., 2014). As shown previously, whole proteome analysis of PLD3 exhibits an increase of PLD3 in neurons compared with undifferentiated iNGNs. PLD3 is known to be transported from the ER and Golgi via endosomes to lysosomes, where the cytosolic N-terminal membrane-bound domain of full-length PLD3 is proteolytically cleaved. The resulting soluble luminal PLD3 containing the putative active site is delivered to the lysosomes (Gonzalez et al., 2018). Both chemical proteomics and whole proteome analysis cannot provide the information on which form of PLD3 is likely to be modified by AMPylation. Thus, we focused on the development of a gel-based methodology that would allow the separation of different protein forms as well as the quantification of the protein AMPylation levels. Previously, isoelectric focusing gels have been
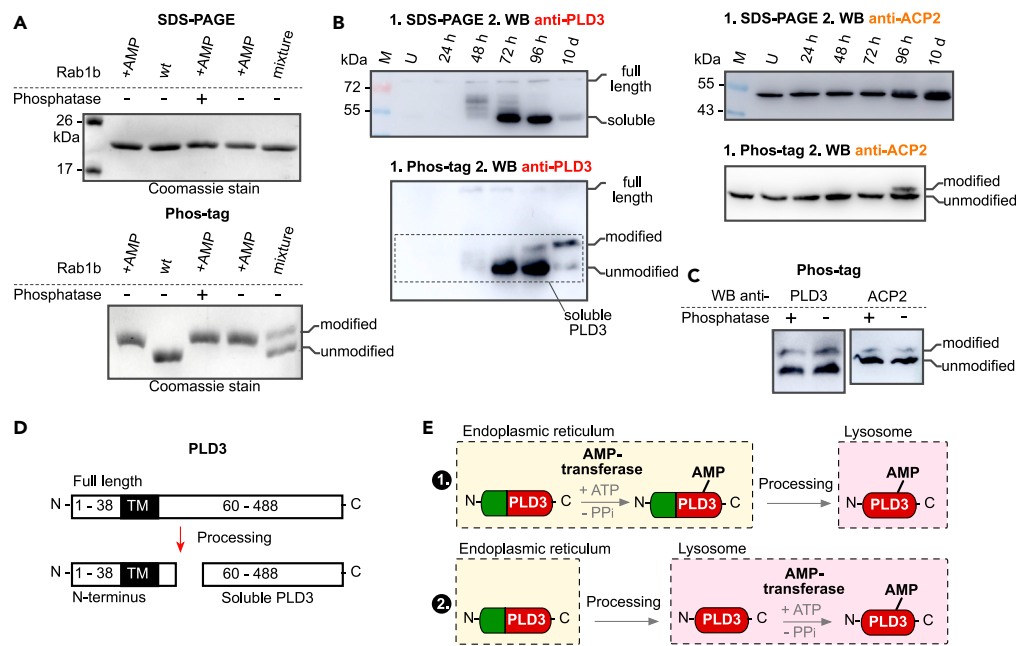
Figure 3. Phos-tag ligand SDS-PAGE distinguishes unmodified and AMPylated proteins
(A) Coomassie-stained SDS-PAGE and Phos-tag ligand containing SDS-PAGE gel separation of unmodified (wt) and AMPylated (+AMP) recombinant Rab1b. With and without treatment with the shrimp alkaline phosphatase (±).
(B) PLD3 and ACP2 Phos-tag ligand SDS-PAGE separation and western blotting during iNGNs differentiation and maturation. Visualization using the anti-PLD3 and anti-ACP2 antibodies.
(C) Phos-tag SDS-PAGE analysis of the PLD3 and ACP2 PTM status after phosphatase treatment.
(D) Processing of the full-length PLD3 into soluble active PLD3.
(E) Scheme showing the two different scenarios of putative intracellular trafficking and AMPylation of the PLD3. M stands for protein marker.

used for the separation of AMPylated HSPA5 from its unmodified form, but the separation of the two species appears to be rather limited (Preissler et al., 2015). Therefore, we explored the possibility to use another gel-based method, which also takes advantage of the presence of a phosphate moiety, but utilize the Phos-tag ligand (Kinoshita et al., 2004, 2006).

Phos-tag ligand-based polyacrylamide gel electrophoresis separates the modified and unmodified proteins

In order to validate our results from chemical proteomics experiments by a complementary method, which would not rely on the pro-N6pA probe and mass spectrometry, we have concentrated on the development of a gel-based approach. Therefore, we have tested the possibility to capture the monophosphate moiety of AMPylated proteins using an alkoxide-bridge Mn^{2+} metal complex bound in the gel by Phos-tag ligand. We have used a recombinant Rab1b protein, which was AMPylated *in vitro* by the recombinant bacterial effector DrrA (Du et al., 2021). The Rab1b identities were confirmed by top-down mass spectrometry (Figure S4). To prepare a Phos-Tag gel, conventional sodium dodecyl sulfate (SDS)-polyacrylamide gels (SDS-PAGEs) were supplemented with the commercially available Mn^{2+} ion-binding Phos-tag ligand. Direct comparison of non-AMPylated, AMPylated, and 1:1 mixture of both modified and unmodified Rab1b showed that indeed the Phos-tag ligand added to standard SDS-PAGE resolves the two species and yields two clearly separated bands as visualized by Coomassie staining (Figure 3A). A control experiment using SDS-PAGE without addition of the Phos-tag ligand showed no separation of the two species (Figure 3A). Because the Phos-tag ligand was initially developed for separation of the phosphorylated proteins, we have treated the *in vitro* AMPylated Rab1b with shrimp alkaline phosphatase to exclude the potential separation due to protein phosphorylation and to show the resistance of the AMP moiety against the phosphatase cleavage (Figure 3A). The activity of the phosphatase was verified by hydrolysis of the phosphorylated ovalbumin (Figure S5). Next, we examined whether it is possible to separate AMPylated HSPA5 from its

unmodified form in HeLa cell lysates. Therefore, after separation the Phos-Tag gel was blotted onto a PVDF membrane and visualized by staining with an anti-HSPA5 antibody. We observed a clear separation of the two forms, confirming our hypothesis that the Phos-tag ligand can be used for the separation of the AMPylated proteins in lysates (Figure S6). Hence, we have established the Phos-tag ligand-functionalized SDS-PAGE separation as a useful method for analysis of AMPylated proteins, which overcomes the necessity for treatment with the probe, and thus it can be used for analysis of protein AMPylation from a wider range of sources, for example, from animal tissues. In our chemical proteomics experiment, we have identified several AMPylated lysosomal proteins, including PLD3 and ACP2. PLD3 can exist either as the full-length protein or as the active soluble form localized in lysosomes, which is obtained by cleavage of the N terminus containing the transmembrane domain (Figure 3D) (Gonzalez et al., 2018). To obtain better insight into the changes of full-length, soluble, and modified PLD3 during neuronal differentiation, we performed the Phos-tag gel-based separation from undifferentiated iNGNs, and iNGNs differentiated between 1 and 10 days. To our surprise, we observed the modified soluble form of PLD3 only after 4 days of differentiation, with a majority of modified PLD3 in 10-day differentiated neurons, which is in line with our observation of increasing pro-N6pA labeling of PLD3 from the chemical proteomics study (Figures 3B and S7). Moreover, analysis by standard SDS-PAGE showed that the soluble lysosomal PLD3 form increases dramatically with increasing time of iNGN differentiation (up to 72 h), whereas it drops substantially upon maturation of iNGN neurons between 4 and 10 days post induction (Figure 3B). In addition to PLD3, a similar AMPylation trend was observed for another lysosomal protein, ACP2, further corroborating the chemical proteomic results and pointing to a specific function of probable protein AMPylation in neuronal maturation (Figure 3B) (Makrypidi et al., 2012). To exclude the separation due to the protein phosphorylation, lysates from the 10-day differentiated iNGNs were treated with the shrimp alkaline phosphatase (Figure 3C). Taken together, the combination of standard SDS-PAGE and Phos-tag-based gel electrophoresis enabled a detailed characterization of PLD3 post-translational processing and AMPylation dynamics during neuronal differentiation and maturation. More specifically, the amount of active soluble PLD3 increases during the differentiation process (up to 72), whereas its AMPylation raises during maturation of iNGN neurons (96 h–10 d). However, the identity and localization of the AMP transferase catalyzing the PLD3 AMPylation remains unknown, thus leaving two plausible scenarios of its AMPylation processing and trafficking into lysosomes (Figure 3E). Even though both methods correlate and point that the observed modification is indeed AMPylation, the exact identity of the PTM has not been confirmed by a direct mass spectrometry experiment. We have attempted to identify the AMPylation site of PLD3. For this, we enriched PLD3 by immunoprecipitation before digestion with trypsin and LC-MS/MS analysis. However, perhaps due to the low sequence coverage (53%) it did not lead to identification of the modification site. In order to specifically enrich only the modified peptides from lysates, we performed enrichment using the desthiobiotin linker, which allows one to selectively elute modified peptides after the enrichment and tryptic digestion (Zanon et al., 2020). Unfortunately, spectra analysis did not lead to identification of either PLD3 or any other site including T518 on the HSPA5. This is likely caused by the low ionization efficiency, fragmentation of the adenosine, and the low abundance of AMPylated peptides.

PTM status correlates with the PLD3 activity in iNGN neurons

To correlate the catalytic activity of PLD3 with its PTM status during the course of iNGNs differentiation, a PLD3-specific acid 5′ exonuclease activity assay was carried out (Cappel et al., 2021). Therefore, whole-cell lysates of iNGN cells were incubated with a fluorophore- and quencher-coupled oligodesoxynucleotide. A 5′ exonucleolytic digest thereby led to a proportional increase in fluorescence signal, measured kinetically over 12 h. The overall activity of PLD3 per total cellular protein increases during the differentiation, which can be accounted for by the increasing levels of its soluble form in lysosomes. In contrast, the PLD3 activity decreases significantly from the 4th day to the 10th day after induction of the differentiation, coinciding with increasing AMPylation of the soluble form during neuronal maturation (Figure 4A). This observation suggests that AMPylation might inhibit PLD3's catalytic activity, which is in line with previous reports on inhibition of the chaperon activity of HSPA5 and the peptidase activity of CTSB.

Directed differentiation of physiological human iPSCs shows the modified PLD3 as the only soluble form in young and mature neurons

In order to compare the AMPylation pattern obtained from the highly homogeneous and robust differentiation and maturation of iNGN cells we have collected lysates from physiological iPSCs, young and mature neurons differentiated for 5 and 10 weeks, respectively (Figures 4B and S8) (Gunhanlar et al., 2018). The modified, likely AMPylated soluble form was the only PLD3 species detectable by Phos-tag SDS-PAGE
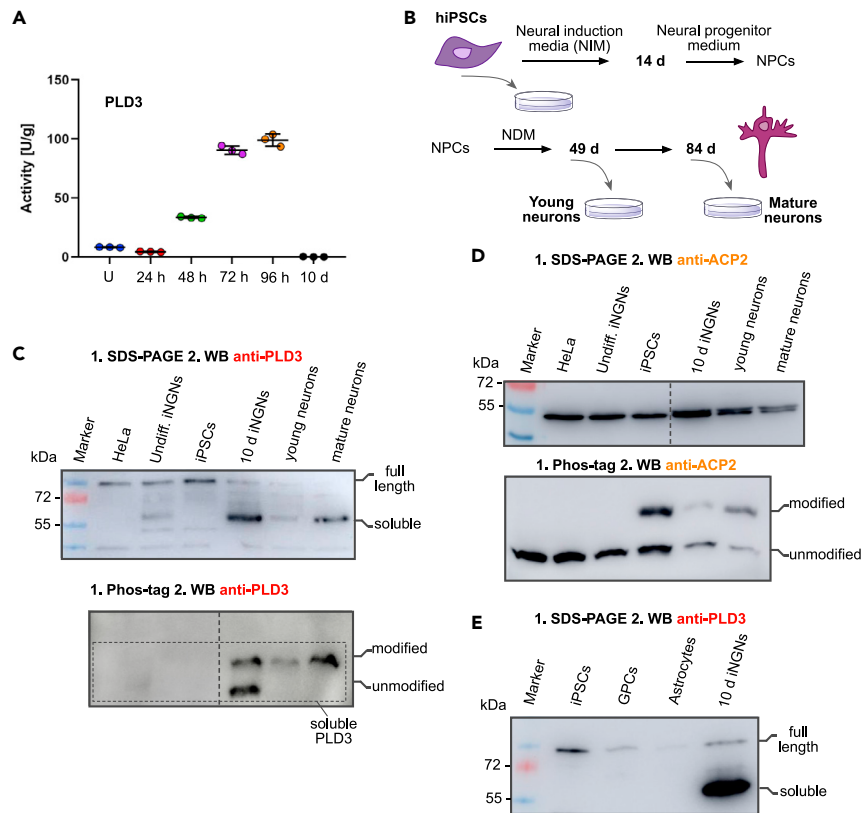
Figure 4. Modification of PLD3 correlates with its activity in neurons

(A) PLD3 5′ exonuclease activity assay correlates the amount of the soluble PLD3 form with its activity and shows the modification to inhibit the catalytic activity ($n = 3$, line indicates the mean, error bars show the standard deviation).

(B) Directed differentiation protocol of human iPSCs into dopaminergic neurons. NDM, neural differentiation medium.

(C) SDS-PAGE and Phos-tag ligand SDS-PAGE separations followed by western blot using the anti-PLD3 antibody show pronounced and nearly quantitative modification, likely AMPylation of PLD3 in dopaminergic neurons compared with the iNGN forward reprogramming.

(D) ACP2 SDS-PAGE and Phos-tag ligand SDS-PAGE separations followed by western blot using anti-ACP2 antibody.

(E) SDS-PAGE followed by western blot using the anti-PLD3 antibody detects solely the full-length PLD3 in iPSCs, GPCs, and astrocytes.

in young as well as in mature neurons, whereas in undifferentiated iPSCs the anti-PLD3 antibody detected no soluble form, whether unmodified or modified (Figures 4C and S9). Comparison of the PLD3 pattern in 10-day differentiated iNGNs, young and mature physiological neurons, with that in lysates from HeLa cells and undifferentiated iPSCs and iNGNs shows that the soluble forms, whether unmodified or AMPylated, are specific to neurons. Moreover, the analysis of PLD3 processing and AMPylation during differentiation of iPSCs into glial progenitor cells (GPCs) and astrocytes confirmed the specificity of this process for the neuronal cell lineage (Figure 4E, see Figure S9 for Phos-tag SDS-PAGE analysis). A similar degree of ACP2 modification was seen in 10-day differentiated iNGN and physiologically mature neural networks (Figure 4D). Taken together, these findings further corroborate the hypothesis that AMPylation plays a specific role in maturation of human neurons.

DISCUSSION

Lysosomal dysfunction is linked to several human pathologies such as LSDs, cancer, neurodegeneration, and aging (Marques and Saftig, 2019). Neuronal cells are particularly sensitive to impaired lysosomal function owing to their tightly controlled differentiation process and postmitotic character. Even though the genetic basis and the biochemistry underlying these diseases are known, the cellular and molecular mechanisms leading to disruption of neuronal viability remain to be understood. Our study describes lysosomal

proteins to be significantly enriched among the AMPylated proteins with changing stoichiometry at various stages of neuronal differentiation, pointing to a specific function of protein AMPylation during differentiation and maturation. The fine-tuning of the lysosomal activity was recently reported to be critical for maintenance of the quiescent neural stem cells fitness and their activation responsiveness (Leeman et al., 2018). Thus, the precise orchestration of these processes might be achieved by adding an extra regulatory layer of protein PTMs, including AMPylation. Furthermore, comparison of the expression levels of the identified AMPylated proteins during the development of mouse and human embryos as well as human organoids shows clear differences between the species and model systems pointing to a regulation of the proteins at different levels and stages of gestation (Figure S10) (Klingler et al., 2021). The gel-based analyses of PLD3 showed a striking increase of the modified soluble form in differentiated iNGN neurons. Comparing the iNGN forward reprogramming model of neuronal differentiation with differentiation of physiological human iPSCs into young and mature neurons showed an even stronger trend with fully modified PLD3 in mature physiological neurons. Prolonging the neurogenic period gives rise to more neurons and expansion of the upper layer of neocortex (Stepien et al., 2020). Thus, lengthening the maturation period of the physiological neurons in comparison with the fast differentiation of iNGN cells may provide more time to establish a more strictly defined modification status of PLD3. In particular, the increased expression of PLD3 was shown to coincide with late neuronal development in the hippocampus and the primary somatosensory cortex (Pedersen et al., 1998). In consequence, the PLD3 AMPylation may modulate PLD3 activity or stability to ensure the proper lysosomal function during the migration and maturation of the basal progenitors from the cortical subventricular zone. The activity assay of PLD3 revealed that the probable AMPylation status correlates with its activity in iNGN neurons. Even though we have characterized the progress of the PLD3 AMPylation in detail, it remains to be elucidated where the AMPylation of PLD3 occurs and which AMP-transferase catalyzes the attachment of this modification (Figure 3E) or if it can be reversed by active deAMPylation. The heterogeneity of the cell types in the human cortex is enormous, including excitatory and inhibitory neurons, astrocytes, microglia, and oligodendrocytes. Analysis of the expression levels of the lysosomal proteins ACP2, ABHD6, PLD3, and CTSC in these cell types exhibits substantial differences (Figure S11) (Kanton et al., 2019). Together with considerable variance in the PTM status, as we have demonstrated for PLD3 in astrocytes and neuronal cell lineage, it highlights the diversity and specificity of the proteoforms in the central nervous system. Therefore, in future studies, it will be important to compare the AMPylation status of these proteins in different cell types and to elucidate the functional consequences. Our study provides evidence that AMPylation is likely a rare modification of the lysosomal proteins and thus gives insight into a putative mechanism modulating the protein activity in this critical cellular department. This finding may open up new possibilities for designing therapeutic strategies against the damage linked to lysosomal dysfunction during neurodevelopment, cancer, and aging.

Limitations of the study

The combination of the two methods, MS-based chemical proteomics and Phos-tag ligand gel analysis, strongly supports the AMPylation of several lysosomal proteins; however, a direct proof for AMPylation is missing in this study. The main bottleneck in using mass spectrometry for validation is the inherent instability of the adenosine monophosphate moiety and the low abundance of the modified peptides. In cells, the pro-N6pA probe is converted to the corresponding N^6-propargyl ATP, which might be further metabolized to N^6-propargyl NAD^+ and incorporated onto proteins as ADP-ribosylation. Although we have shown in our previous study (Kielkowski et al., 2020a) that the amounts are negligible, it cannot be excluded under different experimental conditions. Together, the protein PTM status of lysosomal proteins observed in the chemical proteomic experiments correlates well with the Phos-tag ligand SDS-PAGE analysis. Nevertheless, the limitation of the Phos-tag gel separation is that it cannot distinguish between different NMPylations such as UMPylation or GMPylation. Further development of mass spectrometry experiments should determine the identity of the PTM.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - iNGN cells
  - Generation of NPCs
  - Generation of neurons
  - iPSC culture for astrocytes differentiation
  - Generation of GPCs and astrocytes
  - Culturing of HeLa and SH-SY5Y cells
- METHOD DETAILS
  - Fluorescence imaging
  - Immunoistochemistry – physiological neurons
  - PLD3 activity assay
  - Chemical proteomics
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Calculation chemical proteomics
  - Statistical analysis chemical proteomics
  - Profile plots

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103521.

AUTHOR CONTRIBUTIONS

P.K., M.D., and T.B. conceived the study. T.B. cultured and differentiated the iNGN cells. T.B. carried out all proteomic experiments and gel-based analyses. C.C. performed the fluorescence imaging and PLD3 activity assay. F.D.M. cultured the physiological hiPSCs and differentiated them into neurons. G.S. cultured the hiPSCs and differentiated them into GPCs and astrocytes. E.K. and F.S. helped to establish the iNGN cell culture and reviewed the manuscript. P.K. and T.B. wrote the manuscript. All authors revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Aebersold, R., Agar, J.N., Amster, I.J., Baker, M.S., Bertozzi, C.R., Boja, E.S., Costello, C.E., Cravatt, B.F., Fenselau, C., Garcia, B.A., et al. (2018). How many human proteoforms are there? Nat. Chem. Biol. *14*, 206–214.

Arranz, A.M., Strooper, B., Frigerio, C., Horre, K., Fazzari, P., Saido, T.C., and Saito, T. (2017). PLD3 gene and processing of APP. Nature *541*, E1.

Ayo-Martin, A.C., Kyrousi, C., Giaimo, R.D., and Cappello, S. (2020). GNG5 controls the number of apical and basal progenitors and alters neuronal migration during cortical development. Front. Mol. Biosci. *7*, 578137.

Biedler, J.L., Helson, L., and Spengler, B.A. (1973). Morphology and growth, tumorigenicity, and cytogenetics of human neuroblastoma cells in continuous culture. Cancer Res. *33*, 2643–2652.

Boyer, L.F., Campbell, B., Larkin, S., Mu, Y., and Gage, F.H. (2012). Current Protocols in Stem Cell Biology (Wiley), pp. 1H.6.1–1H.6.11.

Broncel, M., Serwa, R.A., and Tate, E.W. (2012). A new chemical handle for protein AMPylation at the host–pathogen interface. ChemBioChem *13*, 183–185.

Broncel, M., Serwa, R.A., Bunney, T.D., Katan, M., and Tate, E.W. (2016). Global profiling of huntingtin-associated protein E (HYPE)-Mediated AMPylation through a chemical proteomic approach. Mol. Cell Proteomics *15*, 715–725.

Busskamp, V., Lewis, N.E., Guye, P., Ng, A.H., Shipman, S.L., Byrne, S.M., Sanjana, N.E., Murn, J., Li, Y., Li, S., et al. (2014). Rapid neurogenesis through transcriptional activation in human stem cells. Mol. Syst. Biol. *10*, 760.

Cappel, C., Gonzalez, A.C., and Damme, M. (2021). Quantification and characterization of the 5′ exonuclease activity of the lysosomal nuclease PLD3 by a novel cell-based assay. J. Biol. Chem. *296*, 100152.

Casey, A.K., and Orth, K. (2017). Enzymes involved in AMPylation and deAMPylation. Chem. Rev. *118*, 1199–1215.

Cox, J., Hein, M.Y., Luber, C.A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. Mol. Cell Proteomics *13*, 2513–2526.

Cruchaga, C., Karch, C.M., Jin, S., Benitez, B.A., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., et al. (2014). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature *505*, 550.

Du, J., Wrisberg, M.-K., Gulen, B., Stahl, M., Pett, C., Hedberg, C., Lang, K., Schneider, S., and Itzen, A. (2021). Rab1-AMPylation by Legionella DrrA is allosterically activated by Rab1. Nat. Commun. *12*, 460.

Engel, P., Goepfert, A., Stanger, F.V., Harms, A., Schmidt, A., Schirmer, T., and Dehio, C. (2012). Adenylylation control by intra- or intermolecular active-site obstruction in Fic proteins. Nature *482*, 107–110.

Gavin, A.L., Huang, D., Blane, T.R., Thinnes, T.C., Murakami, Y., Fukui, R., Miyake, K., and Nemazee, D. (2021). Cleavage of DNA and RNA by PLD3 and PLD4 limits autoinflammatory triggering by multiple sensors. Nat. Commun. *12*, 5874.

Gonzalez, A.C., Schweizer, M., Jagdmann, S., Bernreuther, C., Reinheckel, T., Saftig, P., and Damme, M. (2018). Unconventional trafficking of mammalian phospholipase D3 to lysosomes. Cell Rep. *22*, 1040–1053.

Grammel, M., Luong, P., Orth, K., and Hang, H.C. (2011). A chemical reporter for protein AMPylation. J. Am. Chem. Soc. *133*, 17103–17105.

Gunhanlar, N., Shpak, G., van der Kroeg, M., Gouty-Colomer, L.A., Munshi, S.T., Lendemeijer, B., Ghazvini, M., Dupont, C., Hoogendijk, W.J.G., Gribnau, J., et al. (2018). A simplified protocol for differentiation of electrophysiologically mature neuronal networks from human induced pluripotent stem cells. Mol. Psychiatr. *23*, 1336–1344.

Ham, H., Woolery, A.R., Tracy, C., Stenesen, D., Krämer, H., and Orth, K. (2014). Unfolded protein response-regulated Drosophila Fic (dFic) protein reversibly AMPylates BiP chaperone during endoplasmic reticulum homeostasis. J. Biol. Chem. *289*, 36059–36069.

Hipp, M.S., Kasturi, P., and Hartl, F.U. (2019). The proteostasis network and its decline in ageing. Nat. Rev. Mol. Cell Biol. *20*, 421–435.

Kanton, S., Boyle, M.J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J.S., Han, D., et al. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature *574*, 418–422.

Kielkowski, P., Buchsbaum, I.Y., Kirsch, V.C., Bach, N.C., Drukker, M., Cappello, S., and Sieber, S.A. (2020a). FICD activity and AMPylation remodelling modulate human neurogenesis. Nat. Commun. *11*, 517.

Kielkowski, P., Buchsbaum, I.Y., Becker, T., Bach, K., Cappello, S., and Sieber, S.A. (2020b). A pronucleotide probe for live-cell imaging of protein AMPylation. Chembiochem *21*, 1285–1287.

Kingdon, H., Shapiro, B., and Stadtman, E. (1967). Regulation of glutamine synthetase. 8. ATP: glutamine synthetase adenylyltransferase, an enzyme that catalyzes alterations in the regulatory properties of glutamine synthetase. Proc. Natl. Acad. Sci. U S A *58*, 1703–1710.

Kinoshita, E., Takahashi, M., Takeda, H., Shiro, M., and Koike, T. (2004). Recognition of phosphate monoester dianion by an alkoxide-bridged dinuclear zinc(ii) complex. Dalton Trans *21*, 1189–1193.

Kinoshita, E., Kinoshita-Kikuta, E., Takiyama, K., and Koike, T. (2006). Phosphate-binding tag, a new tool to visualize phosphorylated proteins. Mol. Cell Proteomics *5*, 749–757.

Klaus, J., Kanton, S., Kyrousi, C., Ayo-Martin, A.C., Giaimo, R.D., Riesenberg, S., O'Neill, A.C., Camp, J.G., Tocco, C., Santel, M., et al. (2019). Altered neuronal migratory trajectories in human cerebral organoids derived from individuals with neuronal heterotopia. Nat. Med. *25*, 561–568.

Klingler, E., Francis, F., Jabaudon, D., and Cappello, S. (2021). Mapping the molecular and cellular complexity of cortical malformations. Science *371*, eaba4517.

Leeman, D.S., Hebestreit, K., Ruetz, T., Webb, A.E., McKay, A., Pollina, E.A., Dulken, B.W., Zhao, X., Yeo, R.W., Ho, T.T., et al. (2018). Lysosome activation clears aggregates and enhances quiescent neural stem cell activation during aging. Science *359*, 1277–1283.

Makrypidi, G., Damme, M., Müller-Loennies, S., Trusch, M., Schmidt, B., Schlüter, H., Heeren, J., Lübke, T., Saftig, P., and Braulke, T. (2012). Mannose 6 dephosphorylation of lysosomal proteins mediated by acid phosphatases Acp2 and Acp5. Mol. Cell Biol. *32*, 774–782.

Marques, A.R.A., and Saftig, P. (2019). Lysosomal storage disorders – challenges, concepts and avenues for therapy: beyond rare diseases. J. Cell Sci. *132*, jcs221739.

Pedersen, K.M., Finsen, B., Celis, J.E., and Jensen, N.A. (1998). Expression of a novel murine phospholipase D homolog coincides with late neuronal development in the forebrain. J. Biol. Chem. *273*, 31494–31504.

Pieles, K., Glatter, T., Harms, A., Schmidt, A., and Dehio, C. (2014). An experimental strategy for the identification of AMPylation targets from complex protein samples. Proteomics *14*, 1048–1052.

Preissler, S., Rato, C., Chen, R., Antrobus, R., Ding, S., Fearnley, I.M., and Ron, D. (2015). AMPylation matches BiP activity to client protein load in the endoplasmic reticulum. ELife *4*, e12621.

Preissler, S., Rato, C., Perera, L., Saudek, V., and Ron, D. (2017). FICD acts bifunctionally to AMPylate and de-AMPylate the endoplasmic reticulum chaperone BiP. Nat. Struct. Mol. Biol. *24*, 23–29.

Puck, T.T., and Marcus, P.I. (1955). A rapid method for viable cell titration and clone production with Hela cells in tissue culture: the use of X-irradiated cells to supply conditioning factors. Proc. Natl. Acad. Sci. U S A *41*, 432–437.

Rahman, M., Ham, H., Liu, X., Sugiura, Y., Orth, K., and Krämer, H. (2012). Visual neurotransmission in Drosophila requires expression of Fic in glial capitate projections. Nat. Neurosci. *15*, 871–875.

Rauh, T., Brameyer, S., Kielkowski, P., Jung, K., and Sieber, S.A. (2020). MS-based in situ proteomics reveals AMPylation of host proteins during bacterial infection. Acs Infect Dis. *6*, 3277–3289.

Santos, R., Vadodaria, K.C., Jaeger, B.N., Mei, A., Lefcochilos-Fogelquist, S., Mendes, A., Erikson, G., Shokhirev, M., Randolph-Moore, L., Fredlender, C., et al. (2017). Differentiation of inflammation-responsive astrocytes from glial progenitors generated from human induced pluripotent stem cells. Stem Cell Rep. *8*, 1757–1769.

Sanyal, A., Chen, A.J., Nakayasu, E.S., Lazar, C.S., Zbornik, E.A., Worby, C.A., Koller, A., and Mattoo, S. (2015). A novel link between fic (filamentation induced by cAMP)-mediated adenylylation/AMPylation and the unfolded protein response. J. Biol. Chem. *290*, 8482–8499.

Sanyal, A., Dutta, S., Camara, A., Chandran, A., Koller, A., Watson, B.G., Sengupta, R., Ysselstein, D., Montenegro, P., Cannon, J., et al. (2019). Alpha-synuclein is a target of fic-mediated adenylylation/AMPylation: possible implications for Parkinson's disease. J. Mol. Biol. *431*, 2266–2282.

Schultz, M.L., Tecedor, L., Chang, M., and Davidson, B.L. (2011). Clarifying lysosomal storage diseases. Trends Neurosci. *34*, 401–410.

Sengupta, R., Poderycki, M.J., and Mattoo, S. (2019). CryoAPEX - an electron tomography tool for subcellular localization of membrane proteins. J. Cell Sci. *132*, jcs222315.

Sieber, S.A., Cappello, S., and Kielkowski, P. (2020). From young to old: AMPylation hits the brain. Cell Chem. Biol. *27*, 773–779.

Sreelatha, A., Yee, S.S., Lopez, V.A., Park, B.C., Kinch, L.N., Pilch, S., Servage, K.A., Zhang, J., Jiou, J., Karasiewicz-Urbańska, M., et al. (2018). Protein AMPylation by an Evolutionarily conserved pseudokinase. Cell *175*, 809–821.e19.

Stadlmann, J., Taubenschmid, J., Wenzel, D., Gattinger, A., Dürnberger, G., Dusberger, F., Elling, U., Mach, L., Mechtler, K., and Penninger, J.M. (2017). Comparative glycoproteomics of stem cells identifies new players in ricin toxicity. Nature *549*, 538–542.

Stepien, B.K., Naumann, R., Holtz, A., Helppi, J., Huttner, W.B., and Vaid, S. (2020). Lengthening neurogenic period during neocortical development causes a hallmark of neocortex expansion. Curr Biol *30*, 4227–4237.e5. https://doi.org/10.1016/j.cub.2020.08.046.

Stoka, V., Turk, V., and Turk, B. (2016). Lysosomal cathepsins and their regulation in aging and neurodegeneration. Ageing Res. Rev. *32*, 22–37.

Truttmann, M.C., Pincus, D., and Ploegh, H.L. (2018). Chaperone AMPylation modulates aggregation and toxicity of neurodegenerative

disease-associated polypeptides. Proc. Natl. Acad. Sci. U S A *115*, E5008–E5017.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat. Methods *13*, 731–740.

Yarbrough, M.L., Li, Y., Kinch, L.N., Grishin, N.V., Ball, H.L., and Orth, K. (2009). AMPylation of rho GTPases by *Vibrio* VopS disrupts effector binding and downstream signaling. Science *323*, 269–272.

Yu, X., and LaBaer, J. (2015). High-throughput identification of proteins with AMPylation using self-assembled human protein (NAPPA) microarrays. Nat. Protoc. *10*, 756–767.

Zanon, P.R.A., Lewald, L., and Hacker, S.M. (2020). Isotopically labeled desthiobiotin azide (isoDTB) tags enable global profiling of the bacterial cysteinome. Angew. Chem. Int. Ed. *59*, 2829–2836.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat polyclonal anti-mouse IgG, AF488-linked	Thermo Fisher Scientific	Cat# A-11001; RRID:AB_2534069
Goat polyclonal anti-rabbit IgG, AF488-linked	Thermo Fisher Scientific	Cat# A-11008; RRID: AB_143165
Goat polyclonal anti-rabbit IgG, HRP-linked	Sigma-Aldrich	Cat# A6667; RRID: AB_258307
Rabbit polyclonal anti-ABHD6	Thermo Fisher Scientific	Cat# PA5-38999; RRID AB_2555591
Rabbit polyclonal anti-ACP2	Thermo Fisher Scientific	Cat# PA5-29961; RRID: AB_2547435
Rabbit polyclonal anti-COX IV	Thermo Fisher Scientific	Cat# ab16056; RRID: AB_443304
Rabbit polyclonal anti-HSPA5	Thermo Fisher Scientific	Cat# PA5-34941; RRID: AB_2552290
Mouse monoclonal anti-KDEL (10C3)	Millipore	Cat# 420400; RRID: AB_212090
Mouse monoclonal anti-LAMP2 (H4B4)	DSHB	Cat# N/A; RRID: AB_528129
Rabbit polyclonal anti-PLD3	Sigma-Aldrich	Cat# HPA012800; RRID: AB_1855330
Mouse monoclonal anti-TUBB3	Sigma-Aldrich	Cat# T8660 RRID: AB_477590
Guinea pig polyclonal anti-DCX	Millipore	Cat# AB2253 RRID: AB_1586992
Rabbit polyclonal anti-TBR1	Millipore	Cat# AB31940 RRID: AB_2200219
Mouse polyclonal anti-MAP2	Millipore	Cat# AB5392 RRID: AB_2138153
Mouse monoclonal anti-Neun	Millipore	Cat# MAB377 RRID: AB_2298772
Goat polyclonal anti-mouse IgG2b, AF647-linked	Thermo Fisher Scientific	Cat# A-21242 RRID: AB_2535811
Goat polyclonal anti-rabbit IgG, AF546-linked	Thermo Fisher Scientific	Cat# A-11010 RRID: AB_2534077
Goat polyclonal anti-Guinea pig IgG, AF647-linked	Thermo Fisher Scientific	Cat# A-21450 RRID: AB_2735091
Goat polyclonal anti-mouse IgG, AF546-linked	Thermo Fisher Scientific	Cat# A-21123 RRID: AB_2535765
Mouse monoclonal anti-S100beta	Sigma-Aldrich	Cat# S2532 RRID: AB_477499
Rabbit polyclonal anti-FGFR3	Santa Cruz	Cat# sc-123 RRID: AB_631511
Goat anti-mouse IgG1, Alexa fluor 488	Thermo Fisher Scientific	Cat# A-21121 RRID: AB_2535764
Rabbit polyclonal anti-Ovalbumin	Thermo Fisher Scientific	Cat# PA5-97525 RRID: AB_2812141
Chemicals, peptides, and recombinant proteins		
Acetone (HPLC grade)	VWR chemicals	Cat# 200067.320; CAS: 67-64-1
Acetonitrile (LC-MS grade)	Fisher Scientific	Cat# A955-212; CAS: 75-05-8
Alanyl-Glutamine	Sigma-Aldrich	Cat# G8541, CAS: 39537-23-0
Ammoniumperoxodisulfat (APS)	Sigma-Aldrich	Cat# 09913; CAS: 7727-54-0
β-Mercaptoethanol	Sigma-Aldrich	Cat# M3148; CAS: 60-24-2

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biotin-PEG$_3$-N$_3$	Carbosynth	Cat# FA34890 CAS: 875770-34-6
Bromphenol blue	Fluka	Cat# 32712 CAS: 115-39-9
BSA	AppliChem	Cat# A6588; CAS: 9048-46-8
Coomassie Blue R-250	Fluka	Cat# 27816; CAS: 6104-59-2
CuSO$_4$ x 5 H$_2$O	Acros	Cat# 10162, 7758-99-8
ddH2O (LC-MS grade)	Honeywell	Cat# 15350 CAS: 732-18-5
DMSO	Sigma-Aldrich	Cat# D4540, CAS: 67-68-5
Dithiothreitol	VWR (AppliChem)	Cat# A2948, CAS: 3483-12-3
Formic Acid (LC-MS grade)	Fisher Scientific	Cat# A117; CAS: 64-18-6
HEPES	Carl Roth	Cat# HN77.5; CAS: 7365-45-9
Iodacetamide	Sigma-Aldrich	Cat# I6125; CAS: 144-48-9
KCl	AppliChem	Cat# A2939; CAS: 7447-40-7
KH$_2$PO$_4$	Sigma-Aldrich	Cat# P9791; CAS: 7778-77-0
NA$_2$HPO$_4$	Sigma-Aldrich	Cat# T876; CAS: 7558-79-4
NA$_2$SeO$_3$	Sigma-Aldrich	Cat# S5261; CAS: 10102-18-8
NaCl	Bernd Kraft GmbH	Cat# KRAF04160; CAS: 7647-14-5
NP40	Sigma-Aldrich	Cat# 74385; CAS: 9016-45-9
Methanol (LC-MS grade)	Fisher Scientific	Cat# A456; CAS: 67-56-1
Powdered milk	AppliChem	Cat# A0830; CAS: 999999-99-4
SDS	AppliChem	Cat# A2572; CAS: 151-21-3
Sodium deoxycholate	Sigma-Aldrich	Cat# 30970; CAS: 302-95-4
TAMRA-N$_3$	Baseclick	CAT# BCFA-008-1
TBTA	TCI	Cat# T2993; CAS: 510758-28-8
TCEP	Carbosynth	Cat# FT01756; CAS: 51805-45-9
TEAB (1 M)	Sigma-Aldrich	Cat# T7408, CAS: 15715-58-9
TEMED	Sigma-Aldrich	Cat# T9281; CAS: 110-18-9
Thiourea	Merck	Cat# 107979; CAS: 62-56-6
Tris-base	Fisher Scientific	Cat# 10344; CAS: 77-86-1
Trypan Blue	Fisher Scientific	Cat# 11886
Tween® 20	VWR (AppliChem)	Cat# A4974; CAS: 9005-64-5
Urea	AppliChem	Cat# A1049, CAS: 57-13-6
DPBS (1×)	Sigma-Aldrich	Cat# D8357
DMEM (1×)	Sigma-Aldrich	Cat# D6546
Ham's F-12 w/o L-Glu	Sigma-Aldrich	Cat# N4888
cOmplete® protease inhibitor	Sigma-Aldrich	Cat# 05001
Normal goat serum	Biozol	Cat# VEC-S-1000
FCS	Thermo Fisher Scientific	Cat# A3840001
hHolo-transferrin	Sigma-Aldrich	Cat# 616424; CAS: 11096-37-0
hFGF-2	MACS Miltenyi Biotec	Cat# 130-104-921
hInsulin	BioXtra	Cat# I9278, CAS: 11061-68-0
hTGF- β1	MACS Miltenyi Biotec	Cat# 130-095-067
Geltrex	Thermo Fisher Scientific	Cat# A1413201
Immobilon® Western HRP sustrate	Merck Millipore	Cat# WBKLS0500
Color prestained protein standard, Broad range (10-250 kDa)	New England BioLabs	Cat# P7719S
Blue prestained protein standard, Broad range (11-250 kDa)	New England BioLabs	Cat# P7718L

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pen-strep	Sigma-Aldrich	Cat# P0781
Rotiphorese.Gel 30 (37, 5:1)	Carl Roth	Cat# 3029.1
TrypLE express	Thermo Fisher Scientific	Cat# 12013
Trypsin	Promega	Cat# V5113
Thiazovivin	Merck Millipore	Cat# 420220; CAS: 1056-71-8
NeuroBrew-21	Miltenyi Biotech	Cat# 130-093-566
4,6-Diamidino-2-phenylindole	Sigma-Aldrich	Cat# D9542
Accutase	Stem Cell Technologies	Cat# 07920
Ascorbic acid	Sigma-Aldrich	Cat# A92902; CAS: 50-81-7
B27-supplement (minus vitamin A)	Thermo Fisher Scientific	Cat# 12010
BDNF	Peprotech	Cat# 450-02
dcAMP	Sigma-Aldrich	Cat# D1256; CAS: 93839-95-3
GDNF	Peprotech	Cat# 450-10
Laminin	Sigma-Aldrich	Cat# L2020; CAS: 114,956-81-9
Matrigel® Basement membrane matrix, LDEV-free	Corning®	Cat# 354234
Minimum essential medium/non-essential amino acid	Thermo Fisher Scientific	Cat# 11050
mTESR1 medium	Stem Cell Technologies	Cat# 85850
N2-supplement	Thermo Fisher Scientific	Cat# 17048
Neurobasal medium	Thermo Fisher Scientific	Cat# 21103049
Polyornithine	Sigma-Aldrich	Cat# P4957
Rock inhibitor Y-27632(2HCl)	Stem Cell Technologies	Cat# 72304
Astrocyte media	ScienCell	Cat# 1801
Recombinant human LIF	Alomone Labs	Cat# L-200
Recombinant human FGF-basic	Peprotech	Cat# AF-100-18 B
Recombinant human EGF	Peprotech	Cat# AF-100-15
Recombinant human noggin	Peprotech	Cat# 120-10C
Recombinant human PDGF-AA	R&D Systems	Cat# 221-AA
Accutase	Thermo Fisher Scientific	Cat# A1110501
Collagenase	Stem Cell Technologies	Cat# 07909
Critical commercial assays		
Pierce® BCA protein Assay kit	Thermo Fisher Scientific	Cat# 23225
Deposited data		
MS raw data and calculation results	ProteomeXchange	PXD023873
Experimental models: Cell lines		
Human: HeLa	Puck and Marcus, 1955	RRID: CVCL_0030
Human: SH-SY5Y	Biedler et al., 1973	RRID: CVCL_0019
Human: iNGNs	Volker Busskamp, CRTD Dresden	
Human: iPSCs	Dr.Micha Drukker HHZ Munich	HMGU No 1 RRID: CVCL_YT30
Software and algorithms		
MaxQuant	Cox et al., 2014	https://www.maxquant.org/download_asset/maxquant/latest
Perseus	Tyanova et al., 2016	https://maxquant.net/download_asset/perseus/latest
Origin	NA	https://www.originlab.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Pavel Kielkowski (pavel.kielkowski@cup.lmu.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Mass spectrometry-based proteomics data have been deposited at ProteomeXchange and are publicly available as of the date of publication. The accession number is listed in the key resources table.

- This paper does not report original code.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

iNGN cells

Coating of petri dishes for iNGN culturing. In order to culture iNGNs, petri dishes have to be coated. For this, 15 mg/mL Geltrex LDEV-free was diluted 1:1000 in cold coating media (49.5% DMEM (1×), 49.5% F-12, 1 % Pen-Strep 100×) and was immediately added onto the petri dish. The coating volumes were as follows: p60, 3.5 mL; p100, 10 mL and p150, 25 mL. After addition of the coating media to the petri dishes, they were incubated at least one hour in the incubator at 37°C and were then ready for use.

Passaging and culturing of iNGN cells in p100 dish. First, media was removed and cells were washed with 4 mL PBS. Then, 1.5 mL TrypLE^{TM} Express was added and the mixture was incubated 7 min at 37°C. The detached cells were resuspended with 2 mL pre-warmed E7 media (49% (v/v) DMEM, 49% (v/v) F-12, 1% (v/v) Alanyl-Glutamine 100×, 77.6 nM Na_2SeO_3, 11.2 mM NaCl, 10 μg/mL hHolo-Transferrin, 10 μg/mL hInsulin) supplemented with 2 μM thiazovivin (E7+TZ) and the cell suspension was transferred to a 15 mL falcon. Before the cell suspension was spinned down 5 min at 600 rpm, the concentration of the cell suspension was determined by mixing 10 μL cell suspension with 10 μL trypan blue. Afterwards, the supernatant was removed and the cell pellet was resuspended in the calculated amount of media to obtain a concentration of 1.5 million cells in 1 mL (For seeding iNGNs for differentiation the concentration was adjusted to 2.5 million cells in 1 mL). Next, the media of the previously coated p100 dish was removed and 9 mL E7+TZ media was added and topped with 1 mL of the cell suspension. Finally, the E7+TZ media was supplemented with 10 μL TGF-beta (2.0 μg/mL) and 10 μL FGF-IS (20 μg/mL) in order to obtain E9+TZ medium. The iNGN cells were cultured overnight at 37°C and 5% CO_2, before the media was exchanged to E9 on the next day.

Differentiation of iNGNs in p100 dish. In order to differentiate iNGNs to neurons during 4 days, the expression of the two transcriptions factors Neurogenin-1 and Neurogenin-2 have to be induced by the addition of doxycycline (Dox). For this, 2.5 million cells in 10 mL E7 media containing 0.5 μg/mL Dox and 2 μM TZ (E7+Dox + TZ) were seeded in a p100 dish. After overnight incubation at 37°C in the incubator, the media was exchanged to 10 mL E7+Dox without TZ. On the next day, 10 mL fresh E7+Dox media was added onto the dish. On day 4, half of the E7+dox media was removed and 5 mL Neurobasal A media supplemented with 4% NS-21 was added. On day 5, the complete media was removed and 10 mL Neurobasal A media supplemented with 2% NS-21 and 2 mM L-alanyl-glutamine was added. On day 7, 2 mL of the media was removed and 2 mL Neurobasal A media supplemented with 2% NS-21 and 2 mM L-alanyl-glutmanine was added. Until day 10, the iNGNs were incubated at 37°C in the incubator without any further media exchange.

iPSC culture for physiological neurons differentiation. iPSCs (RRID: CVCL_YT30) were cultured as previously described in (Ayo-Martin et al., 2020). They were cultured in Matrigel Basement Membrane Matrix, LDEV-free coated plates in mTESR1 medium supplemented with 1× mTESR1 supplement. Media was changed every day. For passaging, the cells were dissociated using Accutase and the collected colonies

were resuspended in mTESR1 with 1× mTESR1 supplement and 10 μM Rock inhibitor Y-27632(2HCl) and diluted in the desired density.

Generation of NPCs

Neural progenitors were generated as previously described with modifications (Boyer et al., 2012; Klaus et al., 2019). In short, embryoid bodies were generated from iPSCs by plating colonies in suspension in neural induction medium consisting of DMEM F12 with N2 and B27 supplements (minus vitamin A). After 7 days in suspension, Embryoid bodies were plated on polyornithine and laminin coated dishes and cultured for 7 days in neural induction medium. Neural rosettes were manually picked, dissociated and plated in a new polyornithine/laminin-coated plate in neural progenitor medium (neural induction medium supplemented with bFGF at 20 ng/mL). For passaging, the cells were dissociated using Accutase and split at a maximum ratio of 1:4.

Generation of neurons

Neurons were generated following the Gunhanlar protocol (Gunhanlar et al., 2018). Briefly, NPCs were plated on poly-L-ornithine and laminin coated dishes in neural differentiation medium consisting of Neurobasal with N2, B27 supplements (minus vitamin A), minimum essential medium/non-essential amino acid and laminin, supplemented with BDNF, GDNF, ascorbic acid and dcAMP. Media was changed every 2–3 days. Young and mature neurons were collected after 5 weeks and 10 weeks in culture respectively.

iPSC culture for astrocytes differentiation

iPSCs were cultured on Geltrex LDEV-Free, Reduced Growth Factor Basement Membrane Matrix coated 6-well plates in mTESR1 medium supplemented with 1× mTESR1 supplement. Media was changed every day. For passaging, the cells were incubated with Collagenase Type IV for 5–7 minutes at 37°C. The collagenase was aspirated and fresh mTESR1 with 1× mTESR1 supplement was added to each well. A cell scraper was used to collect the cells, and they were subsequently plated on a fresh 6-well plate at the desired dilution.

Generation of GPCs and astrocytes

Glial progenitor cells and astrocytes were generated as previously described with modifications (Santos et al., 2017). Briefly, confluent iPSC cultures were dissociated with collagenase, collected with a cell scraper and then cultured in suspension to form embryoid bodies. The first 24hrs the cells were cultured in mTESR1 with 1× mTESR1 supplement and 10 μM Rock inhibitor Y-27632. For the next two weeks the cells were cultured in Astrocyte medium (AM) supplemented with 20ng/mL Noggin and 10 ng/mL PDGFAA, and an additional week with only PDGFAA. The embryoid bodies were then manually dissociated by pipetting and the resulting GPCs were plated on poly-L-ornithine and laminin-coated dishes in AM supplemented with 10ng/mL bFGF and 10ng/mL EGF. Astrocytes were differentiated from GPCs in AM supplemented with 10ng/mL LIF. Media was changed every other day. The GPCs and astrocytes were collected after 5 weeks and 9 weeks of differentiation respectively. The images were obtained on a LSM710 laser-scanning confocal (Carl Zeiss microscope, ZEN software) at ×40 magnification.

Culturing of HeLa and SH-SY5Y cells

HeLa (RRID: CVCL_0030) and SH-SY5Y (RRID: CVCL_0019) cells were cultured in Dulbeccos Modified Eagles Medium – high glucose (DMEM) supplemented with 10% fetal calf serum (FCS) and 2 mM L-alanyl-glutamine at 37°C and 5% CO_2 atmosphere.

METHOD DETAILS

Fluorescence imaging

SH-SY5Y neuroblastoma cells were seeded on glass coverslips and grown in DMEM supplemented with 10% FCS. For Click staining of AMPylated proteins, cells were treated with 100 μM pro-N6pA for 24 h. After washing twice in cold PBS, the cells were fixed with 4% paraformaldehyde (PFA) in PBS for 20 min, washed two more times in PBS and permeabilized by incubation with 0.1% Triton X-100 in PBS for 5 min. The metabolically labelled proteins were coupled to TAMRA via CuAAC using 1 mM $CuSO_4$, 5 mM THPTA, 10 μM TAMRA-N_3 and 100 mM sodium ascorbate in PBS for 1 h. Following washing twice, a second fixation in 4% PFA for 10 min and two more washing steps. For immunostaining, the cells were permeabilized in 0.2% saponin in PBS for 5 min, followed by quenching with 0.12% glycine and 0.2% saponin for 10 min

and incubation in blocking solution for 1 h (10% FCS and 0.2% saponin in PBS). Primary antibodies were diluted in blocking solution and incubated with the coverslips overnight at 4°C. After four washing steps in 0.2% saponin, the cells were incubated with AlexaFluor488 coupled secondary antibodies for 1 h at room temperature, washed four times in 0.2% saponin and twice in H_2O. The coverslips were finally placed in 15 μL mounting medium (167 mg/mL Mowiol 4–88, 3% glycerol, 20 mg/mL DABCO, 1 μg/mL DAPI in PBS) Microscopic images were recorded at an Olympus FV1000 confocal laser scanning microscope using a U Plan S-Apo 100× oil objective (1.40 NA).

Immunoistochemistry – physiological neurons

10 weeks old mature neurons were fixed using 4% PFA for 10 min and permeabilized with 0.3% Triton for 5 min. After fixation and permeabilization, cells were blocked with 0.1% Tween, 10% Normal Goat Serum. Primary and secondary antibodies were diluted in blocking solution. Nuclei were visualized using 0.5 mg/mL 4,6-diamidino-2-phenylindole (DAPI). Stained cells were analyzed using a Leica laser-scanning microscope.

PLD3 activity assay

For quantitative determination of PLD3 5'exonuclease activity, lysates were prepared in TRIS-lysis buffer (TBS with 1% (v/v) Triton X-100 and 1 tablet cOmplete EDTA-free protease inhibitor cocktail). After collecting the whole cell lysates as described above, lysates were diluted in a final volume of 100 μL MES reaction buffer (50 mM MES, 200 mM NaCl) to a final concentration of 50 ng/μL in a lumox multiwell 96 plate (Sarstedt). The reaction was started by addition of 100 pmol quenched FAM-ssDNA substrate (6-FAM-AC CATGACGTTC*C*T*G*-BMN-Q535 (Biomers.net) with * indicating a phosphothioate bond). After a pre-incubation period of 30 min fluorescence emission at 528 nm (following excitation at 485 nm) was measured in a microwell plate reader (SynergyHT from BioTek) from below the wells over a period of 12 h every 5 min while incubation at 37°C. For evaluation, a substrate control without lysate and a lysate control without substrate were measured together with the samples. The 5'exonuclease activity was calculated as the slope of the measured fluorescence in the samples minus both controls.

Chemical proteomics

Seeding of iNGNs for differentiation.   The identification of the AMPylation targets during neuronal differentiation was performed with four controls and four probe treated samples for each time point. In each 10 cm dish 2.5 million iNGNs in 10 mL E7+Dox+TZ media were seeded for differentiation. The media composition that is used for each day of differentiation is described in the section above, differentiation of iNGNs. One day before the cells were harvested, the samples were treated with 5 μL of 100 mM AMPylation probe (pro-N6pA) and the controls with 5 μL DMSO.

Harvesting and cell lysis.   Cells were washed twice with 2 mL PBS. Then, 500 μL lysis buffer (PBS with 1% (v/v) NP40 and 1% (w/v) sodium deoxycholate and 1 tablet protease inhibitors per 10 mL buffer)) was added and cells were scrapped into an Eppendorf tube. The cell suspension was incubated 15 min at 4°C with agitation, before cells were spinned down 10 min at 12,000 rpm and 4°C. Subsequently, the cytosolic fraction of the lysate was transferred into a new 2 mL Eppendorf tube.

Measurement of protein concentrations.   In order to measure the protein concentrations of the lysates bicinchoninic acid assay was performed. First, bovine serum albumin (BSA) standards with concentrations of 12.5, 25, 50, 100, 200 and 400 μg/mL were prepared and samples as well as controls were diluted 40 times to a total volume of 200 μL. To measure standards, samples and controls in triplicates, 50 μL of each was added to three wells of a transparent 96-well plate with flat bottom. Afterwards, 100 μL working reagent (2 μL R2 and 98 μL R1) was added to each well by a multistepper and the plate was incubated 15 min at 60°C. Then, the absorbance at 620 nm was measured by Tecan and the protein concentrations were calculated. For each replicate, except for the 24 h time point (250 μg), 400 μg of protein was used and the volumes were adjusted to a total volume of 970 μL with 0.2% SDS in PBS.

Coupling of AMPylated proteins.   In order to couple the alkyne residue of AMPylated target proteins with Biotin-PEG$_3$-N$_3$, CuAAC was used. For this reaction, 10 μL 10 mM Biotin-PEG$_3$-N$_3$, 10 μL 100 mM TCEP and 1.2 μL 83.5 mM TBTA were added to 970 μL of each lysate. The mixture was vortexed and spinned down before 20 μL 50 mM $CuSO_4$ was added to initiate the reaction. Finally, the reaction mixture was incubated 1.5 h shaking at 600 rpm and 25°C in the dark.

Protein precipitation.   First, each reaction mixture of the previously performed click reaction was transferred to a 15 mL falcon. Then, 4 mL acetone was added to each falcon in order to precipitate the proteins. After 1 h of incubation at −20°C, proteins were spinned down 15 min at 11,000 rpm and 4°C. Supernatant was discarded and each pellet was resuspended in 1 mL methanol by sonicating 3 times for 5 s at 20% intensity. Subsequently, each suspension was transferred to a 1.5 mL Eppendorf tube and was centrifuged 10 min at 11,000 rpm and 4°C. Each pellet was washed again with 1 mL methanol before it was dissolved in 1 mL 0.2% SDS in PBS by sonicating 3 times for 5 s at 20% intensity.

Avidin beads enrichment.   In order to enrich the biotinylated proteins, avidin agarose beads (Sigma-Aldrich) were used. First, avidin beads were thawed on ice before 50 µL avidin bead suspension for each sample or control was washed 3 times with 1 mL 0.2% SDS in PBS to equilibrate the beads. After each addition of washing solution, the Eppendorf tube was carefully inverted 10 times, the suspension was centrifuged 2 min at 2000 rpm at room temperature and the supernatant was discarded. Subsequently, each dissolved protein pellet was spinned down at maximum speed for 2 min and room temperature before the supernatant was added to the equilibrated beads. After each avidin bead suspension was incubated 1 h under continuous mixing at room temperature, the beads of each sample or control were washed 3 times with 1 mL 0.2% SDS in PBS, 2 times with 1 mL 6 M urea in H_2O and 3 times with 1 mL PBS.

On beads digest of enriched proteins.   In order to prepare defined peptide fragments for the following MS-measurements, the enriched proteins were digested with trypsin. First, the washed avidin beads were resuspended in 200 µL Xbuffer (7 M urea, 2 M thiourea in 20 mM HEPES pH 7.5). Then, 0.2 µL 1 M DTT was added to reduce the disulfide bonds. Afterwards, each mixture was vortexed and incubated 45 min at room temperature shaking at 600 rpm. Next, 2 µL 550 mM IAA was added to alkylate the cysteine residues. After each mixture was vortexed and incubated 30 min at room temperature shaking at 600 rpm in the dark, 0.8 µL 1 M DTT was added to quench the alkylation. Subsequently, each mixture was vortexed and incubated for another 30 min at room temperature shaking at 600 rpm in the dark before 600 µL 50 mM TEAB was added to increase the pH value to 8. Afterward, 1.5 µL trypsin (0.5 µg/µL in 50 mM acetic acid) was added to digest the enriched proteins. Finally, each mixture was vortexed and incubated overnight at 37°C shaking at 600 rpm.

On the next day, 4 µL FA was added to stop the digest. Subsequently, each mixture was vortexed and centrifuged 1 min at 2000 rpm before the supernatant was transferred into a new Eppendorf tube. After 50 µL 0.1% FA was added to the avidin beads, each mixture was vortexed, the centrifugation step was repeated and the supernatant was again transferred to the new Eppendorf tube. Then, 50 µL 0.1% FA was added once more to the avidin beads and each mixture was vortexed. Finally, each mixture was centrifuged 3 min at 13,000 rpm and the supernatant was transferred to the new Eppendorf tube as before. For the following desalting step of the digested proteins, the pH was checked to be below 3.

Desalting of peptides mixture.   In order to remove all disturbing salt from the digested proteins, Sep Pak C18 cartridges (50 mg columns, waters) were used. First, cartridges were washed with 1 mL ACN and 1 mL 80% ACN with 0.5% FA. Then, cartridges were equilibrated with 3 mL 0.5% FA before the acidified samples were loaded slowly. Afterwards, cartridges were washed with 3 mL 0.5% FA. The desalted peptides were eluted 2 times with 250 µL 80% ACN with 0.5% FA into a LoBind Eppendorf tube. Finally, the eluates were lyophilized.

Final peptide preparation.   First, 30 µL 1% FA was added to the lyophilized peptides. Then, the mixture was vortexed, spinned down and sonicated 15 min to dissolve the complete peptides. Afterwards, the dissolved peptides were spinned down and added onto centrifugal filter units. Finally, the solution was centrifuged 1 min at 13,000 rpm and the filtrate was transferred into plastic MS vials.

MS-measurement.   MS-measurements were performed on a Q Exactive HF mass spectrometer (Thermo Fisher Scientific) coupled to an UltiMate™ 3000 Nano HPLC (Thermo Fisher Scientific) via an EASY-Spray source (Thermo Fisher Scientific). First, peptides were loaded on a Acclaim PepMap 100 µ-precolumn cartridge (5 µm, 100 Å, 300 µm ID × 5 mm, Thermo Fisher Scientific). Then, peptides were separated at 40°C on a PicoTip emitter (noncoated, 15 cm, 75 µm ID, 8 µm tip, New Objective) that was in house packed with Reprosil-Pur 120 C18-AQ material (1.9 µm, 120 Å, Dr. A. Maisch GmbH). The gradient was run from 1-36% ACN supplemented with 0.1% FA during a 120 min method (0–5 min 1%; 5-8 min to 6%; 8-98 min to 36%; 98-100 min to 85%; 100–105 min wash with 85%; 105-110 min to 1%, 110–120 min with 1%) at a flow rate of

200 nL/min. For measurements of chemical-proteomic samples on Q Exactive HF mass spectrometer, the following settings were used: The Q Exactive HF was operated in dd-MS^2 mode with the following settings: Polarity: positive; MS^1 resolution: 120 k; MS^1 AGC target: 3 × 10^6 charges; MS^1 maximum IT: 20 ms; MS1 scan range: m/z 300–1750; MS^2 resolution: 15 k; MS^2 AGC target: 2 × 10^5 charges; MS^2 maximum IT: 100 ms; Top N: 20; isolation window: m/z 1.6; isolation offset: m/z 0.2; HCD stepped normalised collision energy: 28%; intensity threshold: 5 × 10^4 counts; charge exclusion: unassigned, 1, 7, 8, >8; peptide match: off; exclude isotopes: on; dynamic exclusion: 90 s.

Full proteome analysis.   In order to quantify the proteins levels during neuronal differentiation, a full proteom analysis was performed. For this approach, the same lysates (control samples) that were used for the chemical proteomic experiments were used. First, 100 μg protein of each lysate was diluted to a total volume of 200 μL with 0.2% SDS in PBS. Then, 800 μL acetone was added and incubated 1 h at – 20°C to precipitate proteins. Next, proteins were spinned down 15 min at 11,000 rpm and 4°C and the supernatant was discarded. Each pellet was resuspended in 1 mL methanol by sonicating 3 times for 5 s at 20% intensity and the mixture was spinned down once again as before. The supernatant was discarded and the washed protein pellets were dissolved in 200 μL Xbuffer. The following digest and desalting were performed as described in the sections above. Finally, the lyophilized peptides were dissolved in 200 μL 1% FA and filtered as described in the section, final peptide preparation.

Phosphatase treatment.   In order to ensure that we only observed the separation of AMPylated and not phosphorylated proteins using a Phos-tag gel, the lysates were treated with a phosphatase prior to analysis. For the reaction, 10 μL 10× buffer (0.5 M Tris, 0.1 M MgCl$_2$, pH 9.0), 1 μL shrimp alkaline phosphatase (1000u/mL) and 100 μg lysate were mixed and filled up with H$_2$O to a total volume of 100 μL. The reaction mixture was incubated overnight at 37°C. As a positive control, the phosphorylated protein ovalbumin was used.

Western blot analysis.   For each Western blot analysis, 20 μg cell lysate was used. In order to denature proteins, 4 μL 5× Lämmli buffer (10% (w/v) SDS, 50% (v/v) glycerol, 25% (v/v) β-mercaptoethanol, 0.5% (w/v) bromphenol blue, 315 mM Tris/HCl, pH 6.8) was added to 16 μL lysate solution and the samples were boiled 5 min at 95°C. Afterwards, 20 μL of each sample was loaded onto a 7.5, 10 or 12.5% SDS gel and proteins were separated according to their size by SDS-PAGE. Then, the separated proteins were transferred onto a membrane using a blotting sandwich moistened by blotting buffer (48 mM Tris, 39 mM glycine, 0.0375% (m/v) SDS, 20% (v/v) methanol), which was composed of one extra thick blot paper, the PVDF transfer membrane, the SDS-PAGE gel and again one extra thick blot paper. Before the protein transfer was carried out 45 min at 25 V using a Semi Dry Blotter (Bio-Rad), the transfer membrane was pre-incubated 5 min in methanol. In order to block non-specific binding sites, the membrane was incubated 60 min in blocking solution (0.5 g milk powder in 10 mL PBST (PBS +0.5% Tween)). Subsequently, 10 μL primary antibody with specificity for the protein of interest was added and the mixture was incubated 1 h at room temperature. The membrane was washed 3 times for 10 min with PBST before 1 μL of the secondary HRP antibody in 10 mL blocking solution was added. After 1 h of incubation at room temperature the membrane was washed again 3 times for 10 min with PBST. Then, 400 μL ECL Substrate and 400 μL peroxide solution were mixed and added to the membrane to stain the Western blot. Finally, images of the Western blot were taken by developing machine Amersham Imager 680 (GE Healthcare).

Phos-tag gel.   In order to separate the unmodified from the post-translationally modified protein form, a Phos-tag gel was used. Compared to conventional SDS-gels, Phos-tag gels contain additionally Phos-tag reagent and MnCl$_2$. To pore a 1 mm thick 7.5% Phos-tag separating gel, 1.25 mL 30% acrylamide solution, 1.25 mL 1.5 M Tris pH 8.8 solution, 50 μL 5 mM Phos-tag reagent, 50 μL 10 mM MnCl$_2$, 50 μL 10% SDS, 2.33 mL H$_2$O, 5 μL TEMED and 25 μL 10% APS were mixed. The stacking gel was prepared as for conventional SDS-gels by mixing 0.6 mL 30% acrylamide solution, 1 mL 0.5 M Tris pH 6.8 solution, 40 μL 10% SDS, 2.34 mL H$_2$O, 4 μL TEMED and 20 μL 10% APS. After the Phos-tag gel was polymerized, 20 μg of the lysates were loaded and dependent on the size of the protein of interest, the gel was run for 2 to 3 h at 150 V. Following the separation, the gel was washed 2 times 15 min with blotting buffer (48 mM Tris, 39 mM glycine, 0.0375% (m/v) SDS, 20% (v/v) methanol) supplemented with 10 mM EDTA to remove the manganese-ions. Subsequently, the gel was washed 15 min in blotting buffer before it was blotted as described in the Western blot section.

QUANTIFICATION AND STATISTICAL ANALYSIS

Calculation chemical proteomics

MS Raw files were analysed using MaxQuant software 1.6.12.0 with the Andromeda search engine. Searches were performed against the Uniprot database for Homo sapiens (taxon identifier: 9606, March 2020). At least two unique peptides were required for protein identification. False discovery rate determination was carried out using a decoy database and thresholds were set to 1% FDR both at peptide-spectrum match and at protein levels. LFQ quantification was used as described for each sample.

Statistical analysis chemical proteomics

Statistical analysis of the MaxQuant result table proteinGroups.txt was done with Perseus 1.6.10.43.

First, LFQ intensities were $\log_2$-transformed. Afterwards, potential contaminants as well as reverse peptides were removed. Then, the rows were divided into two groups - DMSO (control) and probe treated sample (sample). Subsequently, the groups were filtered for at least three valid values out of four rows in at least one group and the missing values were replaced from normal distribution. The $-\log_{10}$(p values) were obtained by a two-sided one sample Student's t-test over replicates with the initial significance level of p = 0.05 adjustment by the multiple testing correction method of Benjamini and Hochberg (FDR = 0.05) using the volcano plot function.

Profile plots

In order to visualize the changes of the LFQ-intensities during the iNGN differentiation, profile plots were used. For this, the mean (circle), the median (line) and the whiskers for outliers from n = 4 of each time point were prepared in Origin with box defined by 25th and 75th percentile.

ABBREVIATIONS

-/-	Knockout
ºC	degree Celsius
2i	medium with PD0325901 and CHIR99021
5cadC	5-carboxy-2'-deoxycytidine
5fdC	5-formyl-2'-deoxycytidine
5hmdC	5-hydroxymethyl-2'deoxycytidine
5mdC	5-methyl-2'deoxycytidine
a2i	medium with CGP77675 and CHIR99021
ATP	Adenosine Triphosphate
BER	base excision repair
bp	base pair
C/R	medium with CHIR and IWR1
cDNA	complementary DNA
CHIR	CHIR99021
CHO-K1	Chinese Hamster Ovary K1
CpG	dCdG dinculceotide
CNS	Central Nervous System
dA	2'-deoxyadenosine
dC	2'-deoxycytidine
DNA	Deoxyribonucleic Acid
DNMT	DNA methyltransferase
dG	2'-deoxyguanine
DSB	Double Strand Break

dT	Thymidine
EDTA	Ethylenediaminetetraacetic acid
EpiSC	Epiblast Stem Cell
F-dC	2´-deoxy-2'-(R)-fluoro-cytidine
F-fdC	2´-deoxy-2'-(R)-fluoro-formylcytidine
F-mdC	2´-deoxy-2'-(R)-fluoro-methylcytidine
FBS	Foetal Bovine Serum
FGF	Fibroblast Growth Factor
FOX	Forkhead box proteins
gDNA	genomic DNA
HAT	histone acetyltransferase
HDAC	histone deacetylase
HEK	Human Embryonic Kidney
ICM	Inner Cell Mass
iNGN	Neurogenin 1 and 2-expressing human induced pluripotent stem cells
iPS	induced Pluripotent Stem Cells
JAK	Janus Kinase
JmjC	Jumonji C
K	lysine
KDM	Histone lysine demethylases
KMT	Histone lysine methyltransferase
KO	Knockout
LC-MS/MS	Liquid chromatophrapy tandem mass spectrometry
LIF	Leukemia inhibitory factor

lncRNA	long non-coding RNA
LSD	Lysine-specific demethylase 1
mESC	mouse embryonic stem cell
MeCP2	methyl CpG binding protein 2
miRNA	microRNA
mtDNA	mitochondrial DNA
mRNA	messenger RNA
NAD+	Nicotinamide Adenine Dinucleotide
OCT4	Octamer-binding transcription factor 4
PCR	Polymerase Chain Reaction
PGC	Primordial germ cells
piRNA	piwi-interacting RNA
PRC2	polycomb repressive complex 2
PTM	Post-Translational Modification
R	arginine
RNA	Ribonucleic acid
RT-qPCR	Real-Time quantitative polymerase chain reaction
SAM	S-Adenosyl methionine
siRNA	small interfering RNA
Sirt	Sirtuin
STAT	signal transducer and activator of transcription
TE	Trophoecoderm
TET	Ten-eleven translocation
TDG	Thymine DNA glycosylase

TSS   Transcriptional Start Site

αKG   alpha-Ketogluterate

μL   microlitre

REFERENCES

1. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).

2. Huang, K. & Fan, G. DNA methylation in cell differentiation and reprogramming: an emerging systematic view. *Regen. Med.* **5**, 531–544 (2010).

3. Gopisetty, G., Ramachandran, K. & Singal, R. DNA methylation and apoptosis. *Mol. Immunol.* **43**, 1729–1740 (2006).

4. Brown, S. E., Fraga, M. F., Weaver, I. C. G., Berdasco, M. & Szyf, M. Variations in DNA Methylation Patterns During the Cell Cycle of HeLa Cells. *Epigenetics* **2**, 54–65 (2007).

5. Rottach, A., Leonhardt, H. & Spada, F. DNA methylation-mediated epigenetic control. *JCB* **108**, 43–51 (2009).

6. Waddington, C. H. The epigenotype. 1942. *Int. J. Epidemiol.* **41**, 10–13 (2012).

7. Tronick, E. & Hunter, R. G. Waddington, Dynamic Systems, and Epigenetics. *Front. Behav. Neurosci.* **10**, 107 (2016).

8. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).

9. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

10. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature 1997 389:6648* **389**, 251–260 (1997).

11. Finch, J. T. *et al.* Structure of nucleosome core particles of chromatin. *Nature 1977 269:5623* **269**, 29–36 (1977).

12. Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J. Cell Biol.* **83**, 403–427 (1979).

13. Sajan, S. A. & David Hawkins, R. Methods for identifying higher-order chromatin structure. *Annu Rev Genomics Hum Genet* **13**, 59–82 (2012).

14. Mohtat, D. & Susztak, K. Fine Tuning Gene Expression: The Epigenome. *Semin. Nephrol.* **30**, 468–476 (2010).

15. Holliday, R. Epigenetics: an overview. *Dev. Genet.* **15**, 453–457 (1994).

16. Wu, C. T. & Morris, J. R. Genes, genetics, and epigenetics: a correspondence. *Science* **293**, 1103–1105 (2001).

17. Wagner, M. *et al.* Age-Dependent Levels of 5-Methyl-, 5-Hydroxymethyl-, and 5-Formylcytosine in Human and Mouse Brain Tissues. *Angew. Chem., Int. Ed.* **54**, 12511–12514 (2015).

18. Turpin, M. & Salbert, G. 5-methylcytosine turnover: Mechanisms and therapeutic implications in cancer. *Front. Mol. Biosci.* **9**, (2022).

19.   Vandiver, A. R., Idrizi, A., Rizzardi, L., Feinberg, A. P. & Hansen, K. D. DNA methylation is stable during replication and cell cycle arrest. *Sci. Rep.* **5**, 1–8 (2015).

20.   Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science (1979)* **357**, (2017).

21.   Li, H. *et al.* DNA N6-Methyladenine Modification in Eukaryotic Genome. *Front. Genet.* **13**, (2022).

22.   Ehrlich, M. & Wang, R. Y. H. 5-Methylcytosine in Eukaryotic DNA. *Science (1979)* **212**, 1350–1357 (1981).

23.   Bilyard, M. K., Becker, S. & Balasubramanian, S. Natural, modified DNA bases. *Curr. Opin. Chem. Biol.* **57**, 1–7 (2020).

24.   Tamaru, H. Confining euchromatin/heterochromatin territory: jumonji crosses the line. *Genes Dev.* **24**, 1465 (2010).

25.   Collins, L. J., Schönfeld, B. & Chen, X. S. *Handbook of Epigenetics. The New Molecular and Medical Genetics*. *Handbook of Epigenetics: The New Molecular and Medical Genetics* (Academic Press, 2011). doi:10.1016/B978-0-12-375709-8.00004-6.

26.   Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *JBC* **175**, 315–332 (1948).

27.   Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature 2008 454:7205* **454**, 766–770 (2008).

28.   Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res. Spec. Publ.* **8**, 1499 (1980).

29.   Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).

30.   Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).

31.   Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *PNAS* **90**, 11995 (1993).

32.   Han, L., Su, B., Li, W. H. & Zhao, Z. CpG island density and its correlations with genomic features in mammalian genomes. *Genome Biol* **9**, 1–12 (2008).

33.   Jeziorska, D. M. *et al.* DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. *PNAS* **114**, E7526–E7535 (2017).

34.   Antequera, F. & Bird, A. CpG islands. *EXS* **64**, 169–185 (1993).

35.   Straussman, R. *et al.* Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.* **16**, 564–571 (2009).

36.   Laird, P. W. The power and the promise of DNA methylation markers. *Nat. Rev. Cancer* **3**, 253–266 (2003).

37.   Acosta-Silva, C., Branchadell, V., Bertran, J. & Oliva, A. Mutual relationship between stacking and hydrogen bonding in DNA. Theoretical study of guanine-cytosine, guanine-5-methylcytosine, and their dimers. *J. Phys. Chem.* **114**, 10217–10227 (2010).

38. Squires, J. E. *et al.* Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.* **40**, 5023 (2012).

39. Yang, X. *et al.* 5-methylcytosine promotes mRNA export — NSUN2 as the methyltransferase and ALYREF as an m5C reader. *Cell Res.* **27**, 606–625 (2017).

40. Hussain, S. *et al.* NSun2-Mediated Cytosine-5 Methylation of Vault Noncoding RNA Determines Its Processing into Regulatory Small RNAs. *Cell Rep.* **4**, 255–261 (2013).

41. Hahn, M. A., Szabó, P. E. & Pfeifer, G. P. 5-Hydroxymethylcytosine: A stable or transient DNA modification? *Genomics* **104**, 314–323 (2014).

42. Münzel, M., Globisch, D. & Carell, T. 5-Hydroxymethylcytosine, the sixth base of the genome. *Angew. Chem. Int. Ed. Engl.* **50**, 6460–6468 (2011).

43. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem.* **6**, 1049–1055 (2014).

44. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, 1–8 (2011).

45. Pfeifer, G. P., Xiong, W., Hahn, M. A. & Jin, S. G. The role of 5-hydroxymethylcytosine in human cancer. *Cell and Tissue Research 2014 356:3* **356**, 631–641 (2014).

46. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).

47. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472 (2013).

48. Dawlaty, M. M. *et al.* Loss of Tet enzymes compromises proper differentiation of embryonic stem cells. *Dev. Cell* **29**, 102 (2014).

49. Ghosh, S., Sengupta, S. & Scaria, V. Hydroxymethyl cytosine marks in the human mitochondrial genome are dynamic in nature. *Mitochondrion* **27**, 25–31 (2016).

50. Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nat. Rev. Genet.* **13**, 7–13 (2011).

51. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* **149**, 1368–1380 (2012).

52. Globisch, D. *et al.* Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. *PLoS One* **5**, e15367 (2010).

53. Chen, Z. *et al.* Decreased 5-hydroxymethylcytosine levels correlate with cancer progression and poor survival: a systematic review and meta-analysis. *Oncotarget* **8**, 1944 (2017).

54. Bachman, M. *et al.* 5-Formylcytosine can be a stable DNA modification in mammals. *Nat. Chem. Biol.* **11**, 555–557 (2015).

55. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).

56. Song, C. X. & He, C. Potential functional roles of DNA demethylation intermediates. *Trends Biochem. Sci.* **38**, 480–484 (2013).

57. Neri, F. *et al.* Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. *Cell Rep* **10**, 674–683 (2015).

58. Pfaffeneder, T. *et al.* The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angew. Chem., Int. Ed.* **50**, 7008–7012 (2011).

59. Hong, T. *et al.* 5-Formylcytosine and 5-Carboxylcytosine Significantly Reduce the Catalytic Activity of HhaI DNA Methyltransferase. *Chin. J. Chem.* **35**, 853–856 (2017).

60. Su, M. *et al.* 5-Formylcytosine Could Be a Semipermanent Base in Specific Genome Sites. *Angew. Chem., Int. Ed.* **55**, 11797–11800 (2016).

61. Rahimoff, R. *et al.* 5-Formyl- and 5-Carboxydeoxycytidines Do Not Cause Accumulation of Harmful Repair Intermediates in Stem Cells. *J. Am. Chem. Soc.* **139**, 10359–10364 (2017).

62. Tang, Y., Zheng, S. J., Qi, C. B., Feng, Y. Q. & Yuan, B. F. Sensitive and Simultaneous Determination of 5-Methylcytosine and Its Oxidation Products in Genomic DNA by Chemical Derivatization Coupled with Liquid Chromatography-Tandem Mass Spectrometry Analysis. *Anal. Chem.* **87**, 3445–3452 (2015).

63. Song, C. X. *et al.* Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* **153**, 678–691 (2013).

64. Booth, M. J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat. Chem.* **6**, 435–440 (2014).

65. Raiber, E. A. *et al.* 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* **10**, 1258–1266 (2018).

66. Pujari, S. S. *et al.* Site-Specific 5-Formyl Cytosine Mediated DNA-Histone Cross-Links: Synthesis and Polymerase Bypass by Human DNA Polymerase η. *Angew. Chem., Int. Ed.* **60**, 26489–26494 (2021).

67. Li, F. *et al.* 5-Formylcytosine Yields DNA-Protein Cross-Links in Nucleosome Core Particles. *J. Am. Chem. Soc.* **139**, 10617–10620 (2017).

68. Ji, S. *et al.* 5-Formylcytosine mediated DNA–protein cross-links block DNA replication and induce mutations in human cells. *Nucleic Acids Res. Spec. Publ.* **46**, 6455–6469 (2018).

69. Naldiga, S. *et al.* Error-prone replication of a 5-formylcytosine-mediated DNA-peptide cross-link in human cells. *JBC* **294**, 10619–10627 (2019).

70. Ji, S. *et al.* 5-Formylcytosine-induced DNA–peptide cross-links reduce transcription efficiency, but do not cause transcription errors in human cells. *JBC* **294**, 18387–18397 (2019).

71. Ji, S. *et al.* Transcriptional Bypass of DNA-Protein and DNA-Peptide Conjugates by T7 RNA Polymerase. *ACS Chem. Biol.* **14**, 2564–2575 (2019).

72. Dubini, R. C. A., Schön, A., Müller, M., Carell, T. & Rovó, P. Impact of 5-formylcytosine on the melting kinetics of DNA by 1H NMR chemical exchange. *Nucleic Acids Res. Spec. Publ.* **48**, 8796–8807 (2020).

73. Raiber, E. A. *et al.* 5-Formylcytosine alters the structure of the DNA double helix. *Nat. Struct. Mol. Biol.* **22**, 44–49 (2015).

74. Gao, Y. *et al.* 5-Formylcytosine landscapes of human preimplantation embryos at single-cell resolution. *PLoS Biol* **18**, e3000799 (2020).

75. Runtsch, L. S., Stadlmeier, M., Schön, A., Müller, M. & Carell, T. Comparative Nucleosomal Reactivity of 5-Formyl-Uridine and 5-Formyl-Cytidine. *Chem. Eur. J.* **27**, 12747–12752 (2021).

76. Schelter, F. *et al.* 5-Hydroxymethyl-, 5-Formyl- and 5-Carboxydeoxycytidines as Oxidative Lesions and Epigenetic Marks. *Chem. Eur. J.* **27**, 8100–8104 (2021).

77. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).

78. Zhang, L. *et al.* Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat. Chem. Biol.* **8**, 328–330 (2012).

79. Maiti, A. & Drohat, A. C. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine. *JBC* **286**, 35334–35338 (2011).

80. Jessop, P. & Gering, M. Immunohistochemical Detection of 5-Hydroxymethylcytosine and 5-Carboxylcytosine in Sections of Zebrafish Embryos. *Methods mol. biol.* **2198**, 193–208 (2021).

81. Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res.* **25**, 386 (2015).

82. Abakir, A., Alenezi, F. & Ruzov, A. Analysis of 5-Carboxylcytosine Distribution Using DNA Immunoprecipitation. *Methods in Molecular Biology* **2198**, 311–319 (2021).

83. Pendina, A. A. *et al.* Genomic distribution of 5-formylcytosine and 5-carboxylcytosine in human preimplantation embryos. *Mol. Reprod. Dev.* **85**, 893–895 (2018).

84. Shibutani, T. *et al.* Guanine- 5-carboxylcytosine base pairs mimic mismatches during DNA replication. *Sci. Rep.* **4**, 1–8 (2014).

85. Kellinger, M. W. *et al.* 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **19**, 831–833 (2012).

86. Wang, L. *et al.* Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature* **523**, 621–625 (2015).

87. Konovalov, K. A. *et al.* A comprehensive mechanism for 5-carboxylcytosine-induced transcriptional pausing revealed by Markov state models. *JBC* **296**, 100735–100736 (2021).

88. Lewis, L. C. *et al.* Dynamics of 5-carboxylcytosine during hepatic differentiation: Potential general role for active demethylation by DNA repair in lineage specification. *Epigenetics* **12**, 277–286 (2017).

89. Wheldon, L. M. *et al.* Transient Accumulation of 5-Carboxylcytosine Indicates Involvement of Active Demethylation in Lineage Specification of Neural Stem Cells. *Cell Rep.* **7**, 1353–1361 (2014).

90. Zhang, L. *et al.* Validation and quantification of genomic 5-carboxylcytosine (5caC) in mouse brain tissue by liquid chromatography-tandem mass spectrometry. *Anal. Methods* **8**, 5812–5817 (2016).

91.   Kamińska, E., Korytiaková, E., Reichl, A., Müller, M. & Carell, T. Intragenomic Decarboxylation of 5-Carboxy-2'-deoxycytidine. *Angew. Chem., Int. Ed.* **60**, 23207–23211 (2021).

92.   Wang, Q. *et al.* Site-specific quantification of 5-carboxylcytosine in DNA by chemical conversion coupled with ligation-based PCR. *Chin Chem Lett.* **32**, 3426–3430 (2021).

93.   Lu, X. *et al.* Chemical modification-assisted bisulfite sequencing (CAB-seq) for 5-carboxylcytosine detection in DNA. *JACS* **135**, 9315–9317 (2013).

94.   Shin, J. H., Xu, L. & Wang, D. Mechanism of transcription-coupled DNA modification recognition. *Cell Biosci.* **7**, 1–5 (2017).

95.   Nanan, K. K. *et al.* TET-Catalyzed 5-Carboxylcytosine Promotes CTCF Binding to Suboptimal Sequences Genome-wide. *iScience* **19**, 326–339 (2019).

96.   Maiti, A., Michelson, A. Z., Armwood, C. J., Lee, J. K. & Drohat, A. C. Divergent mechanisms for enzymatic excision of 5-formylcytosine and 5-carboxylcytosine from DNA. *JACS* **135**, 15813–15822 (2013).

97.   Dubini, R. C. A. *et al.* 1H NMR Chemical Exchange Techniques Reveal Local and Global Effects of Oxidized Cytosine Derivatives. *ACS Phys. Chem. Au* **2**, 237–246 (2022).

98.   Dai, Q. *et al.* Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. *ACS Chem Biol* **11**, 470–477 (2016).

99.   Feng, Y. *et al.* Transformation of 5-Carboxylcytosine to Cytosine Through C–C Bond Cleavage in Human Cells Constitutes a Novel Pathway for DNA Demethylation. *CCS Chemistry* **3**, 994–1008 (2021).

100.   Yao, Q., Chen, Y. & Zhou, X. The roles of microRNAs in epigenetic regulation. *Curr. Opin. Chem. Biol.* **51**, 11–17 (2019).

101.   Alavian-Ghavanini, A. & Rüegg, J. Understanding Epigenetic Effects of Endocrine Disrupting Chemicals: From Mechanisms to Novel Test Methods. *Basic Clin. Pharmacol. Toxicol.* **122**, 38–45 (2018).

102.   Qu, S. *et al.* Circular RNA: A new star of noncoding RNAs. *Cancer lett.* **365**, 141–148 (2015).

103.   Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2020).

104.   Arab, K. *et al.* Long Noncoding RNA TARID Directs Demethylation and Activation of the Tumor Suppressor TCF21 via GADD45A. *Mol. Cell* **55**, 604–614 (2014).

105.   Biswas, S. & Rao, C. M. Epigenetic tools (The Writers, The Readers and The Erasers) and their implications in cancer therapy. *Eur. J. Pharmacol.* **837**, 8–24 (2018).

106.   Carter, B. & Zhao, K. The epigenetic basis of cellular heterogeneity. *Nature Reviews Genetics 2020 22:4* **22**, 235–250 (2020).

107.   Du, Q., Wang, Z. & Schramm, V. L. Human DNMT1 transition state structure. *PNAS* **113**, 2916–2921 (2016).

108.   Ciccarone, F. *et al.* Age-dependent expression of DNMT1 and DNMT3B in PBMCs from a large European population enrolled in the MARK-AGE study. *Aging Cell* **15**, 755 (2016).

109. Goto, K. *et al.* Expression of DNA methyltransferase gene in mature and immature neurons as well as proliferating cells in mice. *Differentiation* **56**, 39–44 (1994).

110. Bachmann, S. *et al.* DNA Methyltransferase 1 (DNMT1) Shapes Neuronal Activity of Human iPSC-Derived Glutamatergic Cortical Neurons. *Int. J. Mol. Sci.* **22**, 1–14 (2021).

111. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).

112. Bronner, C., Alhosin, M., Hamiche, A. & Mousli, M. Coordinated Dialogue between UHRF1 and DNMT1 to Ensure Faithful Inheritance of Methylated DNA Patterns. *Genes (Basel)* **10**, (2019).

113. Leonhardt, H., Page, A. W., Weier, H. U. & Bestor, T. H. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* **71**, 865–873 (1992).

114. Bronner, C., Alhosin, M., Hamiche, A. & Mousli, M. Coordinated Dialogue between UHRF1 and DNMT1 to Ensure Faithful Inheritance of Methylated DNA Patterns. *Genes 2019, Vol. 10, Page 65* **10**, 65 (2019).

115. Song, J., Rechkoblit, O., Bestor, T. H. & Patel, D. J. Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. *Science* **331**, 1036 (2011).

116. Liu, L. *et al.* Insufficient DNA methylation affects healthy aging and promotes age-related health problems. *Clin. Epigenetics* **2**, 349–360 (2011).

117. Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M. C. & Leonhardt, H. Recruitment of DNA methyltransferase I to DNA repair sites. *PNAS* **102**, 8905–8909 (2005).

118. Goll, M. G. *et al.* Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science (1979)* **311**, 395–398 (2006).

119. Tuorto, F. *et al.* RNA cytosine methylation by Dnmt2 and NSun2 promotes tRNA stability and protein synthesis. *Nat. Struct. Mol. Biol.* **19**, 900–905 (2012).

120. Raddatz, G. *et al.* Dnmt2-dependent methylomes lack defined DNA methylation patterns. *PNAS* **110**, 8627–8631 (2013).

121. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).

122. Yang, L., Rau, R. & Goodell, M. A. DNMT3A in haematological malignancies. *Nat. Rev. Cancer* **15**, 152–165 (2015).

123. Jeltsch, A., Adam, S., Dukatz, M., Emperle, M. & Bashtrykov, P. Deep Enzymology Studies on DNA Methyltransferases Reveal Novel Connections between Flanking Sequences and Enzyme Activity. *J. Mol. Biol.* **433**, 167186 (2021).

124. Gao, L. *et al.* Comprehensive structure-function characterization of DNMT3B and DNMT3A reveals distinctive de novo DNA methylation mechanisms. *Nat. Commun.* **11**, 1–14 (2020).

125. Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* **19**, 81–92 (2017).

126. Neri, F. *et al.* XDnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell* **155**, 121 (2013).

127. Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H. & Tajima, S. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *JBC* **279**, 27816–27823 (2004).

128. Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983–1993 (2002).

129. Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J. Is H3K4me3 instructive for transcription activation? *BioEssays* **39**, 1–12 (2017).

130. Rada-Iglesias, A. Is H3K4me1 at enhancers correlative or causative? *Nat. Genet.* **50**, 4–5 (2017).

131. Legube, G. & Trouche, D. Regulating histone acetyltransferases and deacetylases. *EMBO Rep* **4**, 944–947 (2003).

132. Ait-Si-Ali, S. *et al.* Histone acetyltransferase activity of CBP is controlled by cycle-dependent kinases and oncoprotein E1A. *Nature* **396**, 184–186 (1998).

133. Tan, L.-M. *et al.* The PEAT protein complexes are required for histone deacetylation and heterochromatin silencing. *EMBO J* **37**, e98770 (2018).

134. Chahrour, M. *et al.* MeCP2, a key contributor to neurological disease, activates and represses transcription. *Science (1979)* **320**, 1224–1229 (2008).

135. Iurlaro, M. *et al.* A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol.* **14**, 1–11 (2013).

136. Fang, Y. *et al.* The H3K36me2 methyltransferase NSD1 modulates H3K27ac at active enhancers to safeguard gene expression. *Nucleic Acids Res.* **49**, 6281–6295 (2021).

137. Murr, R. Interplay Between Different Epigenetic Modifications and Mechanisms. *Adv. Genet.* **70**, 101–141 (2010).

138. Jørgensen, S., Schotta, G. & Sørensen, C. S. Histone H4 Lysine 20 methylation: key player in epigenetic regulation of genomic integrity. *Nucleic Acids Res.* **41**, 2797–2806 (2013).

139. Sriramkumar, S. *et al.* Platinum-induced ubiquitination of phosphorylated H2AX by RING1A is mediated by replication protein A in ovarian cancer. *Mol. Cancer Res.* **18**, 1699–1710 (2020).

140. Zhu, J. K. Active DNA Demethylation Mediated by DNA Glycosylases. *https://doi.org/10.1146/annurev-genet-102108-134205* **43**, 143–166 (2009).

141. Lu, D. *et al.* Harmine-based dual inhibitors targeting histone deacetylase (HDAC) and DNA as a promising strategy for cancer therapy. *Bioorg. Chem.* **120**, 105604 (2022).

142. Agalioti, T., Chen, G. & Thanos, D. Deciphering the transcriptional histone acetylation code for a human gene. *Cell* **111**, 381–392 (2002).

143. Martin, B. J. E. *et al.* Transcription shapes genome-wide histone acetylation patterns. *Nat. Comm.* **12**, 1–9 (2021).

144. Weake, V. M. & Workman, J. L. Histone Ubiquitination: Triggering Gene Activity. *Mol. Cell* **29**, 653–663 (2008).

145. Zhu, D., Zhang, Y. & Wang, S. Histone citrullination: a new target for tumors. *Mol Cancer* **20**, 1–17 (2021).

146. Song, S. & Yu, Y. Progression on citrullination of proteins in gastrointestinal cancers. *Front. Oncol.* **9**, 15 (2019).

147. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* **152**, 1146–1159 (2013).

148. Jin, S. G. *et al.* Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. *Cell Rep.* **14**, 493–505 (2016).

149. Rondelet, G., Dal Maso, T., Willems, L. & Wouters, J. Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *J. Struct. Biol.* **194**, 357–367 (2016).

150. Dhayalan, A. *et al.* The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *JBC* **285**, 26114–26120 (2010).

151. Bröhm, A. *et al.* Methylation of recombinant mononucleosomes by DNMT3A demonstrates efficient linker DNA methylation and a role of H3K36me3. *Commun. Biol.* **5**, 1–12 (2022).

152. Guo, X. *et al.* Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. *Nature 2014 517:7536* **517**, 640–644 (2014).

153. Torres, I. O. & Fujimori, D. G. Functional coupling between writers, erasers and readers of histone and DNA methylation. *Curr. Opin. Struct. Biol.* **35**, 68–75 (2015).

154. Dhalluin, C. *et al.* Structure and ligand of a histone acetyltransferase bromodomain. *Nature 1999 399:6735* **399**, 491–496 (1999).

155. Ruthenburg, A. J. *et al.* Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions. *Cell* **145**, 692–706 (2011).

156. Musselman, C. A., Lalonde, M. E., Côté, J. & Kutateladze, T. G. Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.* **19**, 1218–1227 (2012).

157. Lee, H. S., Park, J. H., Kim, S. J., Kwon, S. J. & Kwon, J. A cooperative activation loop among SWI/SNF, gamma-H2AX and H3 acetylation for DNA double-strand break repair. *EMBO J* **29**, 1434–1445 (2010).

158. Guo, J. U., Su, Y., Zhong, C., Ming, G. L. & Song, H. Emerging roles of TET proteins and 5-hydroxymethylcytosines in active DNA demethylation and beyond. *Cell Cycle* **10**, 2662–2668 (2011).

159. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (1979)* **333**, 1300–1303 (2011).

160. Melamed, P., Yosefzon, Y., David, C., Tsukerman, A. & Pnueli, L. Tet enzymes, variants, and differential effects on function. *Front. Cell Dev. Biol.* **6**, 22 (2018).

161. Ma, L. *et al.* Tet-mediated DNA demethylation regulates specification of hematopoietic stem and progenitor cells during mammalian embryogenesis. *Sci. Adv.* **8**, 3470 (2022).

162. Charlton, J. *et al.* TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers. *Nat. Genet.* **52**, 819 (2020).

163. Parry, A., Rulands, S. & Reik, W. Active turnover of DNA methylation during cell fate decisions. *Nature Reviews Genetics 2020 22:1* **22**, 59–66 (2020).

164. Imai, S. I., Armstrong, C. M., Kaeberlein, M. & Guarente, L. Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature 2000 403:6771* **403**, 795–800 (2000).

165. Kooistra, S. M. & Helin, K. Molecular mechanisms and potential functions of histone demethylases. *Nat. Rev. Mol. Cell Biol.* **13**, 297–311 (2012).

166. Agger, K., Christensen, J., Cloos, P. A. & Helin, K. The emerging functions of histone demethylases. *Curr. Opin. Genet. Dev.* **18**, 159–168 (2008).

167. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010).

168. Capuano, F., Mülleder, M., Kok, R., Blom, H. J. & Ralser, M. Cytosine DNA methylation is found in drosophila melanogaster but absent in saccharomyces cerevisiae, schizosaccharomyces pombe, and other yeast species. *Anal. Chem.* **86**, 3697–3702 (2014).

169. Hu, C. W., Chen, J. L., Hsu, Y. W., Yen, C. C. & Chao, M. R. Trace analysis of methylated and hydroxymethylated cytosines in DNA by isotope-dilution LC-MS/MS: first evidence of DNA methylation in Caenorhabditis elegans. *Biochem. J.* **465**, 39–47 (2015).

170. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **83**, 181–188 (1989).

171. Reik, W., Dean, W. & Walter, J. Epigenetic Reprogramming in Mammalian Development. *Science (1979)* **293**, 1089–1093 (2001).

172. Miller, D., Brinkworth, M. & Iles, D. Paternal DNA packaging in spermatozoa: more than the sum of its parts? DNA, histones, protamines and epigenetics. *Reproduction* **139**, 287–301 (2010).

173. Nabel, C. S. & Kohli, R. M. Demystifying DNA Demethylation. *Science (1979)* **333**, 1229–1230 (2011).

174. Lane, N. *et al.* Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* **35**, 88–93 (2003).

175. Watanabe, D., Suetake, I., Tada, T. & Tajima, S. Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis. *Mech. Dev.* **118**, 187–190 (2002).

176. Chatterjee, A. *et al.* Placental hypomethylation is more pronounced in genomic Loci Devoid of retroelements. *G3: Genes, Genom. Genet.* **6**, 1911–1921 (2016).

177. Reik, W., Dean, W. & Walter, J. Epigenetic Reprogramming in Mammalian Development. *Science (1979)* **293**, 1089–1093 (2001).

178. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).

179. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).

180. Cadet, J. & Wagner, J. R. TET enzymatic oxidation of 5-methylcytosine, 5-hydroxymethylcytosine and 5-formylcytosine. *MRGTEM* **764–765**, 18–35 (2014).

181. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature 2011 473:7347* **473**, 343–348 (2011).

182. Bian, K. *et al.* DNA repair enzymes ALKBH2, ALKBH3, and AlkB oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine in vitro. *Nucleic Acids Res.* **47**, 5522–5529 (2019).

183. Dey, A. S. Bifunctional Role of Fe(II)/2OG-Dependent TET Family 5-Methylcytosine Dioxygenases and ALKBH2,3 in Modified Cytosine Demethylation. *BioChem* **2**, 171–181 (2022).

184. Traube, F. R., Schiffers, S. & Carell, T. Quantification of DNA Methylation and Its Oxidized Derivatives Using LC-MS. *Methods mol. biol.* **2272**, 77–94 (2021).

185. Münzel, M. *et al.* Quantification of the Sixth DNA Base Hydroxymethylcytosine in the Brain. *Angew. Chem., Int. Ed.* **49**, 5375–5377 (2010).

186. Traube, F. R. *et al.* Isotope-dilution mass spectrometry for exact quantification of noncanonical DNA nucleosides. *Nat. Protoc.* **14**, 283–312 (2018).

187. Kok, R. M. *et al.* Global DNA methylation measured by liquid chromatography-tandem mass spectrometry: Analytical technique, reference values and determinants in healthy subjects. *Clin. Chem. Lab. Med.* **45**, 903–911 (2007).

188. Carell, T. *et al.* Structure and Function of Noncanonical Nucleobases. *Angew. Chem., Int. Ed.* **51**, 7110–7131 (2012).

189. Pidugu, L. S. *et al.* Structural Basis for Excision of 5-Formylcytosine by Thymine DNA Glycosylase. *Biochemistry* **55**, 6205–6208 (2016).

190. Pidugu, L. S., Dai, Q., Malik, S. S., Pozharski, E. & Drohat, A. C. Excision of 5-Carboxylcytosine by Thymine DNA Glycosylase. *JACS* **141**, 18851–18861 (2019).

191. Deckard, C. E. & Sczepanski, J. T. Reversible chromatin condensation by the DNA repair and demethylation factor thymine DNA glycosylase. *Nucleic Acids Res.* **49**, 2450 (2021).

192. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67–79 (2011).

193. Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472 (2013).

194. Kawasaki, Y. *et al.* Active DNA demethylation is required for complete imprint erasure in primordial germ cells. *Sci. Rep.* **4**, 1–7 (2014).

195. Ohno, R. *et al.* A replication-dependent passive mechanism modulates DNA demethylation in mouse primordial germ cells. *Development* **140**, 2892–2903 (2013).

196. Smith, Z. D. & Meissner, A. The simplest explanation: passive DNA demethylation in PGCs. *EMBO J* **32**, 318–321 (2013).

197. Iwan, K. *et al.* 5-Formylcytosine to cytosine conversion by C–C bond cleavage in vivo. *Nat. Chem. Bio.* **14**, 72–78 (2017).

198. Schiesser, S. *et al.* Deamination, oxidation, and C-C bond cleavage reactivity of 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine. *JACS* **135**, 14593–14599 (2013).

199. Korytiaková, E., Kamińska, E., Müller, M. & Carell, T. Deformylation of 5-Formylcytidine in Different Cell Types. *Angew. Chem., Int. Ed.* **60**, 16869–16873 (2021).

200. Schön, A. *et al.* Analysis of an Active Deformylation Mechanism of 5-Formyl-deoxycytidine (fdC) in Stem Cells. *Angew. Chem., Int. Ed.* **59**, 5591–5594 (2020).

201. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

202. Busskamp, V. *et al.* Rapid neurogenesis through transcriptional activation in human stem cells. *Mol. Syst.* **10**, 760 (2014).

203. Zhang, Y. *et al.* Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78**, 785–798 (2013).

204. Reyes, J. H. *et al.* Glutamatergic Neuronal Differentiation of Mouse Embryonic Stem Cells after Transient Expression of Neurogenin 1 and Treatment with BDNF and GDNF: In Vitro and In Vivo Studies. *J. Neurosci.* **28**, 12622–12631 (2008).

205. Thoma, E. C. *et al.* Ectopic Expression of Neurogenin 2 Alone is Sufficient to Induce Differentiation of Embryonic Stem Cells into Mature Neurons. *PLoS One* **7**, e38651 (2012).

206. Lam, R. S., Töpfer, F. M., Wood, P. G., Busskamp, V. & Bamberg, E. Functional Maturation of Human Stem Cell-Derived Neurons in Long-Term Cultures. *PLoS One* **12**, e0169506 (2017).

207. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).

208. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**, 471–485 (2015).

209. Sridharan, R. *et al.* Role of the murine reprogramming factors in the induction of pluripotency. *Cell* **136**, 364–377 (2009).

210. Raz, R., Lee, C. K., Cannizzaro, L. A., D'Eustachio, P. & Levy, D. E. Essential role of STAT3 for embryonic stem cell pluripotency. *PNAS* **96**, 2846–2851 (1999).

211. Niwa, H., Burdon, T., Chambers, I. & Smith, A. Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev.* **12**, 2048–2060 (1998).

212. Ying, Q. L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature 2008 453:7194* **453**, 519–523 (2008).

213. Cartwright, P. *et al.* LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* **132**, 885–896 (2005).

214. Sato, N., Meijer, L., Skaltsounis, L., Greengard, P. & Brivanlou, A. H. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of Wnt signaling by a pharmacological GSK-3-specific inhibitor. *Nat. med.* **10**, 55–63 (2004).

215. Fagnocchi, L. *et al.* A Myc-driven self-reinforcing regulatory network maintains mouse embryonic stem cell identity. *Nat. Comm.* **7**, 1–17 (2016).

216. Shimizu, T. *et al.* Dual Inhibition of Src and GSK3 Maintains Mouse Embryonic Stem Cells, Whose Differentiation Is Mechanically Regulated by Src Signaling. *Stem Cells* **30**, 1394–1404 (2012).

217. Li, X. *et al.* Calcineurin-NFAT Signaling Critically Regulates Early Lineage Specification in Mouse Embryonic Stem Cells and Embryos. *Cell Stem Cell* **8**, 46–58 (2011).

218. Yagi, M., Yamanaka, S. & Yamada, Y. Epigenetic foundations of pluripotent stem cells that recapitulate in vivo pluripotency. *Lab. Invest.* **97**, 1133–1141 (2017).

219. Silva, J. & Smith, A. Capturing pluripotency. *Cell* **132**, 532–536 (2008).

220. Spada, F. *et al.* Active turnover of genomic methylcytosine in pluripotent cells. *Nat. Chem. Biol.* **16**, 1411–1419 (2020).

221. Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.* **17**, 155–169 (2016).

222. Tosolini, M. & Jouneau, A. From Naive to Primed Pluripotency: In Vitro Conversion of Mouse Embryonic Stem Cells in Epiblast Stem Cells. *Methods Mol. Biol.* **1341**, 209–216 (2016).

223. Kojima, Y. *et al.* The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak. *Cell Stem Cell* **14**, 107–120 (2014).

224. Wu, J. *et al.* An alternative pluripotent state confers interspecies chimaeric competency. *Nature* **521**, 316–321 (2015).

225. Schröder, A. S. *et al.* Synthesis of (R)-Configured 2'-Fluorinated mC, hmC, fC, and caC Phosphoramidites and Oligonucleotides. *Org. Lett.* **18**, 4368–4371 (2016).

226. Schröder, A. S. *et al.* 2'-(R)-Fluorinated mC, hmC, fC and caC triphosphates are substrates for DNA polymerases and TET-enzymes. *Chem. Comm.* **52**, 14361–14364 (2016).

227. Annesley, T. M. Ion Suppression in Mass Spectrometry. *Clin Chem* **49**, 1041–1044 (2003).

228. Feng, Y. *et al.* Direct decarboxylation of ten-eleven translocation-produced 5-carboxylcytosine in mammalian genomes forms a new mechanism for active DNA demethylation. *Chem. Sci.* **12**, 11322–11329 (2021).

229. Becker, T. *et al.* AMPylation profiling during neuronal differentiation reveals extensive variation on lysosomal proteins. *iScience* **24**, (2021).

APPENDIX

SUPPORTING INFORMATION

1. *"Analysis of an Active Deformylation Mechanism of 5-Formyl-deoxycytidine (fdC) in Stem Cells"*

2. *"Deformylation of 5-Formylcytidine in Different Cell Types"*

3. *"Intragenomic Decarboxylation of 5-Carboxy-2'-deoxycytidine"*

Supporting Information

Analysis of an Active Deformylation Mechanism of 5-Formyl-deoxycytidine (fdC) in Stem Cells

*Alexander Schön[+], Ewelina Kaminska[+], Florian Schelter[+], Eveliina Ponkkonen,
Eva Korytiaková, Sarah Schiffers, and Thomas Carell**

anie_202000414_sm_miscellaneous_information.pdf

Supporting information

Table of contents

Supplementary figures

Figure 1: Quantitative derivatization of a-fdC with MeONH$_2$ after 15 min.
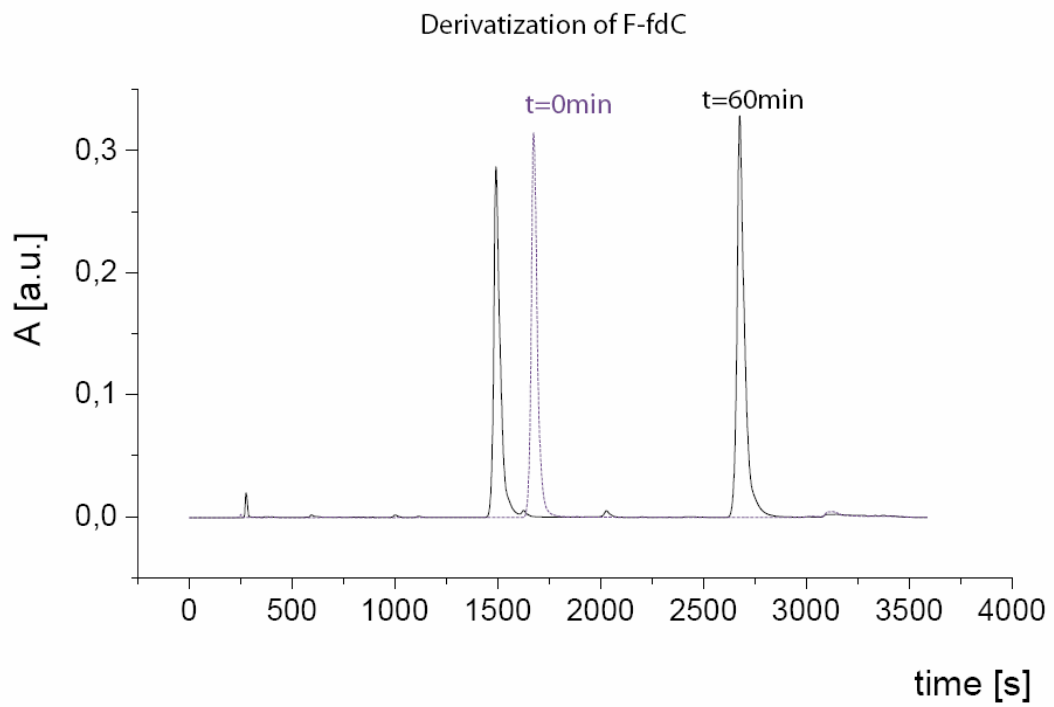
Figure 2: Derivatisation of F-fdC after 60 min, mixture of product and starting material.
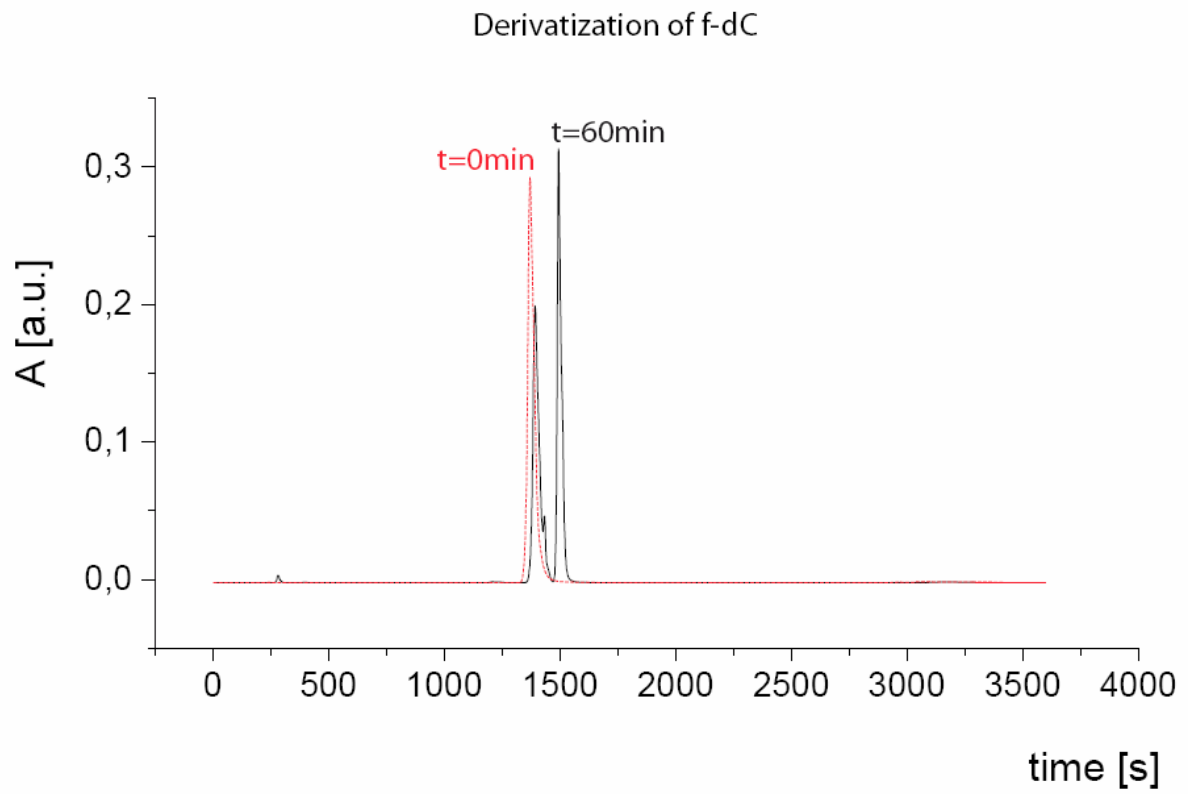
Figure 3: Derivatisation of fdC after 60 min, mixture of product and starting material.
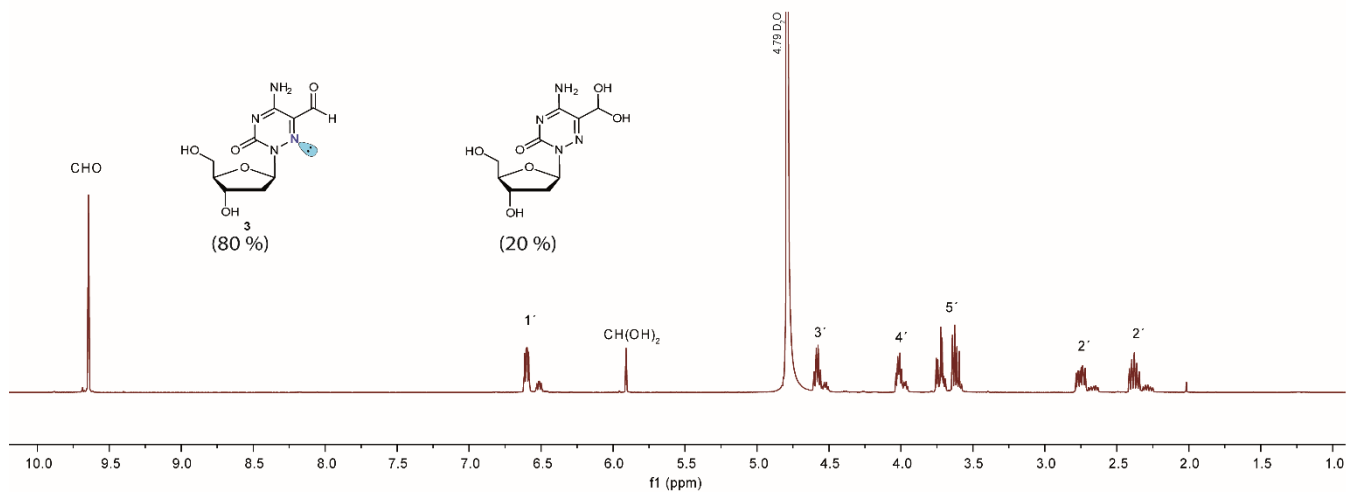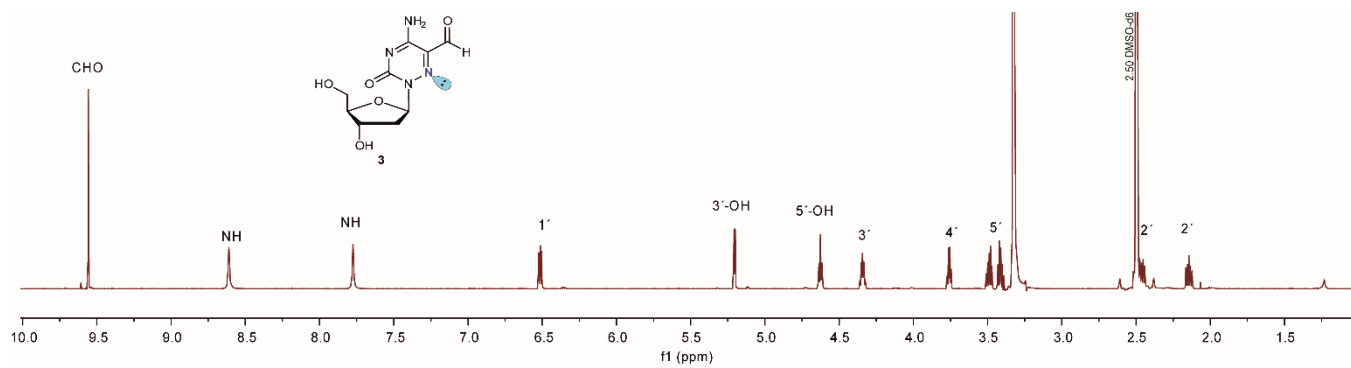
Figure 4: NMR spectra of a-fdC showing the hydrate formation of a-fdC in D₂O. Upper panel: NMR in DMSO-d6. Lower panel: NMR in D₂O.
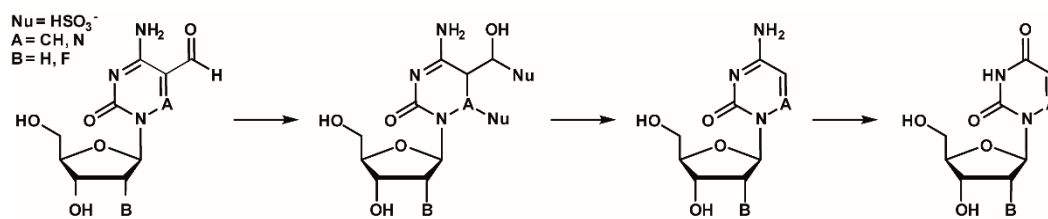
Figure 5: Proposed reaction mechanism with bisulfite

General Methods

Chemicals were purchased from *Sigma-Aldrich*, *TCI*, *Fluka*, *ABCR*, *Carbosynth* or *Acros Organics* and used without further purification. Solvents of reagent grade were purified by distillation. Reactions and column chromatography fractions were monitored by thin-layer chromatography (TLC) on silica gel F254TLC plates from *Merck KGaA*. Flash column chromatography was performed on *Geduran®* Si60 (40-63 µm) silica gel from *Merck KGaA* applying slight nitrogen pressure. Chemical transformations were conducted under nitrogen or argon atmosphere in oven-dried glassware unless otherwise specified. NMR spectra were recorded on *Bruker AVIIIHD 400* (400 MHz) or *Varian NMR-System600* (600 MHz) spectrometers. ^{1}H-NMR shifts were calibrated to the residual solvent resonances: $CDCl_3$ (7.26 ppm), DMSO-d6 (2.50 ppm) and D_2O (4.79 ppm). ^{13}C-NMR shifts were calibrated to the residual solvent: $CDCl_3$ (77.16 ppm), DMSO-d6 (39.52 ppm). All NMR spectra were analysed using the program *MestReNova* 10.0.1 from *Mestrelab Research S. L.* Low resolution mass spectra were measured on a LT Q FT-ICR by *Thermo Finnigan GmbH*. High resolution mass spectra were measured by the analytical section of the Department of Chemistry of the Ludwigs-Maximilians-Universität München on a MAT 90 (ESI) from *Thermo Finnigan GmbH*. IR spectra were recorded on a *PerkinElmer Spectrum BX II FT-IR* system. Substances were applied as a film or directly as solids on the ATR unit. Analytical RP-HPLC was performed on an analytical HPLC *Waters Alliance* (2695 Separation Module, 2996 Photodiode Array Detector) equipped with the column *Nucleosil* 120-2 C18 from *Macherey Nagel* applying eluent flow of 0.5 mL/min. Preparative RP-HPLC was performed on a HPLC *Waters Breeze* (2487 Dual λArray Detector, 1525 Binary HPLC Pump) equipped with the column VP 250/32 C18 from *Macherey Nagel*. A flowrate of 5 mL/min was applied.

Synthetic Procedures

Synthesis of 6-Aza-5-formyl-2´-deoxycytidine 3

3-Bromopyruvic acid semicarbazone **5**
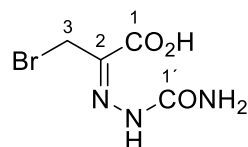
A solution of semicarbazide hydrochloride (3.34 g, 30.0 mmol, 1.0 eq.) and NaOAc (3.20 g, 38.6 mmol, 1.3 eq.) in 25.0 mL ddH$_2$O was added to a solution of 5.00 g 3-bromopyruvic acid (30.0 mmol, 1.0 eq.) in 15.0 mL glacial HOAc and 5.0 mL ddH$_2$O at 0 °C. The mixture was allowed to stand at room temperature for 2.5 h and at 0 °C for 19 h. The resulting white precipitate was filtered, washed with ice cold ddH$_2$O and Et$_2$O and dried subsequently under high vacuum to yield 3.26 g of the semicarbazone **5** as a mixture of (E)- and (Z)-isomers (14.6 mmol, 49 %) as a colorless solid.

^{1}H-NMR (400 MHz, DMSO-*d6*): δ/ppm = 10.58 (s, 0.6H, NN**H**) , 10.52 (s, 0.4H, NN**H**), 7.39 (br s, 0.8H, CON**H$_2$**), 6.92 (br s, 1.3H, CON**H$_2$**), 4.57 (s, 1.3H, 3-H), 4.46 (s, 0.8H, 3-H). **^{13}C-NMR** (101 MHz, DMSO-*d6*): δ/ppm = 163.51 (1-C), 155.63 (1´-C), 132.48 (2-C), 32.48 (3-C). **IR (ATR):** v (cm^{-1}) = 3462 (w), 3245 (w), 2394 (br w), 1894 (br w), 1693 (m), 1445 (m), 1421 (s), 1364 (m), 1297 (m), 1234 (w), 1201 (s), 1164 (m), 1143 (m), 992 (m), 821 (s), 695 (m), 658 (s).

5-Hydroxymethyl-6-azauracil **6**
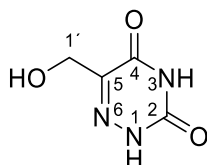
Compound **6** was synthesized according to a modified procedure of *Alekseeva et al.*[1] Under Ar atmosphere 1.40 g semicarbazone **5** (6.2 mmol, 1.0 eq.) were suspended in 50.0 mL SOCl$_2$ and two drops of pyridine were added. The reaction mixture was refluxed at 80 °C for 75 min, allowed to cool to room temperature, filtered through Celite and concentrated *in vacuo* to approximately 15 mL. Crystallization at 0 °C for 4 d resulted in a yellow precipitate which was filtered, dissolved in 30 ml dry CH$_2$Cl$_2$ and concentrated to dryness. After drying under high vacuum the obtained yellow solid was suspended in 20.0 mL ddH$_2$O and refluxed at 110 °C for 17 h. Subsequently, the solution was concentrated *in vacuo*, the residue was dissolved in 15.0 mL ddH$_2$O and lyophillized to yield 657 mg 5-hydroxy-6-azauracil **6** (4.60 mmol, 74 %) as a beige solid.

^{1}H-NMR (400 MHz, DMSO-*d6*): δ/ppm = 12.16 (s, 1H), 11.93 (s, 1H), 4.25 (s, 2H, 1´-H). **^{13}C-NMR** (101 MHz, DMSO-*d6*): δ/ppm = 156.75 (4-C), 149.55 (2-C), 144.33 (5-C), 58.02 (1´-C). **HRMS (ESI$^-$):** calculated for C$_4$H$_4$N$_3$O$_3^-$ [M-H]$^-$ 142.0258; found: 142.0258. **IR (ATR):** v (cm^{-1}) = 3429 (w), 3197 (w), 3034 (w), 2846 (w), 1682 (s), 1476 (m), 1446 (m), 1415 (m), 1244 (m), 1207 (m), 1061 (m), 1030 (m), 869 (m), 811 (m), 742 (s).

5-Hydroxymethyl-(3´,5´-di-*O*-p-toluoyl)-6-aza-2´-deoxyuridine **7**
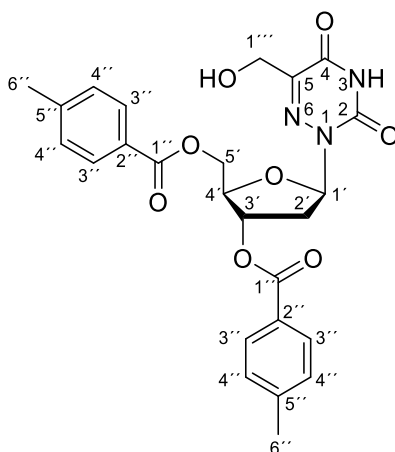
In a heat-dried round bottom flask and under argon atmosphere 290 mg 5-hydroxymethyl-6-azauracil **6** (2.03 mmol, 1.0 eq.) were suspended in 10.64 mL hexamethyldisilazane and 351 µL TMSCl were added. The mixture was refluxed at 135 °C for 75 min, the resulting brown solution was cooled to room temperature and concentrated *in vacuo* at 30 °C. The oily residue was dried under high vacuum for 1 h and subsequently dissolved in 6.38 mL dry CHCl$_3$. 867 mg Hoffer´s chlorosugar (2.23 mmol, 1.1 eq.) were added and the reaction mixture was stirred at room temperature for 17 h. The reaction mixture was poured into 250 mL 0.2 M aq. HCl and extracted with EtOAc (3 x 200 mL). Combined organic layers were dried over Na$_2$SO$_4$ and concentrated to dryness. Purification via silica gel column chromatography (iHex:EtOAc 3:1 → iHex:EtOAc 1:1 → EtOAc) yielded a 9:1 mixture of diastereomeric β- and α- nucleosidation products (760 mg, 1.53 mmol, 75 %) as a colorless solid. Recrystallization from absolute EtOH yielded 558 mg of the pure β-nucleoside **7** (1.13 mmol, 56%) as a colorless solid.

¹H-NMR (600 MHz, CDCl$_3$): δ/ppm = 9.52 (s, 1H, 3-H), 7.96 – 7.88 (m, 4H, 3´´-H), 7.26 – 7.19 (m, 4H, 4´´-H), 6.73 (dd, *J* = 6.7, 6.6 Hz, 1H, 1´-H), 5.65 (ddd, *J* = 6.3, 3.1, 3.0 Hz, 1H, 3´-H), 4.73 (dd, *J* = 11.5, 7.0 Hz, 1H, 5´-H), 4.61 – 4.46 (m, 4H, 4´-H, 5´-H, 1´´´-H), 2.99 (ddd, *J* = 14.3, 6.7, 6.5 Hz, 1H, 2´-H), 2.50 (ddd, *J* = 14.2, 6.8, 3.3 Hz, 1H, 2´-H), 2.42 (s, 3H, 6´´-H), 2.38 (s, 3H, 6´´-H). **¹³C-NMR** (151 MHz, CDCl$_3$): δ/ppm = 166.70 (1´´-C), 166.19 (1´´-C), 155.59 (4-C), 148.47 (2-C), 145.58 (5-C), 144.54 (5´´-C), 144.29 (5´´-C), 129.94 (3´´-C), 129.37 (4´´-C), 129.30 (4´´-C), 126.85 (2´´-C), 126.59 (2´´-C), 86.27 (1´-C), 82.74 (4´-C), 75.20 (3´-C), 64.36 (5´-C), 60.29 (1´´´-C), 35.04 (2´-C), 21.87 (6´´-C), 21.82 (6´´-C). **HRMS (ESI⁺)**: calculated for C$_{25}$H$_{26}$N$_3$O$_8$⁺ [M+H]⁺ 496.1714, found: 496.1715; calculated for C$_{25}$H$_{29}$N$_4$O$_8$⁺ [M+NH$_4$]⁺ 513.1980, found: 513.1975; calculated for C$_{25}$H$_{25}$N$_3$O$_8$Na⁺ [M+Na]⁺ 518.1534, found: 518.1529. **HRMS (ESI⁻)**: calculated for C$_{25}$H$_{24}$N$_3$O$_8$⁻ [M-H]⁻ 494.1566; found: 494.1572. **IR (ATR):** ν (cm⁻¹) = 1714 (s), 1611 (m), 1450 (w), 1273 (s), 1178 (m), 1105 (s), 1020 (w), 753 (m). **R$_f$** (iHex: EtOAc 1:1): 0.27.

5-Methylacetate-(3´,5´-di-*O*-p-toluoyl)-6-aza-2´-deoxyuridine **8**
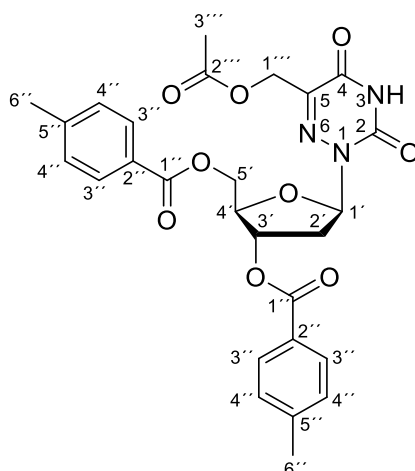
273 mg hydroxymethyl nucleoside **7** (0.55 mmol, 1.0 eq.) were dissolved in 3.30 mL dry pyridine and 220 µL acetic anhydride (2.33 mmol, 4.2 eq.) were added under argon atmosphere. The reaction mixture was stirred at room temperature for 22 h and stopped by addition of 5.0 mL absolute EtOH. The solution was concentrated *in vacuo*, coevaporated twice from dry toluene (2 x 15 mL) and the residue was purified by silica gel column chromatography ($CH_2Cl_2 \rightarrow CH_2Cl_2$:MeOH 20:1) to obtain 283 mg of acetyl protected nucleoside **8** (0.53 mmol, 96 %) as a colorless foam.

¹H-NMR (600 MHz, $CDCl_3$): δ/ppm = 8.54 (s, 1H, 3-H), 7.94 – 7.88 (m, 4H, 3´´-H), 7.26 – 7.18 (m, 4H, 4´´-H), 6.67 (dd, *J* = 6.9, 5.4 Hz, 1H, 1´-H), 5.69 (ddd, *J* = 7.4, 4.5, 3.2 Hz, 1H, 3´-H), 5.04 (d, *J* = 14.0 Hz, 1H, 1´´´-H), 4.91 (d, *J* = 14.0 Hz, 1H, 1´´´-H), 4.58 – 4.49 (m, 3H, 4´-H, 5´-H), 2.96 (ddd, *J* = 14.1, 6.8, 5.4 Hz, 1H, 2´-H), 2.51 (ddd, *J* = 14.1, 7.0, 4.5 Hz, 1H, 2´-H), 2.43 (s, 3H, 6´´-H), 2.39 (s, 3H, 6´´-H), 2.15 (s, 3H, 3´´´-H). **¹³C-NMR** (151 MHz, $CDCl_3$): δ/ppm = 170.33 (2´´´-C), 166.33 (1´´-C), 166.07 (1´´-C), 154.46 (4-C), 147.79 (2-C), 144.52 (5´´-C), 144.13 (5´´-C), 141.45 (5-C), 129.94 (3´´-C), 129.89 (3´´-C), 129.38 (4´´-C), 129.27 (4´´-C), 127.05 (2´´-C), 126.65 (2´´-C), 86.29 (1´-C), 82.54 (4´-C), 74.86 (3´-C), 64.33 (5´-C), 59.91 (1´´´-C), 35.38 (2´-C), 21.88 (6´´-C), 21.82 (6´´-C), 20.73 (3´´´-C). **HRMS (ESI⁺):** calculated for $C_{27}H_{31}N_4O_9^+$ [M+NH₄]⁺ 555.2086, found: 555.2019. **HRMS (ESI⁻):** calculated for $C_{27}H_{26}N_3O_9^-$ [M-H]⁻ 536.1675; found: 536.1677. **IR (ATR)**: ν (cm^{-1}) = 1952 (w), 1711 (s), 1610 (m), 1442 (m), 1380 (m), 1269 (s), 1177 (m), 1100 (s), 1019 (m), 839 (w), 751 (s). **R_f** (CH_2Cl_2:MeOH 20:1): 0.52.

5-Hydroxymethyl-(3´,5´-di-O-p-toluoyl)-6-aza-2´-deoxycytidine 9
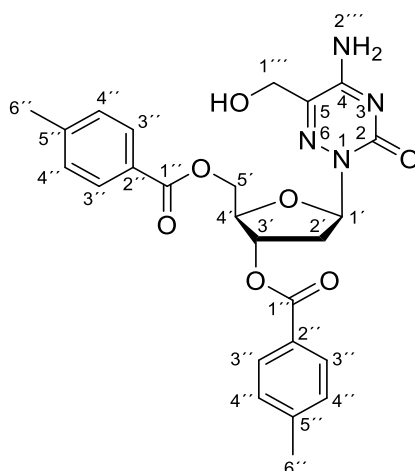
In a heat dried Schlenk-flask 487 mg 1,2,4-triazole (7.05 mmol, 9.0 eq.) were dissolved in 19.50 mL dry MeCN and cooled to 0 °C under Ar atmosphere. 146 µL POCl$_3$ (1.57 mmol, 2.0 eq.) were added dropwise and the mixture was stirred at 0 °C for 10 min. Subsequently, 966 µL NEt$_3$ (6.97 mmol, 8.9 eq.) were added and the mixture was stirred for another 20 min at 0 °C before 421 mg 2´-deoxyuridine derivative **8** (0.78 mmol, 1.0 eq.) were added. The reaction mixture was allowed to warm to room temperature and stirred for 18 h. After complete conversion the mixture was poured into 100 mL saturated aq. NaHCO$_3$ solution and extracted with CH$_2$Cl$_2$ (3 x 150 mL). Combined organic layers were dried over Na$_2$SO$_4$ and concentrated to dryness *in vacuo*.

The residue was dissolved in 15.60 mL 1,4-dioxane, 4.91 mL conc. NH$_4$OH were added and the mixture was stirred at 40 °C for 5 h. After cooling to room temperature the solution was poured into saturated aq. NH$_4$Cl solution (100 mL) and extracted with CH$_2$Cl$_2$ (3 x 100 mL). Combined organic layers were dried over Na$_2$SO$_4$, concentrated to dryness *in vacuo* and purified by silica gel column chromatography (CH$_2$Cl$_2$:MeOH 80:1 → CH$_2$Cl$_2$:MeOH 60:1 → CH$_2$Cl$_2$:MeOH 50:1 → CH$_2$Cl$_2$:MeOH 30:1 → CH$_2$Cl$_2$:MeOH 20:1 → CH$_2$Cl$_2$:MeOH 15:1) to obtain 324 mg of the 2´-deoxycytidine derivative **9** (0.66 mmol, 84 %) as an off-white solid.

¹H-NMR (600 MHz, CDCl$_3$): δ/ppm = 9.96 (s, 1H, 2´´´-H), 9.21 (s, 1H, 2´´´-H), 7.91 – 7.84 (m, 4H, 3´´-H), 7.24 – 7.16 (m, 4H, 4´´-H), 6.67 (d, *J* = 9.6 Hz, 1H, 1´-H), 5.64 (dd, *J* = 6.5, 3.4 Hz, 1H, 3´-H), 4.66 – 4.48 (m, 6H, 4´-H, 5´-H, 1´´´-H), 3.02 – 2.94 (m, 1H, 2´-H), 2.56 – 2.48 (m, 1H, 2´-H), 2.39 (s, 3H, 6´´-H), 2.36 (s, 3H, 6´´-H). **¹³C-NMR** (151 MHz, CDCl$_3$): δ/ppm = 166.45 (1´´-C), 166.15 (1´´-C), 154.50 (4-C), 144.47 (5´´-C), 144.30 (5´´-C), 137.91 (5-C), 129.94 (3´´-C), 129.87 (3´´-C), 129.42 (4´´-C), 129.36 (4´´-C), 126.91 (2´´-C), 126.63 (2´´-C), 87.22 (1´-C), 83.01 (4´-C), 75.07 (3´-C), 64.27 (5´-C), 60.98 (1´´´-C), 35.38 (6´´-C), 21.85 (6´´-C). **HRMS (ESI⁺)**: calculated for C$_{25}$H$_{27}$N$_4$O$_7^+$ [M+H]⁺ 495.1874, found: 495.1872; calculated for C$_{25}$H$_{26}$N$_4$O$_7$Na⁺ [M+Na]⁺ 517.1694, found: 517.1691. **HRMS (ESI⁻)**: calculated for C$_{25}$H$_{25}$N$_4$O$_7^-$ [M-H]⁻ 493.1729; found: 493.1734. **IR (ATR)**: ν (cm⁻¹) = 3229 (br w), 1716 (s), 1611 (m), 1450 (w), 1272 (s), 1178 (m), 1104 (s), 1020 (m), 752 (s). **R$_f$** (CH$_2$Cl$_2$:MeOH 20:1): 0.26.

5-Formyl-(3´,5´-di-*O*-p-toluoyl)-6-aza-2´-deoxycytidine **10**
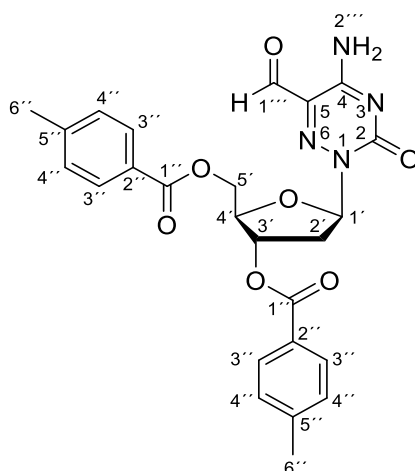
In a heat dried Schlenk flask 306 mg of 5-hydroxymethyl-2´-deoxycytdidine derivative **9** (0.62 mmol, 1.0 eq.) were dissolved in 18.6 mL dry CH_2Cl_2 and cooled to -15 °C in a NaCl/ice cooling bath. Subsequently, 289 mg Dess-Martin periodinane (0.68 mmol, 1.1 eq.) were added, the mixture was allowed to warm to room temperature and stirred for 1 h at room temperature. After complete conversion the reaction was stopped by addition of a solution of 294 mg $Na_2S_2O_3$ (1.86 mmol, 3.0 eq.) in 100 mL saturated aq. $NaHCO_3$ and extracted with with EtOAc (3 x 150 mL). Combined organic layers were dried over Na_2SO_4, concentrated to dryness *in vacuo* and purified by silica gel column chromatography (CH_2Cl_2 → CH_2Cl_2:MeOH 40:1) to yield 272 mg of aldehyde **10** (0.55 mmol, 89 %) as a yellow foam.

¹H-NMR (600 MHz, $CDCl_3$): δ/ppm = 9.40 (s, 1H, 1´´´-H), 8.31 (s, 1H, 2´´´-H), 7.94 – 7.86 (m, 4H, 3´´-H), 7.28 – 7.20 (m, 4H, 4´´-H), 6.87 (dd, *J* = 5.9 Hz, 1H, 1´-H), 5.74 (dd, *J* = 5.9, 2.9 Hz, 1H, 3´-H), 4.72 – 4.52 (m, 3H, 4´-H, 5´-H), 3.01 – 2.94 (m, 1H, 2´-H), 2.66 – 2.59 (m, 1H, 2´-H), 2.43 (s, 3H, 6´´-H), 2.41 (s, 3H, 6´´-H). **¹³C-NMR** (151 MHz, $CDCl_3$): δ/ppm = 188.81 (1´´´-C), 166.26 (1´´-C), 166.17 (1´´-C), 155,17 (4-C), 144.69 (5´´-C), 144.43 (5´´-C), 129.97 (3´´-C), 129.83 (3´´-C), 129.44 (4´´-C), 129.37 (5-C), 126.93 (2´´-C), 126.48 (2´´-C), 88.08 (1´-C), 83.34 (4´-C), 74.73 (3´-C), 63.87 (5´-C), 36.14 (2´-C), 21.89 (6´´-C), 21.84 (6´´-C). **HRMS (ESI⁺)**: calculated for $C_{25}H_{25}N_4O_7^+$ [M+H]⁺ 493.1718, found: 493.1717; calculated for $C_{25}H_{24}N_4O_7Na^+$ [M+Na]⁺ 515.1537, found: 515.1534. **HRMS (ESI⁻)**: calculated for $C_{25}H_{23}N_4O_7^-$ [M-H]⁻ 491.1572; found: 491.1577. **IR (ATR)**: ṽ (cm^{-1}) = 3237 (br w), 1720 (s), 1612 (m), 1272 (s), 1178 (w), 1104 (m), 753 (m). **R_f** (CH_2Cl_2:MeOH 20:1): 0.34.

6-Aza-5-formyl-2´-deoxycytidine 3
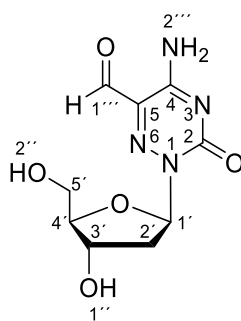
The deprotcetion was performed according to a modified procedure of *Mitchell et al.*[2] 36 mg of toluoyl protected 6-Aza-nucleoside **10** (73 µmol, 1.0 eq.) were dissolved in 0.73 mL dry MeOH and 0.73 dry benzene. To the solution 540 µL 0.5 M NaOMe in dry MeOH (270 µmol, 3.7 eq.) were added and the reaction mixture was stirred for 1.5 h at room temperature. After complete conversion the reaction mixture was neutralized with 2 M aq. HCl and evaporated to dryness. The resulting residue was dissolved in 10 mL ddH$_2$O and extracted with CH$_2$Cl$_2$ (5 x 10 mL). The aqueous layer was lyophyllized and the residue was purified by *reversed phase* HPLC (0 % → 3 % MeCN in H$_2$O in 45 min) to yield 10.1 mg of the desired product **3** (39 µmol, 54 %) as a colorless solid.

^{1}H-NMR (600 MHz, DMSO-*d6*): δ/ppm = 9.56 (s, 1H, 1´´´-H), 8.61 (s, 1H, 2´´´-H), 7.78 (s, 1H, 2´´´-H), 6.52 (dd, *J* = 7.0, 5.0 Hz, 1H, 1´-H), 5.21 (d, *J* = 4.7 Hz, 1H, 1´´-H), 4.63 (t, *J* = 5.9 Hz, 1H, 2´´-H), 4.35 (ddd, *J* = 10.2, 5.3, 5.1 Hz, 1H, 3´-H), 3.76 (dd, *J* = 5.2, 5.1 Hz, 1H, 4´-H), 3.49 (dt, *J* = 11.2, 5.5 Hz, 1H, 5´-H), 3.41 (dt, *J* = 11.8, 6.0 Hz, 1H, 5´-H), 2.47 (ddd, *J* = 13.3, 6.3, 4.9 Hz, 1H, 2´-H), 2.15 (ddd, *J* = 13.1, 7.0, 5.3 Hz, 1H, 2´-H). **^{13}C-NMR** (151 MHz, DMSO-*d6*): δ/ppm = 189.77 (1´´´-C), 156.56 (4-C), 151.66 (2-C), 129.61 (5-C), 87.87 (4´-C), 86.48 (1´-C), 70.50 (3´-C), 62.07 (5´-C), 37.74 (2´-C). **HRMS (ESI$^+$)**: calculated for C$_9$H$_{13}$N$_4$O$_5^+$ [M+H]$^+$ 257.0881, found: 257.0880; calculated for C$_9$H$_{12}$N$_4$O$_5$Na$^+$ [M+Na]$^+$ 279.0700, found: 279.0698. **HRMS (ESI$^-$)**: calculated for C$_9$H$_{11}$N$_4$O$_5^-$ [M-H]$^-$ 255.0735; found: 255.0735.

Synthesis of 6-Aza-2´-deoxycytidine
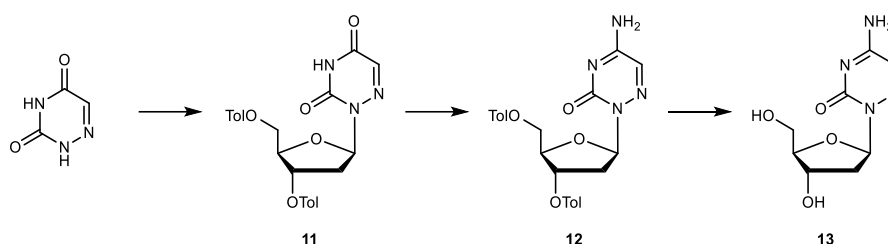
(3´,5´-Di-*O*-p-toluoyl)-6-aza-2´-deoxyuridine **11**
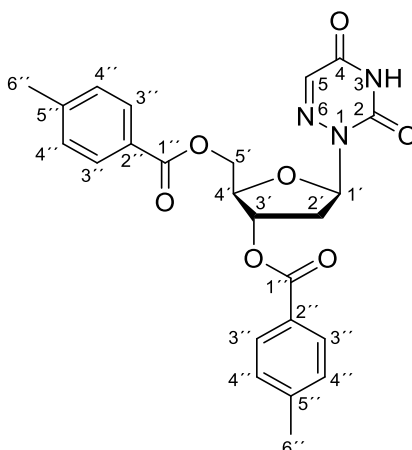
In a heat-dried round bottom flask and under argon atmosphere 300 mg 6-azauracil (2.65 mmol, 1.0 eq.) were suspended in 2.88 mL hexamethyldisilazane and 100 µL TMSCl were added. The mixture was refluxed at 145 °C for 1 h and concentrated *in vacuo* at 30 °C after cooling to room temperature. The obtained dark oil was dried under high vacuum for 1 h and subsequently dissolved in 8.33 mL dry CHCl$_3$. 1.13 g Hoffer´s chlorosugar (2.92 mmol, 1.1 eq.) were added and the reaction mixture was stirred at room temperature for 16 h. The reaction mixture was poured into 250 mL saturated aq. NaHCO$_3$ solution and extracted with EtOAc (3 x 200 mL). Combined organic layers were dried over Na$_2$SO$_4$ and concentrated to dryness. Purification via silica gel column chromatography (iHex:EtOAc 3:1 → iHex:EtOAc 1:1 → iHex:EtOAc 1:2) yielded a mixture of diastereomeric β- and α- nucleosidation products (868 mg, 1.86 mmol, 70 %) as a colorless solid. Recrystallization from absolute EtOH yielded 742 mg of the pure β-nucleoside **11** (1.59 mmol, 60%) as a colorless solid.

^{1}H-NMR (400 MHz, CDCl$_3$): δ/ppm = 8.82 (s, 1H,3-H), 7.97 – 7.91 (m, 4H, 3´´-H), 7.29 – 7.20 (m, 5H, 5-H, 4´´-H), 6.68 (dd, *J* = 6.9, 6.7 Hz, 1H, 1´-H), 5.72 (ddd, *J* = 6.8, 3.9, 3.5 Hz, 1H, 3´-H), 4.66 (dd, *J* = 11.6, 4.3 Hz, 1H, 5´-H), 4.53 (dd, *J* = 9.7, 4.4 Hz, 1H, 4´-H), 4.46 (dd, *J* = 11.6, 5.0 Hz, 1H, 5´-H), 2.98 (dt, *J* = 13.4, 6.5 Hz, 1H, 2´-H), 2.48 (ddd, *J* = 14.1, 6.8, 3.9 Hz, 1H, 2´-H), 2.43 (s, 3H, 6´´-H), 2.41 (s, 3H, 6´´-H). **^{13}C-NMR** (101 MHz, CDCl$_3$): δ/ppm = 166.36 (1´´-C), 166.16 (1´´-C), 155.30 (4-C), 147.77 (2-C), 144.56 (5´´-C), 144.11 (5´´-C), 136.10 (5-C), 129.94 (3´´-C), 129.90 (3´´-C), 129.39 (4´´-C), 129.27 (4´´-C), 127.11 (2´´-C), 126.61 (2´´-C), 86.05 (1´-C), 82.73 (4´-C), 75.03 (3´-C), 64.06 (5´-C), 35.17 (2´-C), 21.90 (6´´-C), 21.86 (6´´-C). **HRMS (ESI$^+$)**: calculated for C$_{24}$H$_{24}$N$_3$O$_7^+$ [M+H]$^+$ 466.1609, found: 466.1610; calculated for C$_{24}$H$_{27}$N$_4$O$_7^+$ [M+NH$_4$]$^+$ 483.1874, found: 483.1874. **HRMS (ESI$^-$)**: calculated for C$_{24}$H$_{22}$N$_3$O$_7^-$ [M-H]$^-$ 464.1463; found: 464.1469. **IR (ATR)**: ν (cm^{-1}) = 3200 (w), 1727 (s), 1699 (s), 1610 (w), 1437 (w), 1392 w), 1374 (w), 1329 (m), 1265 (s), 1173 (w), 1100 (s), 1080 (s), 954 (m), 808 (m), 752 (s). **R$_f$** (iHex: EtOAc 1:1): 0.44.

<u>(3´,5´-Di-*O*-p-toluoyl)-6-aza-2´-deoxycytidine **12**</u>
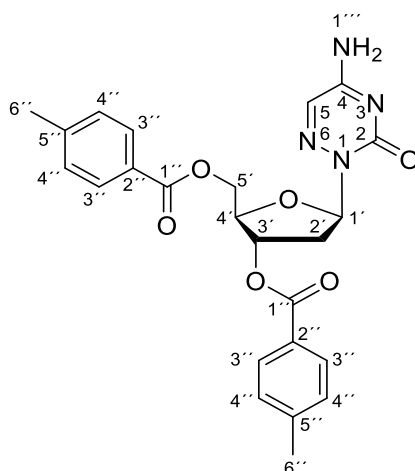
In a heat dried Schlenk-flask 668 mg 1,2,4-triazole (9.67 mmol, 9.0 eq.) were dissolved in 26.80 mL dry MeCN and cooled to 0 °C under Ar atmosphere. 201 µL POCl$_3$ (2.15 mmol, 2.0 eq.) were added dropwise and the mixture was stirred at 0 °C for 10 min. Subsequently, 1.33 mL NEt$_3$ (9.56 mmol, 8.9 eq.) were added and the mixture was stirred for another 20 min at 0 °C before 500 mg of protected 6-aza-2´-deoxyuridine **11** (1.07 mmol, 1.0 eq.) were added. The reaction mixture was allowed to warm to ambient temperature and stirred for 17 h at room temperature. After complete conversion the mixture was poured into 250 mL saturated aq. NaHCO$_3$ solution and extracted with CH$_2$Cl$_2$ (3 x 200 mL). Combined organic layers were washed with brine (150 mL), dried over Na$_2$SO$_4$ and concentrated to dryness *in vacuo*.

The residue was dissolved in 21.40 mL 1,4-dioxane and 6.74 mL conc. NH$_4$OH were added. After stirring the mixture at 45 °C for 3 h the solution was poured into saturated aq. NH$_4$Cl solution (150 mL) and extracted with EtOAc (3 x 150 mL). Combined organic layers were washed with brine (150 mL), dried over Na$_2$SO$_4$ and concentrated to dryness *in vacuo*. The crude product was purified by silica gel column chromatography (CH$_2$Cl$_2$:MeOH 40:1 → CH$_2$Cl$_2$:MeOH 20:1 → CH$_2$Cl$_2$:MeOH 10:1) to obtain 460 mg of the amination product **12** (0.99 mmol, 93 %) as a slightly yellowish solid.

¹H-NMR (600 MHz, CDCl$_3$): δ/ppm = 8.57 (s, 1H, 1´´´-H)), 7.92 (m, 4H, 3´´-H), 7.71 (s, 1H, 5-H), 7.32 (s, 1H, 1´´´-H), 7.25 – 7.16 (m, 4H, 4´´-H), 6.72 (dd, *J* = 6.6, 6.3 Hz, 1H, 1´-H), 5.73 (ddd, *J* = 7.2, 3.9, 3.1 Hz, 1H, 3´-H), 4.58 (ddd, *J* = 11.1, 5.4, 1.4 Hz, 1H, 5´-H), 4.55 – 4.52 (m, 1H, 4´-H), 4.47 (ddd, *J* = 11.2, 5.2, 1.2 Hz, 1H, 5´-H), 3.01 (ddd, *J* = 13.6, 6.8, 6.4 Hz, 1H, 2´-H), 2.45 (ddd, *J* = 13.9, 6.9, 4.2 Hz, 1H, 2´-H), 2.41 (s, 3H, 6´´-H), 2.36 (s, 3H, 6´´-H). **¹³C-NMR** (151 MHz, CDCl$_3$): δ/ppm = 166.60 (1´´-C), 166.20 (1´´-C), 159.07 (4-C), 154.24 (2-C), 144.34 (5´´-C), 143.96 (5´´-C), 129.94 (3´´-C), 129.94 (3´´-C), 129.32 (4´´-C), 129.21 (4´´-C), 128.61 (5-C), 127.10 (2´´-C), 126.84 (2´´-C), 87.76 (1´-C), 82.23 (4´-C), 75.41 (3´-C), 64.55 (5´-C), 35.17 (2´-C), 21.85 (6´´-C), 21.79 (6´´-C). **HRMS (ESI⁺)**: calculated for C$_{24}$H$_{25}$N$_4$O$_6$⁺ [M+H]⁺ 465.1769, found: 465.1770. **HRMS (ESI⁻)**: calculated for C$_{24}$H$_{23}$N$_4$O$_6$⁻ [M-H]⁻ 463.1623; found: 463.1630. **IR (ATR)**: ν (cm⁻¹) = 3305 (br w), 3059 (br w), 1715 (m), 1643 (m), 1610 (m), 1536 (w), 1466 (w), 1334 (w), 1265 (s), 1178 (m), 1102 (s) 1020 (m), 961 (w), 839 (w), 730 (s). **R$_f$** (CH$_2$Cl$_2$:MeOH 20:1): 0.30.

6-Aza-2´-deoxycytidine 13
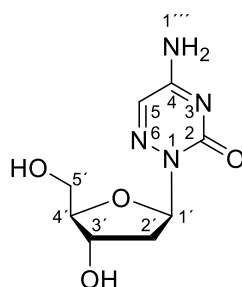
In a round bottom flask 10.0 mL of a solution of MeNH$_2$ in EtOH (33%) were added to 200 mg of toluoyl protected compound **12** (0.43 mmol, 1.0 eq.) and the mixture was stirred at room temperature for 14 h. As reaction control still showed protected starting material another 5.0 mL 33 % MeNH$_2$ in EtOH were added and the mixture was heated to 45 °C for 1 h. The solution was evaporated to dryness, dissolved in 10 mL ddH$_2$O and 10 mL CH$_2$Cl$_2$. Phases were separated, the aqueous layer was extracted with CH$_2$Cl$_2$ (4 x 10 mL) and subsequently concentrated *in vacuo*. The crude product was dissolved in ddH$_2$O (24 mL) and purified by reversed phase HPLC (0 % → 5 % MeCN in H$_2$O in 45 min) to yield 57.6 mg of the deprotected 6-Aza-nucleoside **13** as a white solid (0.25 mmol, 58 %).

¹H-NMR (600 MHz, D$_2$O): δ/ppm = 7.75 (s, 1H, 5-H), 6.57 (dd, *J* = 7.2, 4.8 Hz, 1H, 1´-H), 4.57 (ddd, *J* = 6.9, 4.4, 1.7 Hz, 1H, 3´-H), 4.05 – 4.01 (m, 1H, 4´-H), 3.85 – 3.69 (m, 1H, 5´-H), 3.71 – 3.61 (m, 1H, 5´-H), 2.72 (ddd, *J* = 13.9, 7.1, 4.8, 2.0 Hz, 1H, 2´-H), 2.35 (ddd, *J* = 14.7, 7.4, 4.0, 1.9 Hz, 1H, 2´-H). **¹³C-NMR** (151 MHz, D$_2$O): δ/ppm = 158.97 (4-C), 155.80 (2-C), 129.37 (5-C), 86.52 (1´-C), 86.44 (4´-C), 70.93 (3´-C), 61.80 (5´-C), 36.67 (2´-C). **HRMS (ESI⁻)**: calculated for C$_8$H$_{11}$N$_4$O$_4$⁻ [M-H]⁻ 227.0786; found: 227.0787.

Synthesis of 6-Aza-5-formyl-2´-deoxyuridine
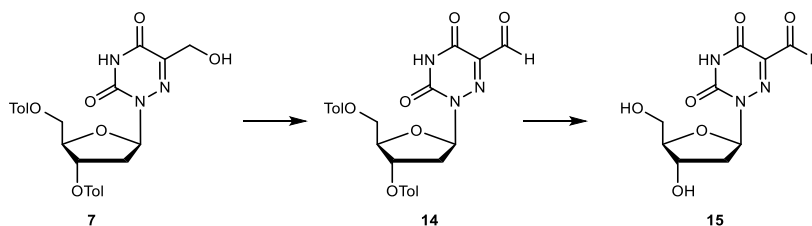
5-Formyl-(3´,5´-di-O-p-toluoyl)-6-aza-2´-deoxyuridine 14
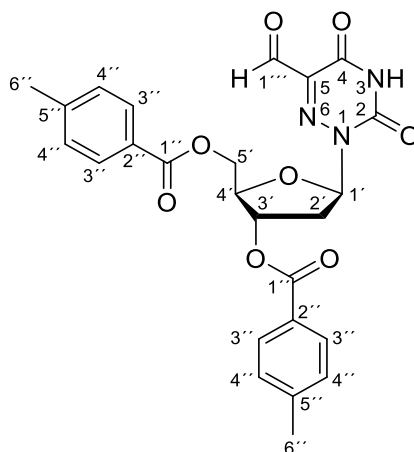
In a Schlenk flask 158 mg of 5-hydroxymethyl-2´-deoxyuridine derivative **7** (0.32 mmol, 1.0 eq.) were dissolved in 9.6 mL dry CH_2Cl_2 and cooled to -15 °C. At this temperature 162 mg Dess-Martin periodinane (0.38 mmol, 1.2 eq.) were added, the mixture was allowed to warm to room temperature slowly and was furthe stirred for 2.5 h. The reaction was stopped by addition of a solution of 182 mg $Na_2S_2O_3$ (1.15 mmol, 3.6 eq.) in 50 mL saturated aq. $NaHCO_3$ and the mixture was extracted with with EtOAc (3 x 100 mL). Combined organic layers were dried over Na_2SO_4, concentrated to dryness *in vacuo* and purified by silica gel column chromatography (iHex:EtOAc 1:1 → EtOAc) to yield 152 mg of aldehyde **14** (0.31 mmol, 96 %) as an off-white foam.

¹H-NMR (400 MHz, $CDCl_3$): δ/ppm = 9.58 (s, 1H, 1´´´-H), 9.02 (s, 1H, 3H), 8.01 – 7.79 (m, 4H, 3´´-H), 7.29 – 7.18 (m, 4H, 4´´-H), 6.70 (t, $J = 6.3$ Hz, 1H, 1´-H), 5.72 (dd, $J = 6.7, 3.4$ Hz, 1H, 3´-H), 4.70 – 4.49 (m, 3H, 4´-H, 5´-H), 3.07 (dt, $J = 14.2, 6.4$ Hz, 1H, 2´-H), 2.58 (ddd, $J = 14.2, 6.7, 4.0$ Hz, 1H, 2´-H), 2.43 (s, 3H, 6´´-H), 2.40 (s, 3H, 6´´-H). **¹³C-NMR** (101 MHz, $CDCl_3$): δ/ppm = 184.16 (1´´´-C), 166.33 (1´´-C), 166.14 (1´´-C), 152.79 (4-C), 147.15 (2-C), 144.68 (5´´-C) , 144.40 (5´´-C), 137.22 (5-C), 129.94 (3´´-C), 129.81 (3´´-C), 129.42 (4´´-C), 129.40 (4´´-C), 126.81 (2´´-C), 126.41 (2´´-C), 87.16 (1´-C), 83.20 (4´-C), 74.59 (3´-C), 63.74 (5´-C), 35.40 (2´-C), 21.91 (6´´-C), 21.85 (6´´-C). **HRMS (ESI⁺)**: calculated for $C_{25}H_{27}N_4O_8^+$ [M+NH₄]⁺ 511.1823, found:511.1823. **HRMS (ESI⁻)**: calculated for $C_{25}H_{22}N_3O_8^-$ [M-H]⁻ 492.1412; found: 492.1413. **IR (ATR)**: ν (cm⁻¹) = 2963 (w), 1712 (s), 1611 (m), 1439 (w), 1398 (w), 1309 (w), 1260 (s), 1178 (m), 10963 (s), 1020 (s), 909 (m), 800 (s), 752 (s), 732 (s). **R_f** (iHex: EtOAc 1:1): 0.17.

5-Formyl-6-aza-2´-deoxyuridine 15
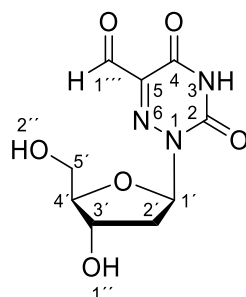
97 mg of toluoyl protected 6-Aza-nucleoside **14** (0.20 mmol, 1.0 eq.) were dissolved in 2.0 mL dry MeOH and 2.0 dry benzene. To the solution freshly prepared 0.5 M NaOMe in dry MeOH (1.46 mL, 0.73 mmol, 3.7 eq.) was added and the reaction mixture was stirred for 2 h at room temperature. Subsequently, the reaction mixture was neutralized with 2 M aq. HCl, evaporated to dryness and redissolved in 10 mL ddH$_2$O. The mixture was extracted with CH$_2$Cl$_2$ (4 x 10 mL). The aqueous layer was lyophyllized and the residue was purified by *reversed phase* HPLC (0 % $\rightarrow$ 3 % MeCN in H$_2$O in 45 min) to yield 16.3 mg of the product **15** (63 µmol, 32 %) as a colorless solid.

^{1}H-NMR (400 MHz, D$_2$O): δ/ppm = 9.62 (s, 0.2H, 1´´´-H), 6.55 (dd, *J* = 7.2, 4.5 Hz, 0.2H, 1´-H), 6.48 (dd, *J* = 7.2, 5.2 Hz, 0.8H, 1´´-H), 5.87 (s, 0.7H, 1´´´-H), 4.62 – 4.48 (m, 1H, 3´-H), 3.96 (m, 1H, 4´-H), 3.75 – 3.55 (m, 2H, 5´-H), 2.71 (m, 1H, 2´-H), 2.28 (m, 1H, 2´-H). **^{13}C-NMR** (101 MHz, D$_2$O): δ/ppm = 190.31 (1´´´-C), 165.12 (4-C), 162.87 (4-C)), 156.31 (2-C), 155.12 (2-C), 142.78, 137.09 (5-C), 87.81 (4´-C), 86.93 (4´-C), 86.48 (1´-C), 86.42 (1´-C), 86.14 (1´´´-C), 71.05 (3´-C), 70.83 (3´-C), 61.81 (5´-C), 61.70 (5´-C), 36.89 (2´-C), 36.57 (2´-C). **HRMS (ESI⁻)**: calculated for C$_9$H$_{10}$N$_3$O$_6^-$ [M-H]⁻ 256.0575; found: 256.0576.

Reactions of **1**, **2**, **3** and **15** with Methoxyamine

16   **17**   **18**   **19**
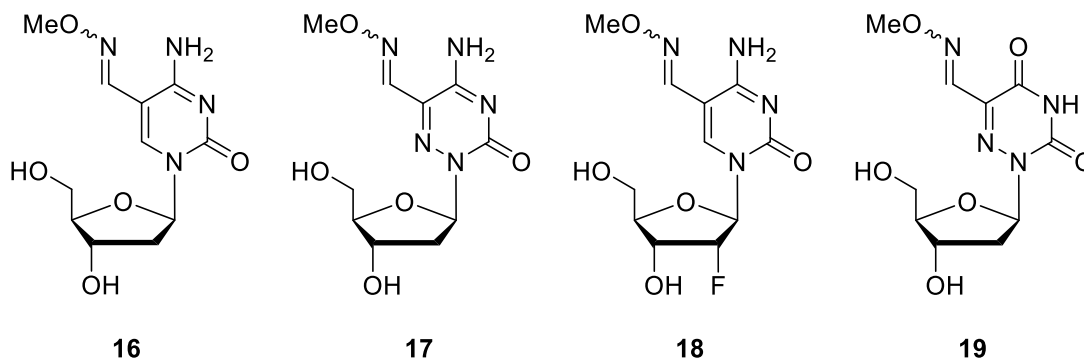
To a solution of 100 nmol nucleoside (1.0 eq.) in 35.0 µL aq. NaOH (pH = 10) was added 33.33 µL aq. MeONH$_2$ (150 mM, 50.0 eq.) and the mixture was incubated at 25 °C for 1 h. The reaction was stopped by the addition of 35.0 µL of aq. HCOOH (pH = 3) and analyzed by reversed phase HPLC. The resulting reaction products were collected and analyzed via HRMS.

	16	17	18	19
Conversion of SM	50 %	quant.	57 %	quant.
HPLC gradient [MeCN in H$_2$O in 45 min]	0 % → 15 %	0 % → 30 %	0 % → 13 %	0 % → 30 %
HRMS (calcd.) ESI$^{(+)}$	285.1194 C$_{11}$H$_{17}$O$_5$N$_4^+$ [M+H]$^+$	286.1146 C$_{10}$H$_{16}$O$_5$N$_5^+$ [M+H]$^+$	303.1099 C$_{11}$H$_{16}$O$_5$N$_4$F$^+$ [M+H]$^+$	
ESI$^{(-)}$	283.1048 C$_{11}$H$_{15}$O$_5$N$_4^-$ [M-H]$^-$	284.1000 C$_{10}$H$_{14}$O$_5$N$_5^-$ [M-H]$^-$		285.0841 C$_{10}$H$_{13}$O$_6$N$_4^-$ [M-H]$^-$
HRMS (found) ESI$^{(+)}$	285.1193 [M+H]$^+$	286.1151 [M+H]$^+$	303.1098 [M+H]$^+$	
ESI$^{(-)}$	283.1048 [M-H]$^-$	284.1000 [M-H]$^-$		285.0840 [M-H]$^-$

Reaction of **1**, **2** and **3** with NaHSO₃

20 mM nucleoside solution in ddH$_2$O was shaken with 4.36 M aq. NaHSO$_3$ (435 eq, pH = 5) at 55 °C. Samples were taken at time points of 0, 5 min, 20 min, 30 min, 40 min, 1 h, 2 h, 3 h, 4 h, 5 h, 6 h and 15 h. Bisulfite was cleaved by preparing 1 mM solution of reaction mixture and 2 M NaOH (1700 eq), shaken for 10 min at rt and followed by immediate injection to HPLC (Gradients: **1** 0% to 13% MeCN in H$_2$O in 45 min, **2** 0% to 15% MeCN in H$_2$O in 45 min, **3** 0% to 3% MeCN in H$_2$O in 45 min).

Chromatograms of each time point were normalized to the integral of the reagent and corrected accordingly. Conversion was determined by the decrease of the corrected integrals of the starting materials.

Cell culture

DMEM high glucose (Sigma Aldrich) containing 10% heat-inactivated FBS (Gibco), 100 U/mL penicillin, 100 µg/mL streptomycin, 2 mM L-glutamine were used to culture Neuro-2a and RBL-2H3 cell lines. DMEM high glucose and Ham's Nutrient Mixture F12 (Sigma Aldrich) containing 10% heat-inactivated FBS (Gibco), 100 U/mL penicillin, 100 µg/mL streptomycin, 2 mM L-glutamine were used to culture CHO-K1 cells. Metabolic-labelling experiments were conducted by plating cells in their dedicated medium containing 350 µM of F-fdC or a-fdC (in case of the co-feeding study: 350 µM of each) for three days.

DMEM high glucose (Sigma Aldrich) containing 10% FBS (PAN Biotech), 2 mM L-glutamine, 1x MEM Non-essential Amino Acid Solution and 0.1 mM β-mercaptoethanol (Sigma Aldrich) were used as basal medium for E14 TDG +/- and -/- (obtained from Cortázar *et al.)* mESC cultures. The mESC lines were maintained in naïve state on gelatin coated plates by supplementing basal medium with 1000 U/mL LIF (ORF Genetics), 3.0 µM GSK3 inhibitor CHIR99021 and 1.0 µM MNK inhibitor CGP57380 (a2i medium). Metabolic labelling experiments with isotope-labeled nucleosides were performed by plating mESCs in priming conditions consisted of basal mESC medium supplemented with 3. µM GSK3 inhibitor CHIR99021 and 2.5 µM tankyrase inhibitor IWR1-endo. Labelled nucleosides were added at the concentration of 350 µM each to the primming medium over three days.

All samples were washed with PBS (Sigma Aldrich) once befor harvesting and lysed directly in the plates by adding RLT buffer (Qiagen) supplemented with 400 µM 2,6-di-tert-butyl-4-methylphenol (BHT) and desferoxamine mesylate (DM). Next, DNA was sheered in MM400 bead mill (Retsch) at 30 Hz for 1 min in 2 mL microcentrifuge tubes (Eppendorf) with 5 mm diamiter stainless steel beads (one per tube) and centrifuged at 21000rcf for 5 minutes. Genomic DNA (gDNA) was extrated using Zymo Quick gDNA mini-prep® kit according to the protocol with an addition of RNAse A treatment (Qiagen) at 0.2 mg/mL in Genomic Lysis Buffer – 400 µL of the solution transferred directly on the column and incubated for 15 minutes. All samples were eluted in DNAse-free ddH$_2$O with 20 µM BHT, the concentration of gDNA was measured on Nanodrop.

While investigating the deformylation levels using F-fdC as a metabollic label, the following conditions were used for the gDNA digestion procedure: Degradase (1.5 µL) per up to 10 µg of F-fdC labelled gDNA in 35 µL in H$_2$O. The digestion misture was incubated at 37 °C for 4 hours. Then, the sampes were filtered using AcroPrep Advance 0.2 µm 96-well filter plate (Pall Life Sciences) prior to LC-MS/MS analysis (39 µL injection volume at 4 °C). This method was compared to NEB enzyme mix digestion (described below) and showed no

difference in the QQQ measurement outcome, thus we concluded that both can be used at equal efficiency.

Chemical labelling and spiking

Chemical labelling of a-fdC was needed to receive a sharper signal during the MS analysis. Therefore, methoxyamine was used as a derivatization reagent. We optimized the reaction conditions including time and temperature.

The quantification of the the Fluoro compounds as well as the other known nucleosides was done by spiking of heavy labelled nucleosides which were synthesized by ourselves.

For the determination of a-fdC in the gDNA samples we did an external calibration.

Analysis of labelled and unlabelled products of F-fdC and a-fdC

The analysis was performed using an UHPLC-QQQ-MS/MS system consisting of a Triple Quad™ 6490 mass spectrometer (Agilent) with an ESI source and an Agilent Infinity 1290 UHPLC. The elution was monitored at 260 nm (Agilent InfinityLab Deuterium Lamp G1314). Data Acquisition and processing were performed using MassHunter Workstation Software Version B.07.01 (Agilent).

The UHPLC separation was performed for the Fluoro-fdC as well as mdC, hmdC, fdC on an InfinityLab Poroshell 120 SB-C8 column (2.1 mm x 150 mm, 2.7 µm, Agilent Technologies, USA) at 35 °C. Water containing 0.0085% FA (v/v, solvent A) and MeCN containing 0.0085% FA (v/v, solvent B) was used as the mobile phase. A gradient of 0 - 3.5% B for 4 min, 3.5 - 5% B for 2.9 min, 5 - 80% B for 0.3 min, 80% B for 3.3 min was used. The flow rate of the mobile phase was set to 0.35 mL min^{-1}.

The derivatized a-fdC, F-fdC and fdC were separated on an InfinityLab Poroshell 120 SB-C18 column (2.1 mm x 150 mm, 2.7 µm, Agilent Technologies, USA) at 35 °C. Water containing 0.0085% FA (v/v, solvent A) and MeCN containing 0.0085% FA (v/v, solvent B) was used as the mobile phase. A gradient of 0% B for 3 min, 0 - 15% B for 27 min, 15 - 100% B for 5 min was used. The flow rate of the mobile phase was set to 0.35 mL min^{-1}.

The mass spectrometry detection was performed under positive ESI mode. The nucleosides and labelled products were monitored using the multiple reaction monitoring (MRM) mode. The MRM parameters were optimized to achieve maximal detection sensitivity (Tables 2 and 3).

Table 2 MRM parameters for the detection of F-fdC and its derivatives

Name	Transition	Scan	Type	Precursor Ion	Product Ion	Ion Polarity
UV-dG	0,0 -> 0,0	MRM	Target	0,0	0,0	Positive
UV-dC	0,0 -> 0,0	MRM	Target	0,0	0,0	Positive
UV total	0,0 -> 0,0	MRM	Target	0,0	0,0	Positive
Fluoro-fdC-dN	274,1 -> 140,1	MRM	Target	274,1	140,1	Positive
Fluoro-fdC-dN-^{15}N$_2$	276,1 -> 142,0	MRM	ISTD	276,1	142,0	Positive
Fluoro-dU	245,1 -> 225,1	MRM	Target	245,1	225,1	Negative
Fluoro-dC	246,1 -> 112,1	MRM	Target	246,1	112,1	Positive
F-fdU	273,1 -> 253,1	MRM	Target	273,1	253,1	Negative
fdC-dN-^{15}N$_2$	258,1 -> 142,0	MRM	ISTD	258,1	142,0	Positive
fdC-dN	256,1 -> 140,1	MRM	Target	256,1	140,1	Positive
^{15}N$_2$-FdC	248,1 -> 114,0	MRM	ISTD	248,1	114,0	Positive

Table 3 MRM parameters for the derivatized a-fdC samples

Name	Transition	Scan	Type	Precursor Ion	Product Ion	Ion Polarity
MeON-a-fdC	286,1 -> 170,1	MRM	Target	286,1	170,1	Positive
a-fdC	257,1 -> 141,1	MRM	Target	257,1	141,1	Positive
a-dC	229,1 -> 188,1	MRM	Target	229,1	188,1	Positive
a-dC	229,1 -> 112,1	MRM	Target	229,1	112,1	Positive

Digestion of the DNA of N2a cell line

As shown in Table 4 we used the following chemicals for the analysis of our samples. Reaction buffer 10X and Enzyme mix was bought as a Nucleoside Digestion Mix (M0649S) kit (*New England BioLabs Inc.*). The nucleosides mix was prepared by ourselve and contains heavy labelled mdC, hmdC, fdC, cadC, 8oxodG and hmdU. Furthermore we spiked heavy labelled Fluoro-dC and Fluoro-fdC for later quantification of the Fluoro-compounds. The concentration of the DNA for the digestion was received from the Nanodrop. The total volume was then incubated at 37 °C for 1.5 hours.

Table 4: Digestion sheet of the gDNA samples

	Sample	c(ng/µL)	m (DNA) [ng]	V (H2O) [µL]	V (DNA in H₂O) [µL]	V (Reaction buffer 10X)	V (Enzyme mix)	V (Nucleosides mix)	Spiking heavy Fluoro-dC	Spiking heavy Fluoro-fdC	Total Volume [µL]	1 h 30 min 37 °C	Injection volume [µL]
Blank 1	1	0	0	38,4	0,0	5,0	3,0	3,0	0,35	0,25	50,0		39
Blank 2	2	0	0	38,4	0,0	5,0	3,0	3,0	0,35	0,25	50,0		39
Control N2a Cofeeding 10 µg	3	528	10000	19,5	18,94	5,0	3,0	3,0	0,35	0,25	50,0		39
Control N2a Cofeeding 10 µg	4	528	10000	19,5	18,94	5,0	3,0	3,0	0,35	0,25	50,0		39
Control N2a Cofeeding 5 µg	5	528	5000	28,9	9,47	5,0	3,0	3,0	0,35	0,25	50,0		39
Control N2a Cofeeding 5 µg	6	528	5000	28,9	9,47	5,0	3,0	3,0	0,35	0,25	50,0		39
Sample N2a Cofeeding 10 µg	7	412	10000	14,1	24,27	5,0	3,0	3,0	0,35	0,25	50,0		39
Sample N2a Cofeeding 10 µg	8	412	10000	14,1	24,27	5,0	3,0	3,0	0,35	0,25	50,0		39
Sample N2a Cofeeding 5 µg	9	412	5000	26,3	12,14	5,0	3,0	3,0	0,35	0,25	50,0		39
Sample N2a Cofeeding 5 µg	10	412	5000	26,3	12,14	5,0	3,0	3,0	0,35	0,25	50,0		39

Derivatization of a-fdC in digested gDNA of N2a cell line

Derivatization of a-fdC from the digested DNA (5 µg) of biological samples by methoxyamine was performed under the following conditions. Briefly, the nucleoside mixture was derivatized in 50 µL H_2O with 40 µL aq. NaOH (pH 10) and 8 µL methoxyamine (163 mM in H_2O) for 45 min at 25 °C. Afterwards the solution was neutralized with 40 µL aq. FA (pH 3). The derivatized nucleosides were then lyophylized and resuspended in 50 µL H_2O.

After resuspension the samples were filtered utilizing a 0.2 µm Supor filtration plate (*Pall Corporation*) and subjected to UHPLC-QQQ-MS/MS.

The determination of F-fdC and some other nucleosides in gDNA (10 µg) was performed by using heavy labelled compounds. After the digestion the samples were filtered utilizing a 0.2 µm Supor filtration plate (*Pall Corporation*) and subjected to UHPLC-QQQ-MS/MS.

Calibration curve for external quantification of a-fdC in gDNA of N2a cell lines

The quantification of a-fdC was done by an external calibration.[3] The external calibration curve was done by serially diluting pure a-fdC (see Table 5) and measured in technical triplicates prior to each measurment. Linear regression was done by Microsoft Excel (Figure 6). The injection volume was 29 µL.

Table 5 Measured values for the external calibration curve of a-fdC

n [fmol]	Average area	standard deviation
0,11	1006	110
0,22	1822	98
0,44	3160	414
0,88	4754	428
1,77	8439	333
3,53	14692	346
7,07	28982	522
14,14	56445	1107
28,28	111343	2041
56,55	220616	2561
113,10	439962	2460
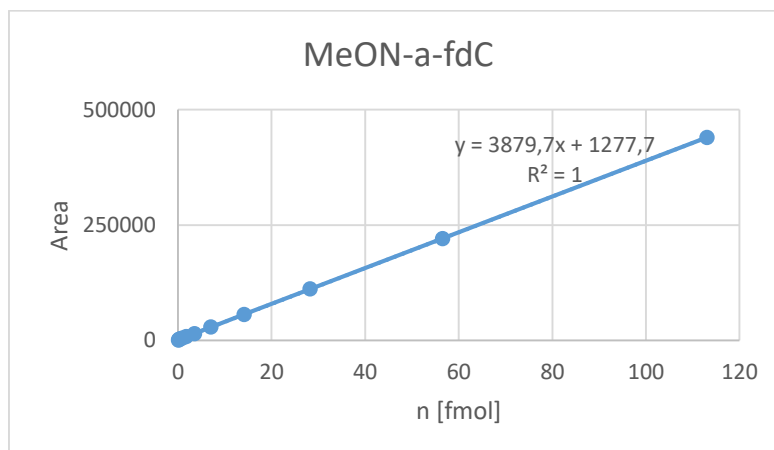
Figure 6: Calibration curve for the external quantification of a-fdC

Quantifiation of concentration of the fed compounds

The quantification of the fed compounds was performed in two technical replicates of a biological triplicate.

For the quantification of a-fdC one has to say, that there was a small background in the a-fdC peak, which was continously substracted from the a-fdC value. This value was determined by the control sample.

The quantification of the Fluoro-fdC and the deformylated Fluoro-dC was done over the ratio of unlabelled to labelled compound. Furthermore it was compared to the amount of deoxy-cytidine and deoxy-Guanosine and dN. The final results can be seen in Figures 7.
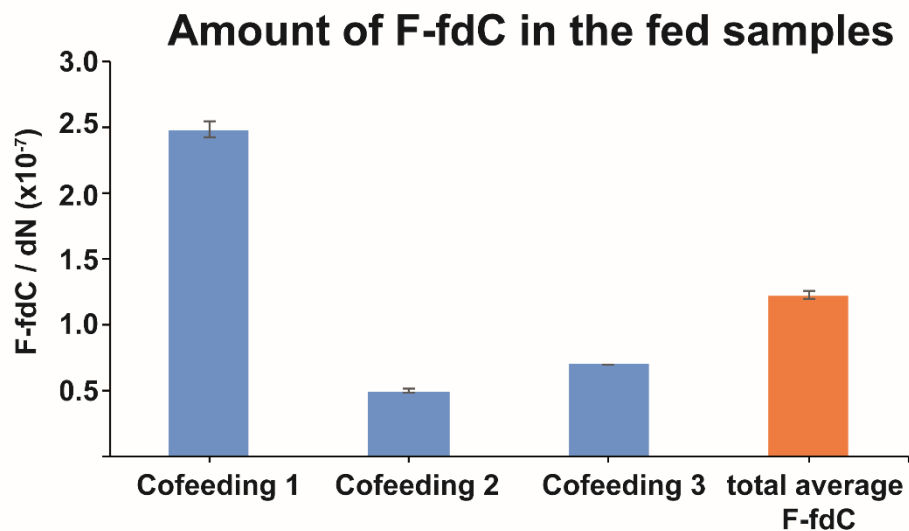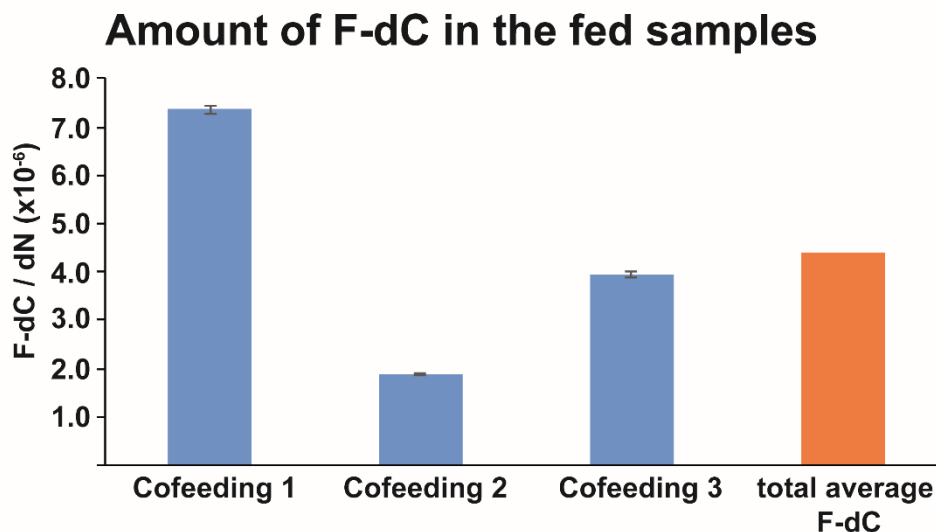
Figure 7 Measured amount of F-fdC (A) and F-dC (B) in the fed samples of N2a cells

Furthermore we were able to calculate the amount of deformylation of Fluoro-fdC by the amounts of Fluoro-fdC and Fluoro-dC. This was able to be done by the unnatural Fluoro compound. In Figure 8 the deformylation rates are shown.

Figure 8 Calculated deformylation rate of F-fdC in N2a cells

NMR spectra of synthesized compounds

¹H NMR and ¹³C NMR spectra

6

¹H NMR peaks: 12.16, 11.93, 4.25, 2.50 DMSO-d6

¹³C NMR peaks: 156.75, 149.55, 144.33, 58.02, 39.52 DMSO-d6

¹H NMR (top) and ¹³C NMR (bottom) spectra of compound **7**.

8

8

9

9.96
9.21
7.91
7.90
7.89
7.88
7.85
7.23
7.21
7.19
7.17
6.71
6.69
6.68
6.67
5.64
5.63
4.66
4.65
4.64
4.62
4.61
4.58
4.57
4.56
4.55
4.54
4.52
4.51
4.50
4.49
3.00
2.99
2.98
2.97
2.53
2.50
2.49
2.48
2.47
2.46
2.43
2.42
2.40
2.39
2.36

13 12 11 10 9 8 7 6 5 4 3 2 1 0 -1
f1 (ppm)

9

166.45
166.15
154.50
144.47
144.30
137.91
129.94
129.87
129.42
129.36
126.91
126.63
87.22
83.01
77.16 CDCl3
75.07
64.27
60.98
35.38
21.85
1.17

230 220 210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 -10
f1 (ppm)

NMR spectra of compound **10**

¹H NMR peaks: 9.41, 9.40, 9.36, 8.31, 8.01, 7.95, 7.94, 7.94, 7.93, 7.92, 7.91, 7.90, 7.89, 7.87, 7.81, 7.26, 6.87, 5.74, 5.73, 5.71, 4.71, 4.70, 4.69, 4.69, 4.67, 4.65, 4.64, 4.63, 4.62, 4.61, 4.60, 4.59, 4.58, 4.57, 4.56, 4.55, 4.54, 4.54, 4.53, 3.00, 2.98, 2.64, 2.62, 2.60, 2.59, 2.43, 2.41, 2.38

¹³C NMR peaks: 188.81, 166.26, 166.17, 144.69, 144.43, 129.97, 129.83, 129.44, 126.93, 126.48, 88.08, 83.34, 77.16 CDCl₃, 74.73, 65.76, 63.87, 62.29, 62.03, 36.14, 21.89, 21.84

^{1}H NMR spectrum (top) and ^{13}C NMR spectrum (bottom) in DMSO-d6

^{1}H NMR peaks: 9.58, 9.56, 9.54, 8.61, 7.78, 6.53, 6.52, 6.51, 5.21, 5.20, 4.64, 4.63, 4.62, 4.36, 4.35, 4.34, 4.33, 3.77, 3.77, 3.76, 3.75, 3.50, 3.49, 3.48, 3.47, 3.43, 3.42, 3.41, 3.40, 3.39, 3.33, 2.50 DMSO-d6, 2.49, 2.48, 2.47, 2.46, 2.45, 2.17, 2.16, 2.15, 2.14, 2.13

^{13}C NMR peaks: 189.77, 156.56, 151.66, 129.61, 87.87, 86.48, 70.50, 62.07, 40.06, 39.80, 39.52 DMSO-d6, 37.74

2

¹H NMR peaks: 8.86, 8.82, 8.82, 7.96, 7.95, 7.94, 7.93, 7.92, 7.91, 7.90, 7.26 CDCl3, 6.70, 6.68, 6.67, 5.73, 5.73, 5.72, 5.71, 5.70, 4.68, 4.66, 4.65, 4.64, 4.55, 4.53, 4.52, 4.51, 4.48, 4.47, 4.45, 4.44, 3.02, 3.00, 2.98, 2.97, 2.95, 2.51, 2.50, 2.50, 2.49, 2.48, 2.47, 2.46, 2.45, 2.43, 2.41

¹³C NMR peaks: 166.36, 166.16, 155.30, 147.77, 144.56, 144.11, 136.10, 129.94, 129.90, 129.39, 129.27, 127.11, 126.61, 86.05, 82.73, 77.16 CDCl3, 75.03, 64.06, 35.17, 21.90, 21.86

Top spectrum (^{1}H NMR) peak labels:
8.57, 7.94, 7.94, 7.93, 7.93, 7.92, 7.92, 7.91, 7.91, 7.71, 7.26 CDCl3, 7.25, 7.24, 7.24, 7.24, 7.23, 7.23, 7.23, 7.23, 7.19, 7.18, 7.18, 7.18, 7.17, 7.17, 7.17, 7.17, 7.17, 6.73, 6.72, 6.71, 5.74, 5.74, 5.73, 5.72, 5.72, 4.59, 4.59, 4.58, 4.58, 4.57, 4.56, 4.55, 4.54, 4.54, 4.53, 4.52, 4.52, 4.48, 4.48, 4.47, 4.46, 4.46, 4.46, 3.03, 3.02, 3.01, 3.00, 2.99, 2.47, 2.46, 2.46, 2.45, 2.45, 2.44, 2.44, 2.43, 2.41, 2.36

Bottom spectrum (^{13}C NMR) peak labels:
166.60, 166.20, 159.07, 154.24, 144.34, 143.96, 129.94, 129.94, 129.32, 129.21, 128.61, 127.10, 126.84, 87.76, 82.23, 77.16 CDCl3, 75.41, 64.55, 35.17, 31.07, 21.85, 21.79

Structure label: 12

13

13

Bibliography

[1]   I. V. Alekseeva, A. S. Shalamai, V. S. Shalamai, V. P. Chernetskii, *Ukr. Khim. Zh.* **1976**, *42*, 398-401.

[2]   W. L. Mitchell, P. Ravenscroft, M. L. Hill, L. J. S. Knutsen, B. D. Judkins, R. F. Newton, D. I. C. Scopes, *J. Med. Chem.* **1986**, *29*, 809-816.

[3]   F. R. Traube, S. Schiffers, K. Iwan, S. Kellner, F. Spada, M. Müller, T. Carell, *Nat. Protoc.* **2019**, *14*, 283-312

Supporting Information

Deformylation of 5-Formylcytidine in Different Cell Types

Eva Korytiaková+, Ewelina Kamińska+, Markus Müller, and Thomas Carell*

anie_202107089_sm_miscellaneous_information.pdf

Table of Contents

Supplementary figures

Figure SI-1: Levels of incorporated 2´-F-fdC (**1**) in the genome (left) and of the deformylation product 2´-F-dC (**2**, right) of different cells after feeding 2´F-fdC at a concentration of 350 µM. The error bards represent the propagated error based on the standard deviations of three biological replicates.
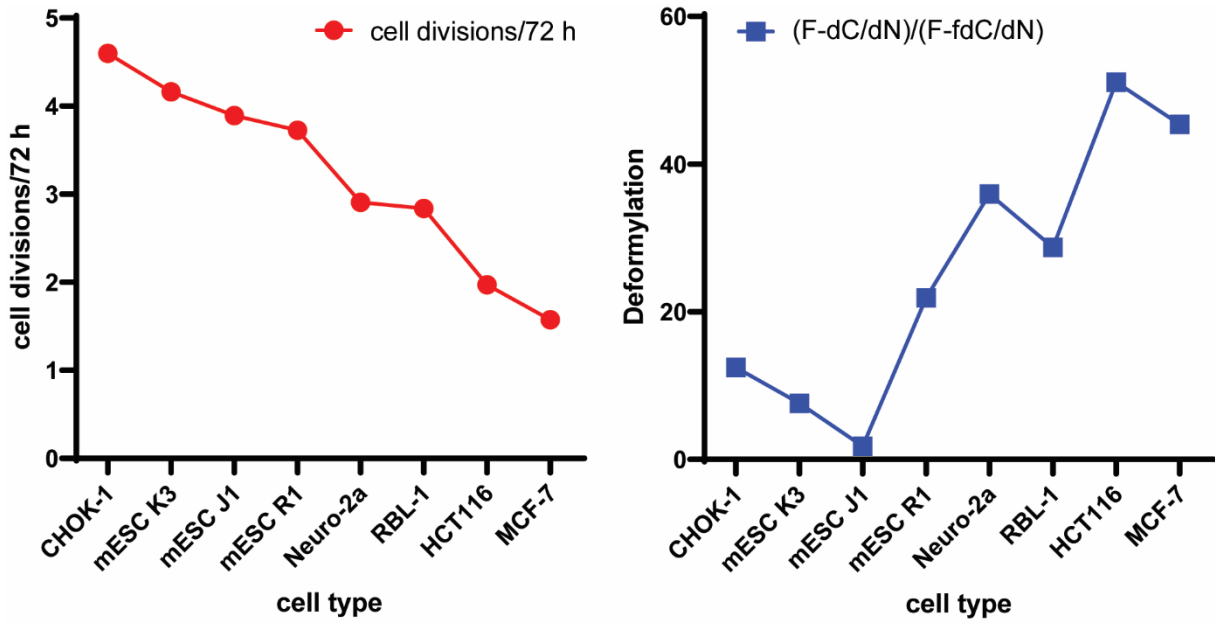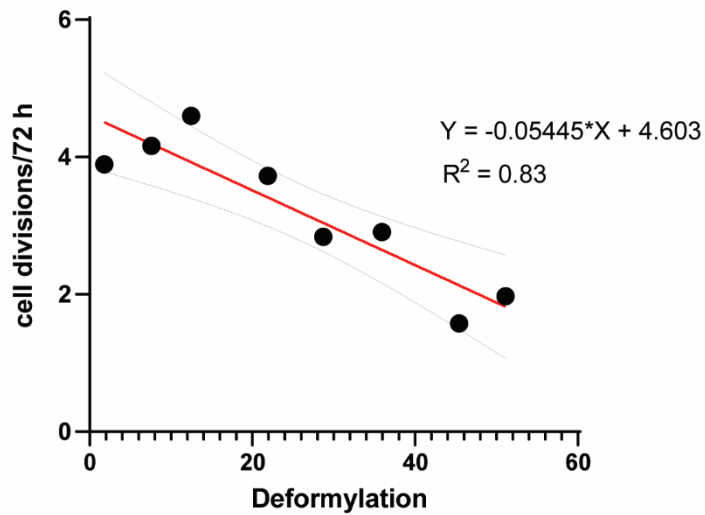
Figure SI-2A: Upper graphs represent the absolute number of the cell divisions calculated as $doubling\ time = \frac{72*\log(2)}{\log_2(F)-\log_2(I)}$, F = final number of cells, I = initial number of cells. and the rates of deformylation. The lower graph shows the inverse correlation between cell doublings over the duration of the experiment and their corresponding deformylation rates.
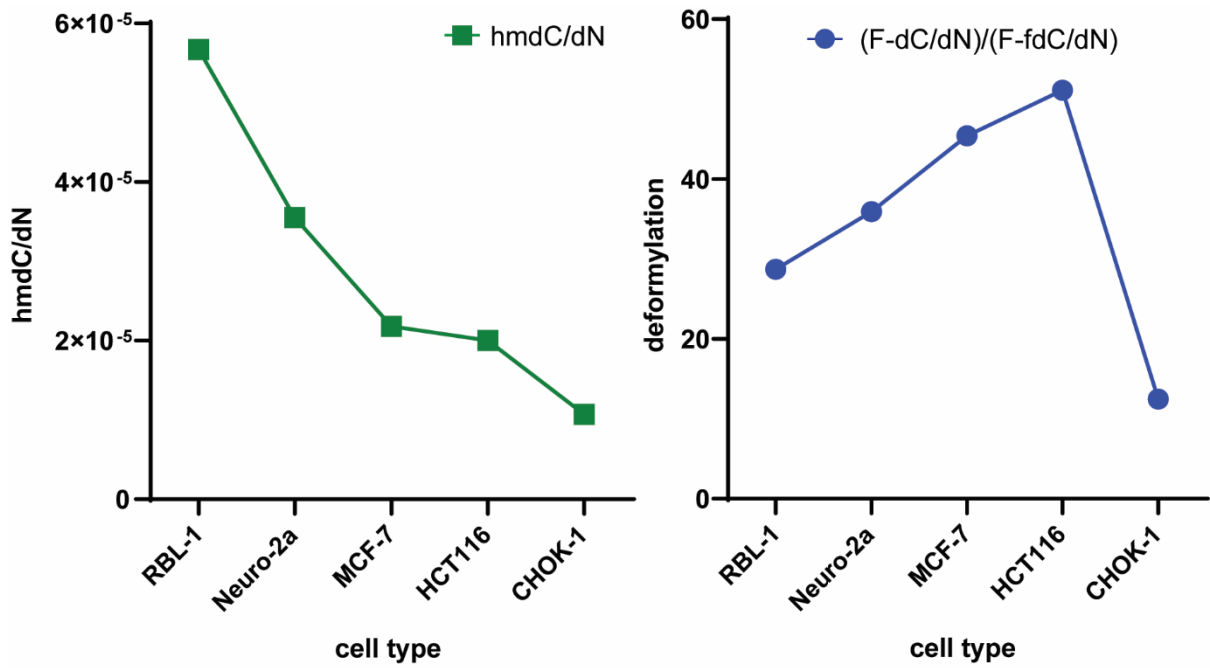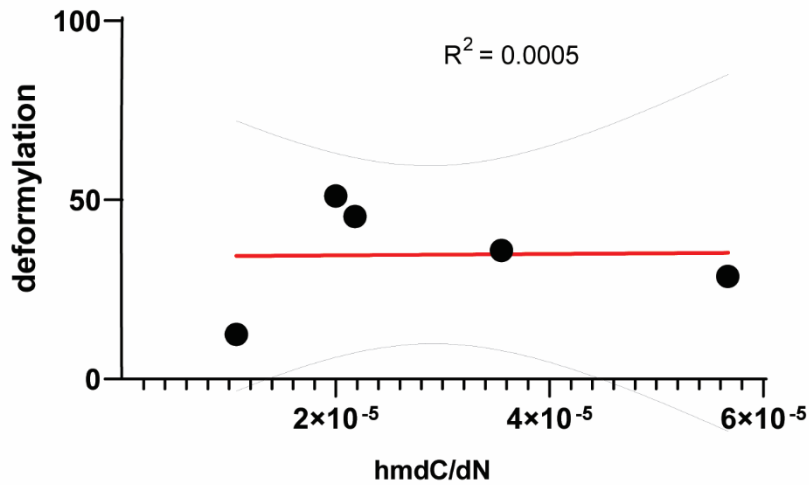
Figure SI-2B: Upper graphs represent the levels of genomic natural hmdC per dN in different cell lines and their corresponding rates of deformylation. The lower graph shows no correlation between Tet – enzymes activity and deformylation as a result of the analysis.
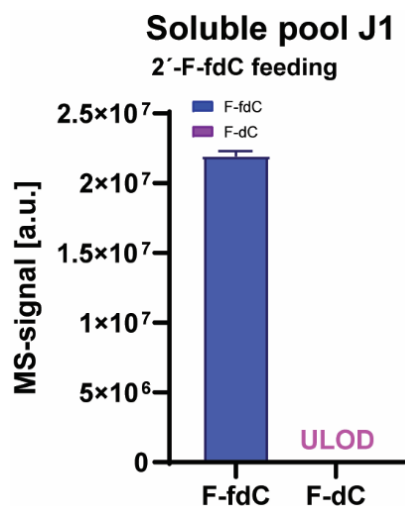
Figure SI-3A: *In vivo* stability assay of 2´F-fdC (**1**). **1** was incubated for 72 h in the cellular medium of J1 at 37 °C. Soluble pool content after feeding 2´F-fdC (**1**) at a concentration of 350 µM, levels of 2´F-fdC (**1**) vs. 2´F-dC (**2**). Error bars represent standard deviation.
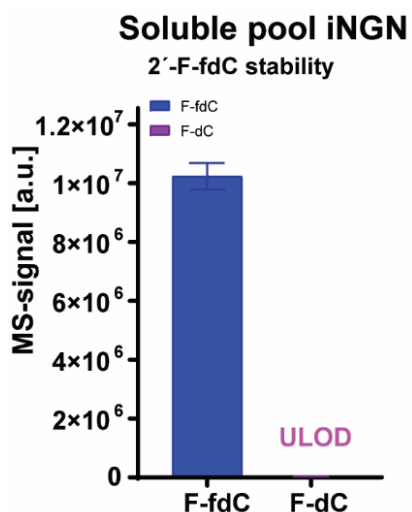
Figure SI-3B: *In vivo* Stability assay of 2´F-fdC (**1**). **1** was incubated for 24 h in the cellular medium of iNGNs at 37 °C. Levels of 2´F-fdC (**1**) vs. 2´F-dC (**2**) analyzed by UHPLC-MS/MS. Error bars represent standard deviation.
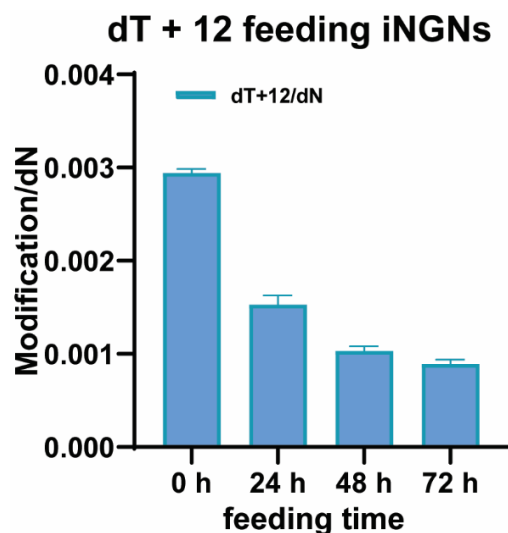
Figure SI-4: Amount of dT+12 (**10**) incorporated in the genome of iNGN cells to different time points after feeding **10**. The error bars represent standard deviation.
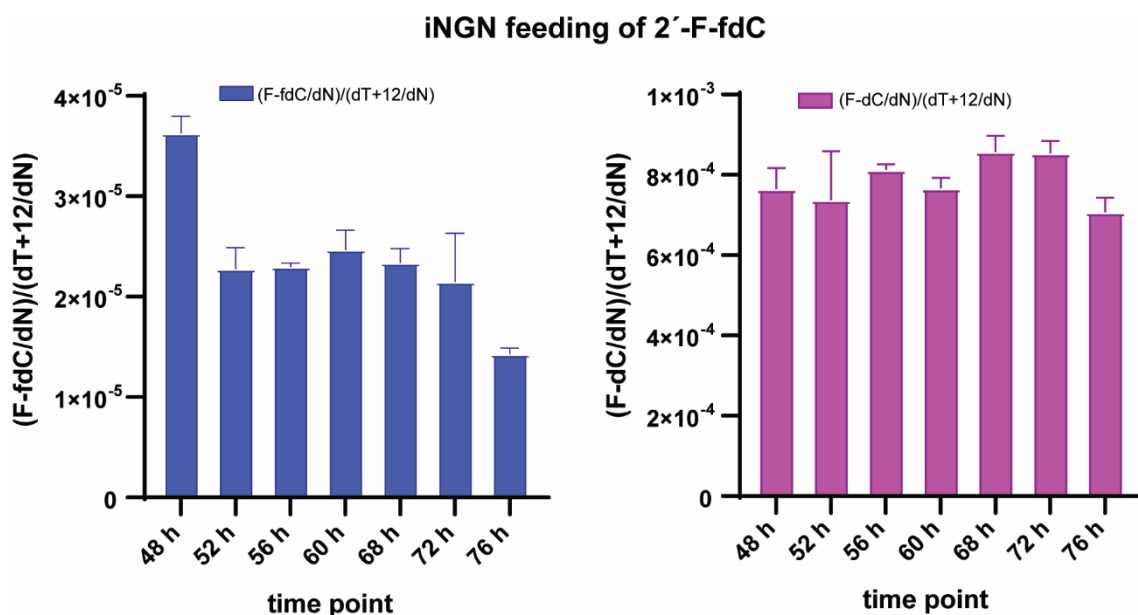
Figure SI-5: Levels of incorporated 2´-F-fdC (**1**) in the genome (left) and of the deformylation product 2´-F-dC (**2**, right) of iNGNs normalized to cell division (dT+12/dN) to different time points after feeding of 2´F-fdC (**1**) at a concentration of 350 µM. While levels of 2´F-fdC/dN decrease, this cannot be directly mirrored by the increase in the levels of 2´F-dC/dN between different time points, due to the fact, that the levels of these modifications are one order of magnitude apart. The error bars represent the standard deviation-
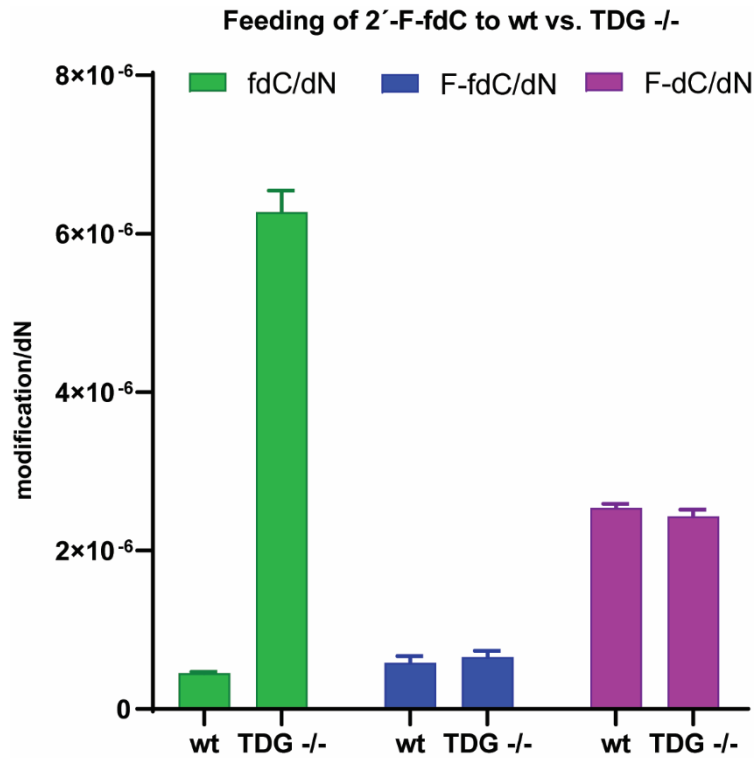
Figure SI-6: Levels of natural fdC (left, green), incorporated modified 2´-F-fdC (**1**) in the genome (middle, blue) and of the deformylation product 2´-F-dC (**2**, right, magenta) of a wildtype and Tdg -/- cell line. Levels of genomic natural fdC increase dramatically in absence of Tdg activity, while the levels of fluorinated-derivatives (**1** and **2**) stay constant. The error bars represent the standard deviation.
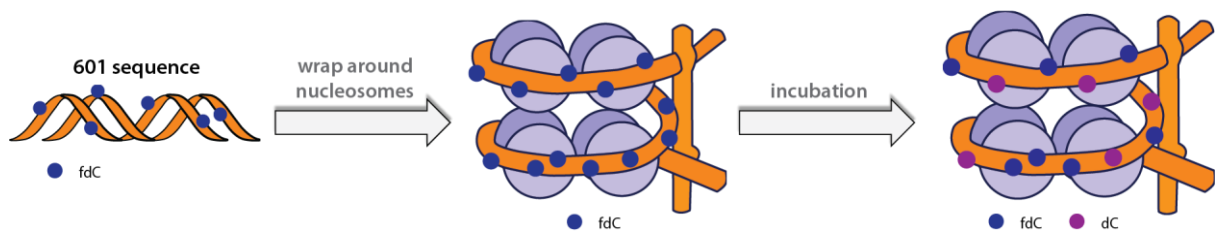
Figure SI-7A: Experimental workflow of PCR-produced 601 Widom sequence[1] (shown below) and its incubation with nucleosomes, to determine the involvement of nucleosomes on deformylation of fdC.

601 sequence:
5´-ATCGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAA CGCGGGGGACAG-3´

Forward Primer: 5´-ATCGATGTATATATCTGACACGTGC-3´

Reverse Primer: 5´-ATCAGAATCCCGGTGCCGAG-3´

7
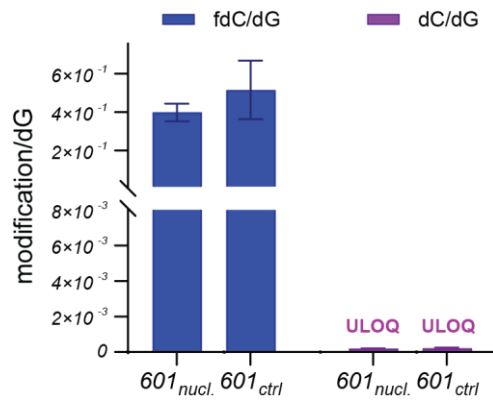
Figure SI-7B: Levels of fdC and dC found in the 601 sequence with or without incubation with nucleosomes. Error bars represent standard deviation of three technical replicates.
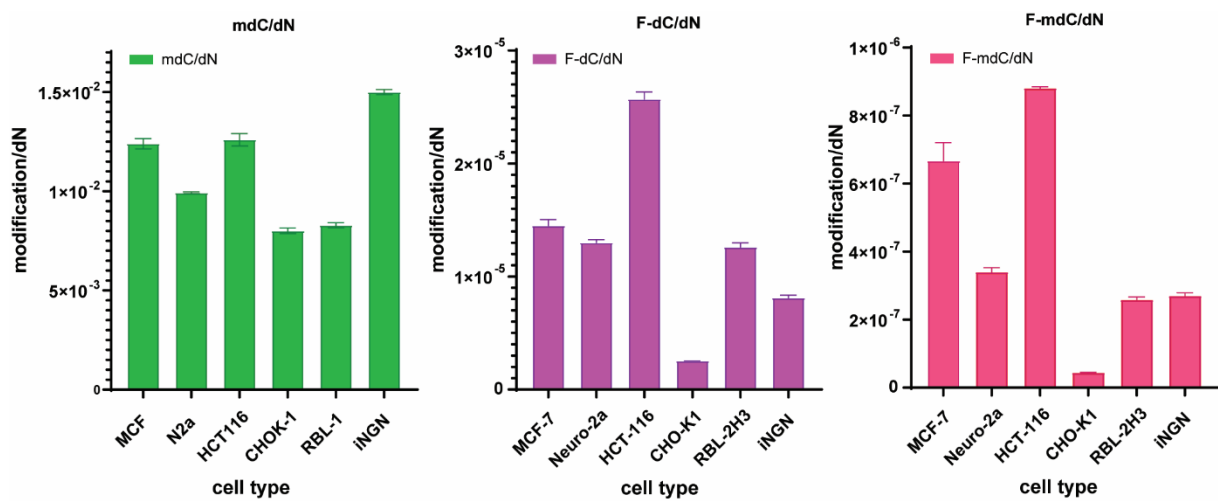
Figure SI-8A: Levels of incorporated natural mdC (green, left), 2′-F-dC (**2**, purple, middle) and 2′-F-mdC (**9**, pink, right) in different cell lines. Error bars represent standard deviation of three biological replicates.
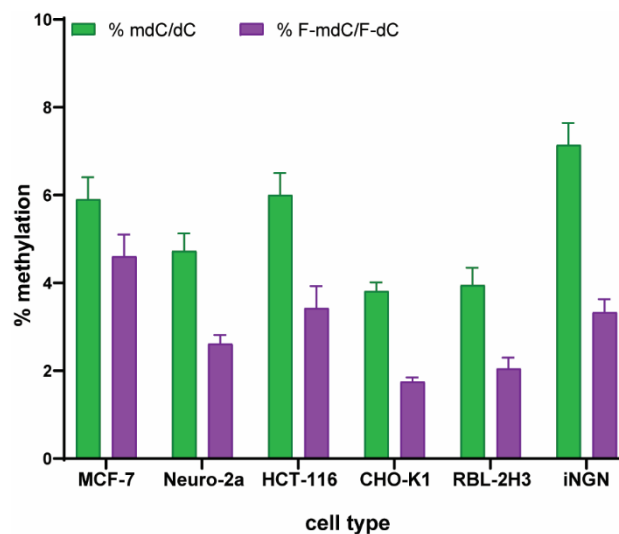
Figure SI-8B: Methylation levels of natural dC and F-dC in different cell lines. Error bars represent standard deviation of three biological replicates.
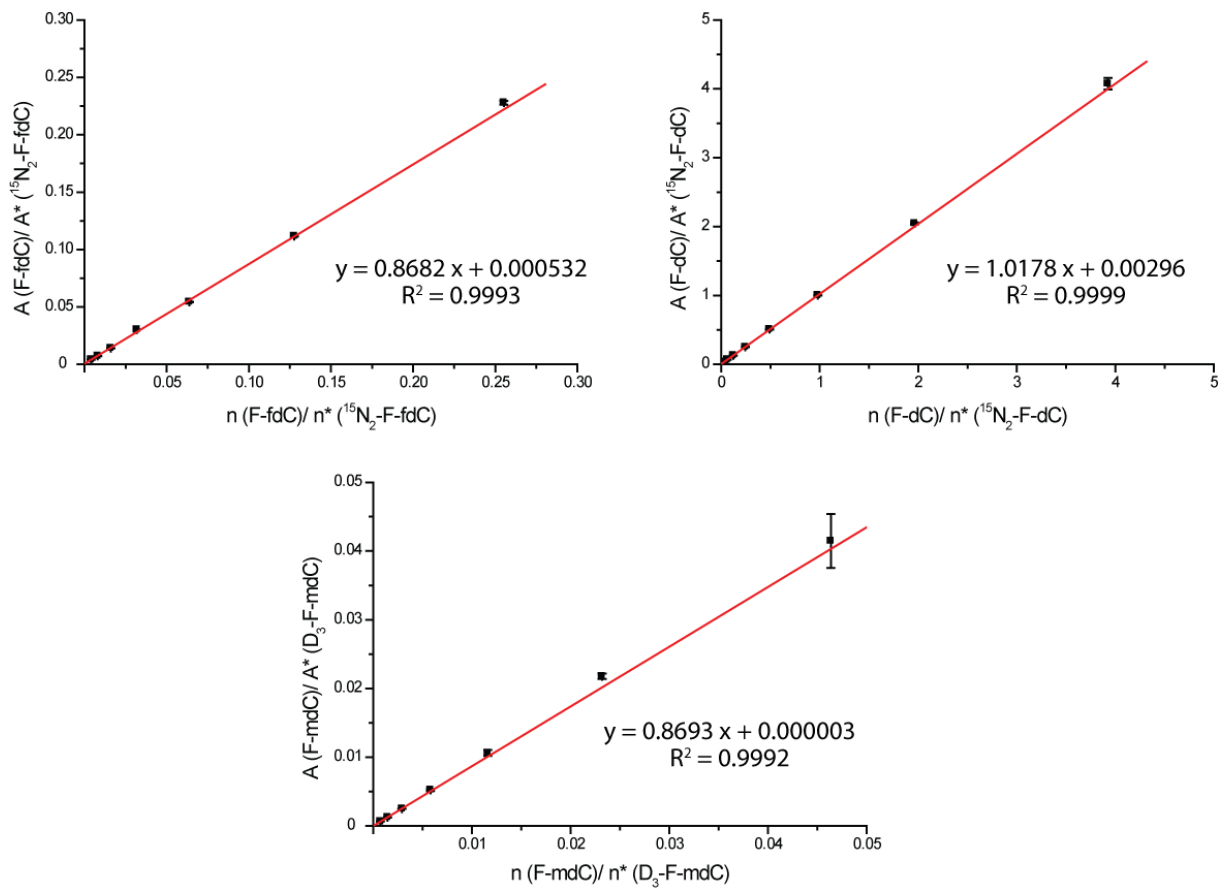
Figure SI-9: Internal calibration curves for the exact quantification of 2'-F-fdC (**1**), 2'-F-dC (**2**) and 2'-F-mdC (**9**) with the corresponding linear equation and coefficient of determination.[2]

Methods and Materials

Chemical Synthesis

Unless noted otherwise, all chemical transformations were performed using flame- or oven dried glassware under an atmosphere of argon. Chemicals were purchased from *Sigma-Aldrich*, *TCI*, *Fluka*, *ABCR*, *Carbosynth* or *Acros Organics* and used without further purification. Technical grade solvents were distilled prior to extraction or chromatography. Reactions and column chromatography fractions were monitored by thin-layer chromatography (TLC) on silica gel F_{254}TLC plates from Merck KGaA. Visualization of the developed TLC plates was achieved through UV-absorption or through staining with Hanessian's stain or $KMnO_4$ stain. Flash column chromatography was performed on Geduran®Si60 (40-63 µm) silica gel from Merck KGaA applying slight nitrogen pressure. NMR spectra were recorded on *Bruker Avance III HD 400 (*400 MHz), *Varian NMR-System 600* (600 MHz) and *Bruker Avance III HD* with *Cryo-Kopf 800* (800 MHz) spectrometers. ^{1}H-NMR shifts were calibrated to the residual solvent resonances: $CDCl_3$ (7.26 ppm), DMSO-d_6 (2.50ppm) and D_2O (4.79 ppm). ^{13}C-NMR shifts were calibrated to the residual solvent: $CDCl_3$ (77.16 ppm), DMSO-d_6 (39.52 ppm). All NMR spectra were analysed using the program MestReNova 10.0.1 from Mestrelab NMR data are reported as follows: chemical shift (multiplicity, coupling constants where applicable, number of hydrogens, assignment). Splitting is reported with the following symbols: s = singlet, d = doublet, t = triplet, dd = doublet of doublets, dt = doublet of triplets, ddd = doublet of doublets of doublets, m = multiplet. Whenever a signal is not clearly distinguishable from related nuclides, roman numerals (e.g. "I", "IV") are attached to indicate interchangeability. Research S. L. Low resolution mass spectra were measured on a LTQ FT-ICR by Thermo Finnigan GmbH. High resolution mass spectra were measured by the analytical section of the Department of Chemistry of Ludwig-Maximilians-Universität München on aMAT 90 (ESI) from Thermo Finnigan GmbH. IR spectra were recorded on a PerkinElmer Spectrum BX II FT-IR system. Substances were applied as a film or directly as solids on the ATR unit. Analytical RP-HPLC was performed on an analytical HPLC Waters Alliance(2695 Separation Module, 2996 Photodiode Array Detector) equipped with the column Nucleosil120-2 C18 from Macherey Nagel applying an eluent flow of 0.5 mL/min. Preparative RP-HPLC was performed on HPLC Waters Breeze (2487 Dual λ Array Detector, 1525 Binary HPLC Pump) equipped with the column VP 250/32 C18 from Macherey Nagel. A flowrate of 5mL/min was applied. HPLC grade MeCN was purchased from VWR.

Cell culture

DMEM high glucose (Sigma Aldrich) containing 10% heat-inactivated FBS (Gibco), 100 U/mL Penicillin-Streptomycin (Gibco), 2 mM L-glutamine were used to culture Neuro-2a, RBL-2H3 ad MCF-7 cell lines.

DMEM high glucose and Ham's Nutrient Mixture F12 (Sigma Aldrich) mixture 50/50 containing 100 U/mL Penicillin-Streptomycin (Gibco), 2 mM L-glutamine were used to culture CHO-K1 cells.

Minimum Essential Medium Eagle (MEM) (Sigma Aldrich) containing 10% heat-inactivated FBS (Gibco), 100 U/mL Penicillin-Streptomycin (Gibco), 2 mM L-glutamine and were used to culture HCT-116 cells.

iNGN cell line[3] at somatic cell state was cultured in "E7 media" consisted of DMEM high glucose (Sigma Aldrich) and Ham's Nutrient Mixture F12 (Sigma Aldrich) 50/50 mixture supplemented with 100 U/mL Penicillin-Streptomycin (Gibco), 64mg/L L -Ascorbic acid 2-phosphate, 0.1mM Nicotinamide (Sigma Aldrich), 78nM Sodium selenite (Sigma Aldrich), 0.64g/L NaCl (Sigma Aldrich), 20 µg/mL Insulin in 25mM HEPES buffer pH 8.2 (Sigma Aldrich), 10 µg/mL Holo-transferrin (Sigma Aldrich). Additionally, with freshly added growth factors 20 ng/mL FGF-2 and TGF-β 10 ng/mL (MACS) making "E9" media. After every splitting the cells were treated with the adherence factor – Thiazovivin (TOCRIS) at 1uM added freshly to the plate. During differentiation the cells were cultured in E7 medium complemented by doxycycline 0,5 µg/mL (Sigma Aldrich) for the first 72 hours. After 72nd hour of differentiation the medium was changed into Neurobasal A (Gibco) + 2% NeuroBrew21 (MACS). The iNGNs were always plated on Geltrex coating (Thermo Fisher) according to manufactures protocol.

mES cells were cultured in three different media – slitting, maintenance (a2i) and priming (C/R) media. The maintenance medium consisted of DMEM high glucose (Sigma Aldrich), 10% heat-inactivated FBS (Gibco), 0.1mM β-mercaptoethanol (Sigma Aldrich), 2mM L-Alanyl-L-Glutamine 200mM (Gibco), 1x Nonessential amino acids 100x (Sigma Aldrich) and 100 U/mL Penicillin-Streptomycin (Gibco). The culture medium consists of the splitting medium supplemented with LIF 1000U/mL (ORF Genetics), 3.0 µM CGP 77675 (Sigma Aldrich) and 3.0 µM GSK3 inhibitor CHIR-99021 (Sigma Aldrich). The priming media consisted of the splitting media supplemented with LIF 1000U/mL (ORF Genetics), 2.5 µM tankyrase inhibitor IWR-1 endo (Sigma Aldrich) and 3.0 µM CHIR-99021 (Sigma Aldrich). The cells were maintained in a2i medium and primed for 3.0 days via plating in C/R medium.

Metabolic-labelling experiments were conducted by plating cells in their dedicated medium containing 350µM of F-fdC for 3 days. mESCs were fed during priming without medium change until the moment of harvesting. The pulse-chase experiment using iNGN cells was performed so that the cells were fed dT+12 and F-fdC on the day of induction and the medium with nucleosides was replaced after 24hours and from the 48th hour the chase has started. The harvesting timepoints from the first hour of induction were: 48h, 52h, 56h, 60h, 68h, 72h and 76h.

The cells were counted before plating and prior to harvesting (at timepoints 0 h and 72 h) using Countess II automatic cell counter (LifeTechnologies. ThermoFisher) via trypan blue staining. The values were recorded and the doubling rates calculated according to the following equation:

$$doubling\ time = \frac{72 * \log(2)}{log_2(F) - log_2(I)}$$

F = final number of cells, I = initial number of cells.

Cell Harvesting and Isolation of genomic DNA

All samples were washed with PBS (Sigma Aldrich) once before harvesting and lysed directly in the plates by adding RLT buffer (Qiagen) supplemented with 400µM 2,6-di-tert-butyl-4-methylphenol (BHT) and desferoxamine mesylate (DM). Next, DNA was sheered in MM400 bead mill (Retsch) at 30 Hz for 1 min in 2mL microcentrifuge tubes (Eppendorf) with 5mm diameter stainless steel beads (one per tube) and centrifuged at 21000rcf for 5 minutes. Genomic DNA (gDNA) was extracted using Zymo Quick gDNA mini-prep® kit according to the protocol with an addition of RNAse A treatment (Qiagen) at 0.2mg/mL in Genomic Lysis Buffer – 400µL of the solution transferred directly on the column and incubated for 15 minutes. All samples were eluted in DNAse-free ddH$_2$O with 20µM BHT, the concentration of gDNA was measured on Nanodrop.

DNA digestion

While investigating the deformylation levels using F-fdC as a metabolic label, the following conditions were used for the gDNA digestion procedure: 1 – 10 µg of genomic DNA in 35 µL H$_2$O were digested as follows: 2 µL of Degradase plus (*Zymo Research*), 5 µL of Degradase plus reaction buffer (*Zymo Research*), 0.5 µL of Benzonase nuclease (*Merck, Serratia marcescens*) and a specific amount of isotopically labelled internal standards were added. The

mixture was incubated for 4 h at 37 °C and then stored at -20 °C. Prior to LC-MS/MS analysis, samples were filtered by using an AcroPrep Advance 9 filter plate 0.2 μm Supor (*Pall Life Science*).

LC/MS-MS analysis of DNA samples[4]

Quantitative UHPLC-MS/MS analysis of digested DNA samples was performed using an Agilent 1290 UHPLC system equipped with a UV detector and an Agilent 6490 triple quadrupole mass spectrometer. Nucleosides of interest were quantified using the stable isotope dilution technique.[2] The source-dependent parameters were as follows: gas temperature 80 °C, gas flow 15 L/min (N$_2$), nebulizer 30 psi, sheath gas heater 275 °C, sheath gas flow 11 L/min (N2), capillary voltage 2.500 V in the positive ion mode, capillary voltage −2.250 V in the negative ion mode and nozzle voltage 500 V. The fragmentor voltage was 380 V/ 250 V. Delta EMV was set to 500 (positive mode) and 800 (negative mode). Compound-dependent parameters are summarized in Supplementary Tables 1-3. Chromatography was performed by a Poroshell 120 SB-C8 column (Agilent, 2.7 μm, 2.1 mm × 150 mm) at 35 °C using a gradient of water and MeCN, each containing 0.0085% (v/v) formic acid, at a flow rate of 0.35 mL/min: 0 →4 min; 0 →3.5% (v/v) MeCN; 4 →7.9 min; 3.5 →5% MeCN; 7.9 →8.2 min; 5 →80% MeCN; 8.2 →11.5 min; 80% MeCN; 11.5 →12 min; 80 →0% MeCN; 12 →14 min; 0% MeCN. The autosampler was cooled to 4 °C. The injection volume was amounted to 39 μL.

Quantification of nucleosides in DNA samples

Nucleosides were quantified using internal calibration curves and the stable isotope dilution technique as described in the literature.[2] Used internal calibration curves for F-nucleosides are shown in Supplementary Figure 4.

Soluble nucleoside pool extraction and purification

The soluble pool extraction protocol started with a cell harvest using typical cell passaging trypsinisation technique rather than a usual cell lysis method (using TrypLE reagent, Gibco). The cells were counted with trypan blue via Countess II FL Automated Cell Counter, Life Technologies (ThermoFisher Scienitfic) and pelleted via centrifugation at 260g for 3min. The pellet was resuspended with the ice-cold mixture of acetonitrile and water (1:1 ratio), 1mL per 1x10^{6}cells and gently vortexed. The cell suspension was incubated on ice for 15min and centrifuged at 16 000g for 10min at 4°C. The supernatant containing the cytosolic fraction was

collected and lyophilised, while the cell pellet was treated with the lysis buffer for subsequent gDNA isolation.

Samples were dissolved in H_2O (pH = 4, 100 µL H_2O for soluble pool from 400 000 cells). Supel-Select HBL 60 mg columns were equilibrated with 1.5 mL MeOH and 3 mL H_2O (pH = 4). Samples were loaded the surface of the column, washed with 4 mL H_2O (pH = 4). Cartridges were dried under high vacuum for 15 min. Sample was eluted with 1.5 mL MeOH/MeCN = 1:1. Eluent was lyophilized, dissolved in H_2O and analysed by UHPLC-MS/MS as described for genomic DNA samples.[2]

Synthetic procedures
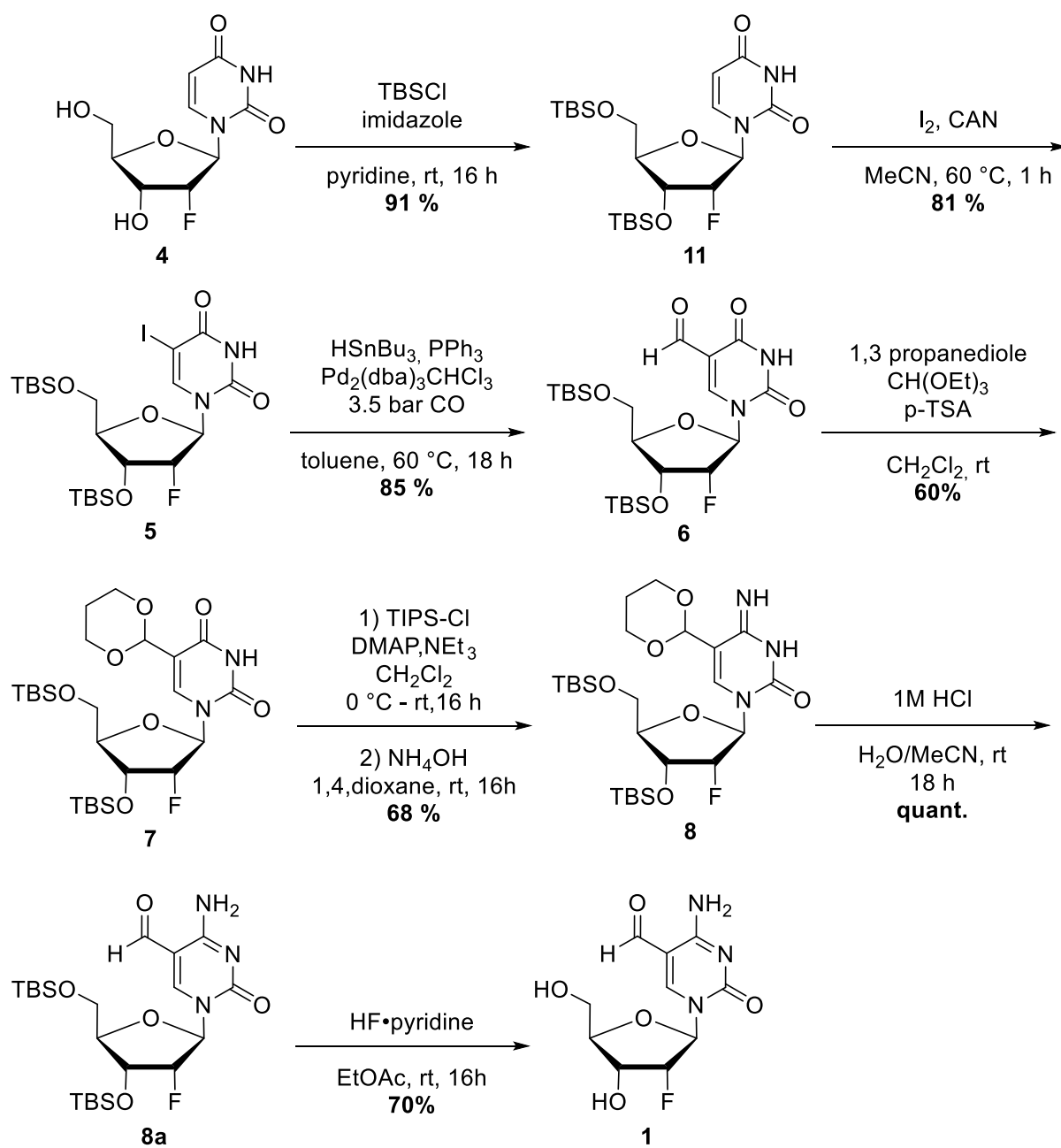
Scheme SI-1: Synthesis pathway towards 2'-F-fdC (**1**).

Synthesis of O3´,5´-Bis(tertbutyldimethylsilyl)-2´-fluorouridine (11)

11
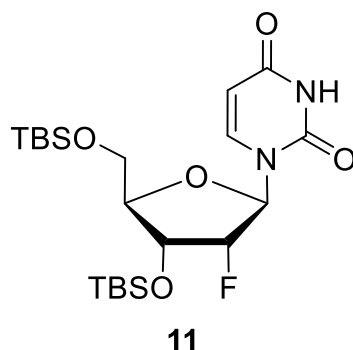
According to a modified literature [5] procedure 2'-deoxy-2'-fluoruridine (2.0 g, 8.1 mmol, 1.0 equiv.), TBS-Cl (3.7 g, 24.4 mmol, 3.0 equiv.) and imidazole (2.5 g, 36.5 mmol, 4.5 equiv.) were combined in a Schlenk flask an dissolved in pyridine (32 mL). The resulting reaction mixture was stirred at room temperature for 18 h. The reaction was stopped by addition of NaHCO$_3$ (sat. aq., 150 mL), extracted with dichloromethane (3 x 40 mL). The combined organic layers were washed with brine (150 mL), dried over Na$_2$SO$_4$ and the solvent was removed *in vacuo*. Residual pyridine was co-evaporated with toluene yielding a colourless crude. Purification by flash column chromatography (SiO$_2$ hexanes/ethyl acetate 2:1) yielded target compound **11** as a colourless solid (3.8 g, 8.0 mmol, 99 %).
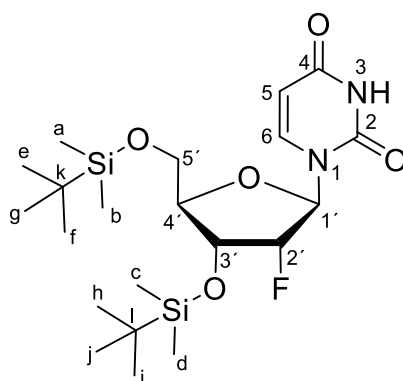
R$_f$ = 0.48 (CH$_2$Cl$_2$/MeOH 9:1)

^{1}H-NMR (599 MHz, CDCl$_3$, ppm): δ = 8.06 (s, 1H, N3-H), 7.91 (s, 1H, $^3J_{H6\text{-}H5}$ = 8.0 Hz, $^4J_{H6\text{-}H1'}$ = 0.9 Hz, C6-H), 6.06 (dd, $^3J_{H1'\text{-}F2'}$ = 15.2 Hz, $^3J_{H1'\text{-}H2'}$ = 2.0 Hz, 1H, C1'-H), 5.68 (dd, $^3J_{H5\text{-}H6}$ = 8.0 Hz, $^4J_{H5\text{-}H3}$ = 0.9 Hz, C5-H), 4.77 (ddd, $^2J_{H2'\text{-}F2'}$ = 52.4 Hz, $^3J_{H2'\text{-}H1'}$ = 2.0 Hz, $^3J_{H2'\text{-}H3}$ = 4.3 Hz, 1H, C2'-H), 4.29 (ddd, $^3J_{H3'\text{-}F2'}$ = 18.7 Hz, $^3J_{H3'\text{-}H2'}$ = 4.2 Hz, $^3J_{H3'\text{-}H4'}$ = 7.2 Hz, 1H, C3'-H), 4.08

16

(dd, $^3J_{H4'-H3'}$ = 7.2 Hz, $^3J_{H4'-H5'}$ = 1.6 Hz, 1H, C4'-H), 4.05 (dd, $^2J_{H5'a-H5'b}$ = 11.7 Hz, $^3J_{H5'a-H4'}$ = 2.0 Hz, 1H, C5'-H$_a$)^I, 3.78 (dd, $^2J_{H5'b-H5'a}$ = 11.8 Hz, $^3J_{H5'b-H4'}$ = 1.7 Hz, 1H, C5'-H$_b$)^I, 0.93 (s, 9H, C$_{e-g}$-H3)^{II}, 0.91 (s, 9H, C$_{h-j}$-H3)^{II}, 0.12 (s, 6H, C$_{a,b}$-H3)^{III}, 0.12 (s, 3H, C$_c$-H3)^{III}, 0.11 (s, 3H, C$_d$-H3)^{III}.

^{13}C-NMR (101 MHz, CDCl$_3$, ppm): δ = 162.6 (C4), 149.9 (C2), 139.9 (C6), 102.5 (C5), 93.2 (d, $^1J_{C2'-F2'}$ = 193.5 Hz, C2'), 87.9 (d, $^2J_{C1'-F2'}$ = 33.8 Hz, C1'), 84.0 (C4'), 68.8 (d, $^2J_{C3'-F2'}$ = 16.3 Hz, C3'), 60.9 (C5'), 26.1 (3C, C$_{e-g}$)^I, 25.8 (3C, C$_{h-j}$)^I, 18.6 (C$_k$)^{II}, 18.3 (C$_l$)^{II}, -4.5 (C$_a$)^{III}, -4.9 (C$_b$)^{III}, -5.3 (C$_c$)^{III}, -5.3 (C$_d$)^{III}.

^{19}F-NMR (377 MHz, CDCl$_3$, ppm): δ = -202.6 (ddd, $^2J_{F2'-H2'}$ = 52.5 Hz, $^3J_{F2'-H3'}$ = 18.7 Hz, $^3J_{F2'-H1'}$ = 15.2 Hz, C2'-F).

HRMS (ESI+): calc. for C$_{21}$H$_{40}$FN$_2$O$_5$Si$_2^+$ [M+H]$^+$: 475.2454, found: 475.2453.

HRMS (ESI-): calc. for C$_{21}$H$_{38}$FN$_2$O$_5$Si$_2^-$ [M-H]$^-$: 473.2309, found: 473.2319.

IR (ATR): ν /cm^{-1} = 3390 (w), 3261 (w), 2593 (w), 2928 (w), 2857 (w), 1759 (w), 1712 (s), 1707 (s), 1695 (vs), 1628 (w), 1455 (m), 1251 (m), 1158 (m), 1120 (s), 994 (m), 978 (m), 880 (ms), 863 (m), 828 (vs), 803 (s) 760 (s), 665 (m).

***Synthesis of* O3´,5´-Bis(tertbutyldimethylsilyl)-2´-fluor-5-iodouridine (5)**

5
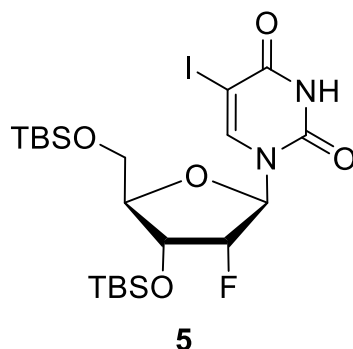
TBS-protected uridine **11** (2.1 g, 4.4 mmol, 1.0 equiv.) was dissolved in dry acetonitrile (200 mL), ceric ammonium nitrate (5.3 g, 9.6 mmol, 2.2 equiv.) and iodine (2.4 g, 9.6 mmol, 2.2 equiv.) were added, and the resulting reaction mixture was stirred for one hour at 60 °C. Upon complete conversion of starting material, the reaction was stopped by addition of $NaHCO_3$ (sat. aq., 100 mL) and $Na_2S_2O_3$ (sat. aq., 100 mL). Reaction mixture was filtered over a plug of SiO_2. The obtained filtrate was concentrated *in vacuo*, the aqueous residue was extracted with dichloromethane (4 x 40 mL). The combined organic layers were washed with brine (150 mL), dried over Na_2SO_4 volatiles were removed under reduced pressure. Resulting crude product was purified by flash column chromatography (SiO_2, hexanes/ethyl acetate 4:1) yielding **5** as a colourless foam (2.1 g, 3.5 mmol, 81 %).
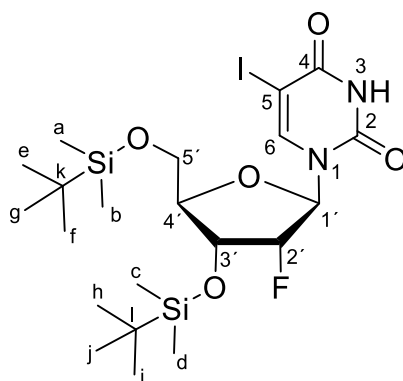
R_f = 0.75 (hexanes/ethyl acetate 2:1)

¹H-NMR (599 MHz, $CDCl_3$, ppm): δ = 8.16 (s, 1H, N3-H), 7.98 (s, 1H, C6-H), 6.09 (dd, $^3J_{H1'-F2'}$ = 14.0 Hz, $^3J_{H1'-H2'}$ = 4.1 Hz, 1H, C1'-H), 4.82 (dt, $^2J_{H2'-F2'}$ = 52.6 Hz, $^3J_{H2'-H1',H3'}$ = 4.4 Hz, 1H, C2'-H), 4.29 (dt, $^3J_{H3'-F2'}$ = 11.3 Hz, $^3J_{H3'-H2',H4'}$ = 4.8 Hz, 1H, C3'-H), 4.08 (dd, $^3J_{H4'-H3'}$ = 4.9 Hz, $^3J_{H4'-H5'}$ = 2.1 Hz, 1H, C4'-H), 3.98 (dd, $^2J_{H5'a-H5'b}$ = 11.8 Hz, $^3J_{H5'a-H4'}$ = 1.9 Hz, 1H, C5'-Ha)ᴵ, 3.77

(dd, $^2J_{H5'b\text{-}H5'a}$ = 11.8 Hz, $^3J_{H5'b\text{-}H4'}$ = 2.2 Hz, 1H, C5'-H$_b$)^I , 0.97 (s, 9H, C$_{e\text{-}g}$-H$_3$)^{II}, 0.91 (s, 9H, C$_{h\text{-}j}$-H$_3$)^{II}, 0.18 (s, 3H, C$_a$-H$_3$)^{III}, 0.17 (s, 3H, C$_b$-H$_3$)^{III} 0.13 (s, 3H, C$_c$-H$_3$)^{III}, 0.11 (s, 3H, C$_d$-H$_3$)^{III}.

^{13}C-NMR (101 MHz, CDCl$_3$, ppm): δ = 159.5 (C4), 149.6 (C2), 144.1 (C6), 92.5 (d, $^1J_{C2'\text{-}F2'}$ = 195.7 Hz, C2'), 87.5 (d, $^2J_{C1'\text{-}F2'}$ = 33.1 Hz, C1'), 85.5 (C5), 69.9 (d, $^2J_{C3'\text{-}F2'}$ = 15.4 Hz, C3'), 69.2 (C4´), 62.0 (C5'), 26.5 (3C, C$_{e\text{-}g}$)^I , 25.8 (3C, C$_{h\text{-}j}$)^I, 18.8 (C$_k$)^{II}, 18.3 (C$_l$)^{II}, -4.5 (C$_a$)^{III}, -4.9 (2C, C$_{b,c}$)^{III}, -5.3 (C$_d$)^{III}.

^{19}F-NMR (377 MHz, CDCl$_3$, ppm): δ = -205.7 (dt, $^2J_{F2'\text{-}H2'}$ = 52.6 Hz, $^3J_{F2'\text{-}H1',H3'}$ = 12.6 Hz, C2'-F).

HRMS (ESI+): calc. for C$_{21}$H$_{39}$FIN$_2$O$_5$Si$_2^+$ [M+H]$^+$: 601.1421, found: 601.1419.
HRMS (ESI-): calc. for C$_{21}$H$_{37}$FIN$_2$O$_5$Si$_2^-$ [M-H]$^-$: 599.1275, found: 599.1292.

IR (ATR): ṽ /cm^{-1} = 3171(w), 3067 (w), 2953 (w), 2928 (m), 2856 (w), 1690 (vs), 1606 (m), 1471 (w), 1426 (w), 1407 (w), 1360 (w), 1253 (s), 1166 (m), 1127 (s), 1072 (s), 996 (w), 680 (w), 889 (m), 858 (m), 802 (vs), 778 (vs), 670 (m).

Synthesis of O3´,5´-Bis(tertbutyldimethylsilyl)-2´-fluor-5-formyluridine (6)

6
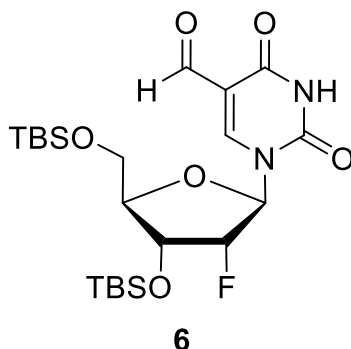
According to a modified literature procedure[5] Pd$_2$dba$_3$•CHCl$_3$ (0.14 g, 0.13 mmol, 0.1 equiv.), PPh$_3$ (0.21 g, 0.80 mmol, 0.6 equiv.) and **5** (0.80 g, 1.3 mmol, 1.0 equiv.) were charged to an autoclave vessel and dissolved in dry toluene (26 mL). The autoclave apparatus was sealed, the reaction mixture was purged with carbon monoxide (3 x 3.5 bar) and subsequently heated to 60 °C. A solution of HSnBu$_3$ (0.47 g, 0.43 mL, 1.6 mmol, 1.2 equiv.) in toluene (0.4 mL) was added via a syringe pump (55 μL/h) over a period of 15 h. After complete addition, reaction mixture was let to cool to room temperature and filtered over a plug of silica (eluted with 350 mL ethyl acetate). The resulting crude product was purified with flash column chromatography (SiO$_2$, 10 w% K$_2$CO$_3$, hexanes/ethyl acetate 4:1) yielding the formylated compound **6** as an colourless foam (560 mg, 1.12 mmol, 84 %).
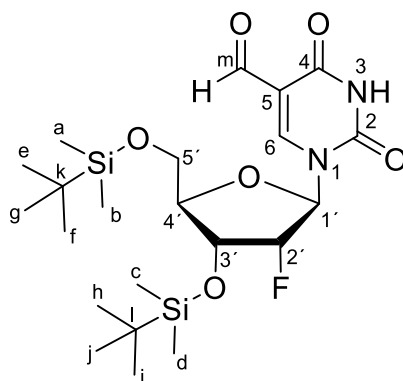
R$_f$ = 0.75 (hexanes/ethyl acetate 2:1)

¹H-NMR (599 MHz, CDCl$_3$, ppm): δ = 10.22 (s, 1H, C$_m$-H), 8.46 (s, 1H, N3-H), 8.42 (s, 1H, C6-H), 6.12 (dd, $^3J_{H1'-F2'}$ = 13.9 Hz, $^3J_{H1'-H2'}$ = 4.0 Hz, 1H, C1'-H), 4.89 (dt, $^2J_{H2'-F2'}$ = 52.5 Hz, $^3J_{H2'-H1',H3'}$ = 4.3 Hz, 1H, C2'-H), 4.31 (dt, $^3J_{H3'-F2'}$ = 11.2 Hz, $^3J_{H3'-H2',H4'}$ = 4.8 Hz, 1H, C3'-H), 4.16-4.12 (m, 1H, C4'-H), 3.99 (dd, $^2J_{H5'a-H5'b}$ = 11.9 Hz, $^3J_{H5'a-H4'}$ = 2.1 Hz, 1H, C5'-H$_a$)^I, 3.78 (dd,

$^2J_{H5'b-H5'a}$ = 11.8 Hz, $^3J_{H5'b-H4'}$ = 2.4 Hz, 1H, C5'-H$_b$)^I , 0.93 (s, 9H, C$_{e-g}$-H$_3$)^{II}, 0.91 (s, 9H, C$_{h-j}$ - H$_3$)^{II}, 0.15 (s, 6H, C$_{a,b}$-H$_3$)^{III}, 0.13 (s, 3H, C$_c$-H$_3$)^{III}, 0.11 (s, 3H, C$_d$-H$_3$)^{III}.

^{13}C-NMR (101 MHz, CDCl$_3$, ppm): δ = 185.5 (C$_m$), 161.1 (C4), 149.0 (C2), 145.3 (C6), 111.8 (C5), 92.7 (d, $^1J_{C2'-F2'}$ = 196.4 Hz, C2'), 88.3 (d, $^2J_{C1'-F2'}$ = 33.4 Hz, C1'), 86.0 (C4´), 70.0 (d, $^2J_{C3'-F2'}$ = 15.4 Hz, C3'), 61.9 (C5´), 26.2 (3C, C$_{e-g}$)^I , 25.8 (3C, C$_{h-j}$)^I, 18.7 (C$_k$)^{II}, 18.3 (C$_l$)^{II}, -4.5 (C$_a$)^{III}, -4.9 (C$_b$)^{III}, -5.3 (C$_c$)^{III} -5.4 (C$_d$)^{III}.

^{19}F-NMR (377 MHz, CDCl$_3$, ppm): δ = -205.3 (dt, $^2J_{F2'-H2'}$ = 52.2 Hz, $^3J_{F2'-H1',H3'}$ = 12.2 Hz, C2'-F).

HRMS (ESI+): calc. for C$_{22}$H$_{40}$FN$_2$O$_6$Si$_2$$^+$ [M+H]$^+$: 503.2403, found: 503.2400.

IR (ATR): v /cm^{-1} = 2949 (w), 2926 (w), 2907 (w), 2889 (w), 2169 (m), 1692 (s), 1599 (m), 1470 (m), 1379 (w), 1361 (w), 1254 (m), 1233 (w), 1121 (m), 1073 (m), 997 (w), 982 (w), 830 (vs), 777 (vs), 673 (m), 658 (w).

Synthesis of O3´,5´-Bis(tertbutyldimethylsilyl)-2´-fluor-5-(1,3-dioxan-2-yl)uridine (7)

7

Nucleoside **6** (310 mg, 0.617 mmol, 1.0 equiv.) was dissolved in dry dichloromethane (12 mL) and cooled to 0 °C. At 0 °C 1,3-propanediol (188 mg, 0.180 mL, 2.47 mmol, 4.0 equiv.), triethylorthoformiate (183 mg, 0.210 mL, 1.23 mmol, 2.0 equiv.) and p-TSA (24.0 mg, 0.123 mmol, 0.2 equiv.) were added. The reaction mixture was stirred for 30 min at 0 °C and 60 min at room temperature. Reaction was quenched by addition of H_2O (25 mL) and extracted with CH_2Cl_2 (3 x 30 mL). Combined organic layers were washed with brine (2 x 80 mL) and dried over Na_2SO_4. After all volatiles were removed under reduced pressure, crude was purified via flash column chromatography (SiO_2, hexanes/ethyl acetate 2:1) to yield **7** as a colourless foam (218 mg, 0.389 mmol, 63 %).

R_f = 0.25 (hexanes/ethyl acetate 2:1)

^{1}H-NMR (599 MHz, $CDCl_3$, ppm): δ = 8.06 (s, 1H, N3-H), 7.66 (d, $^4J_{H6-Hm}$ = 0.8 Hz, 1H, C6-H), 5.97 (dd, $^3J_{H1´-F2´}$ = 15.7 Hz, $^3J_{H1´-H2´}$= 4.3 Hz, 1H, C1'-H), 5.52 (s, 1H, CH_m), 4.91 (dt, $^2J_{H2´-F2´}$ = 53.3 Hz, $^3J_{H2´-H1´;H3´}$ = 4.6 Hz, 1H, C2'-H), 4.35 (dt, $^3J_{H3´-F2´}$ = 10.7 Hz, $^3J_{H3´-H2´;H4´}$ = 5.2 Hz, 1H,

C3'-H), 4.20-4.15 (m, 2H, C_n-H_2)[I] , 4.01 (dp, $^3J_{H4'\text{-}H3'}$ = 5.2 Hz, $^3J_{H4'\text{-}H5'a;H5'b}$ = 1.7 Hz, 1H, C4'-H), 3.98-3.91 (m, 2H, C_p-H_2)[I], 3.90 (dd, $^2J_{H5'a\text{-}H5'b}$ = 11.7 Hz, $^3J_{H5'a\text{-}H4'}$ = 2.3 Hz, 1H, C5'-H_a)[II], 3.76 (dd, $^3J_{H5'b\text{-}H5'a}$ = 11.7 Hz, $^3J_{H5'b\text{-}H4'}$ = 3.4 Hz, 1H, $C_{5'}$-H_b)[II], 2.16-2.08 (m, 1H, C_o-H_2)[III], 1.43-1.38 (m, 1H, C_o-H_2)[III], 0.93 (s, 9H, $C_{e\text{-}g}$-H_3)[IV], 0.91 (s, 9H, $C_{h\text{-}j}$ -H_3)[IV], 0.12 (s, 3H, C_a-H_3)[V], 0.11 (s, 6H, $C_{b;c}$-H_3)[V], 0.11 (s, 3H, C_d-H_3)[V].

^{13}C-NMR (101 MHz, CDCl$_3$, ppm): δ = 161.1 (C4), 149.6 (C2), 169.7 (C6), 113.2 (C5), 95.5 (C_{acetal}) 91.8 (d, $^1J_{C2'\text{-}F2'}$ = 194.5 Hz, C2'), 88.7 (d, $^2J_{C1'\text{-}F2'}$ = 33.4 Hz, C1'), 86.0 (C4'), 69.9 (d, $^2J_{C3'\text{-}F2'}$ = 15.3 Hz, C3'), 68.7 (C_n)[I], 67.7 (C_p)[I], 62.3 (C5'), 26.2 (3C, $C_{e\text{-}g}$)[II] , 25.9 (3C, $C_{h\text{-}j}$)[II], 18.7 (C_k)[III], 18.4 (C_l)[III], -4.6 (C_a) [VI], -4.9 (C_b)[IV], -5.1 (C_c) [VI] -5.4 (C_d)[IV].

^{19}F-NMR (377 MHz, CDCl$_3$, ppm): δ = -204.7 (ddd, $^2J_{F2'\text{-}H2'}$ = 53.4 Hz, $^3J_{F2'\text{-}H3'}$ = 15.6 Hz, $^3J_{F2'\text{-}H1'}$ = 11.1 Hz C2'-F).

HRMS (ESI+): calc. for $C_{25}H_{46}FN_2O_7Si_2^+$ [M+H]$^+$: 561.2822, found: 561.2824.
HRMS (ESI-): calc. for $C_{25}H_{44}FN_2O_7Si_2^-$ [M-H]$^-$: 559.2677, found: 559.2690.

IR (ATR): ν /cm^{-1} = 3191 (w), 2954 (w), 2928 (w), 2887 (w), 2856 (w), 1697 (s), 1470 (m), 1463 (m), 1362 (m), 1257 (m), 1127 (m), 1092 (s), 1004 (m), 966 (m), 860 (m), 802 (vs), 777 (vs), 724 (m), 671 (m).

Synthesis of O3´,5´-Bis(tertbutyldimethylsilyl)-2´-fluor-5-(1,3-dioxan-2-yl)cytidine (8)

8

Nucleoside **7** (370 mg, 0.66 mmol, 1.0 equiv.) and DMAP (16.0 mg, 0.13 mmol, 0.2 equiv.) were dissolved in dry dichloromethane (7 mL), triethylamine (0.9 mL, 0.67 g, 6.6 mmol, 10 equiv.) was added. Reaction mixture was cooled to 0 °C. At 0 °C triisopropylsulfonylchloride (300 mg, 0.99 mmol, 1.5 equiv.) was added in three portions over 30 min at 0 °C. The reaction mixture was let warm to rt and stirred for 16 h at rt. Reaction was quenched by addition of NH_4Cl (sat. aq. 25 mL) and extracted with CH_2Cl_2 (3 x 50 mL). Combined organic layers were washed with brine (50 mL) and dried over Na_2SO_4. All volatiles were removed under reduced pressure. Isolated product **7a** was used for next reaction step without further purification.

Nucleoside **7a** (0.66 mmol, 1.0 equiv.) was dissolved in dry 1,4-dioxane (13 mL), an aqueous NH_3-solution (25%, 4 mL) was added. Reaction solution was stirred at rt for 20 h. After full conversion of starting material, reaction was stopped by addition of NH_4Cl (sat. aq. 50 mL). Mixture was extracted with ethyl acetate (3 x 100 mL), combined organic layers were washed with brine (100 mL) and dried over Na_2SO_4. After all volatiles were removed under reduced pressure, crude was purified via flash column chromatography (SiO_2, CH_2Cl_2/MeOH 50:1 → 20:1) to yield **8** as a yellow foam (253 mg, 0.452 mmol, 68 %).

R$_f$ = 0.56 (CH$_2$Cl$_2$/MeOH 10:1)

^{1}H-NMR (599 MHz, CDCl$_3$, ppm): δ = 7.72 (s, 1H, C6-H), 7.24 (s, 1H, C4-NH), 6.56 (s, 1H, N3-H), 5.95 (dd, $^3J_{H1'-F2'}$ = 17.4 Hz, $^3J_{H1'-H2'}$ = 1.7 Hz, 1H, C1'-H), 5.29 (s, 1H, C-H$_m$), 4.86 (ddd, $^2J_{H2'-F2'}$ = 52.6 Hz, $^3J_{H2'-H1'}$ = 4.5 Hz, $^3J_{H2'-H3'}$ = 1.7 Hz, 1H, C2'-H), 4.25-4.16 (m, 3H, C$_n$, C$_p$, C3'-H)^I 4.06-4.01 (m, 2H, C4'-H, C5'-H) , 3.93-3.87 (m, 2H, C$_n$, C$_p$)^I, 3.81-3.77 (m, 1H, C5'-H), 2.18 (dtt, ^2J = 13.7 Hz, ^3J = 2.6 Hz, ^3J = 1.4 Hz, 1H, C$_o$)^{II}, 1.46 (dtt, ^2J = 13.7 Hz, ^3J = 2.6 Hz, ^3J = 1.4 Hz, 1H, C$_o$)^{II}, 0.95 (s, 9H, C$_{e-g}$-H$_3$)^{III}, 0.89 (s, 9H, C$_{h-j}$ -H$_3$)^{III}, 0.13 (s, 3H, C$_a$-H$_3$)^{IV}, 0.12 (s, 3H, C$_b$-H$_3$)^{IV}, 0.10 (s, 3H, C$_c$-H$_3$)^{IV} ,0.11 (s, 3H, C$_d$-H$_3$)^{IV}.

^{13}C-NMR (101 MHz, CDCl$_3$, ppm): δ = 163.2 (C4), 154.5 (C2), 140.5 (C6), 104.3 (C5), 99.3 (C$_m$) 92.2 (d, $^1J_{C2'-F2'}$ = 192.2 Hz, C2'), 89.7 (d, $^2J_{C1'-F2'}$ = 34.2 Hz, C1'), 83.7 (C4´), 69.0 (d, $^2J_{C3'-F2'}$ = 16.1 Hz, C3'), 67.5 (C$_n$)^I, 61.5 (C5´), , 26.3 (3C, C$_{e-g}$)^{II}, 25.8 (3C, C$_{h-j}$), 25.6 (C$_p$)^I 18.8 (C$_k$)^{III}, 18.2 (C$_l$)^{III}, -4.4 (C$_a$) ^{VI}, -4.9 (C$_b$)^{IV}, -5.1 (C$_c$) ^{VI} -5.1 (C$_d$)^{IV}.

^{19}F-NMR (377 MHz, CDCl$_3$, ppm): δ = -201.3 (dt, $^2J_{F2'-H2'}$ = 52.6 Hz, $^3J_{F2'-H3´,H4'}$ = 17.4 Hz, C2'-F).

HRMS (ESI+): calc. for C$_{25}$H$_{47}$N$_3$O$_6$FSi$_2$$^+$ [M+H]$^+$: 560.2982, found: 560.2979.
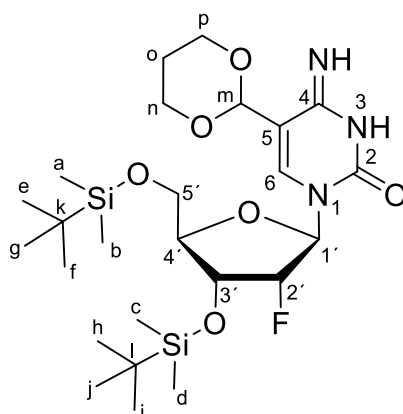
IR (ATR): ν /cm^{-1} = 2954 (w), 2928 (w), 2891 (w), 2857 (w), 1694 (s), 1463 (m), 1362 (m), 1257 (m), 1127 (m), 1093 (s), 1004 (m), 966 (m), 938 (w), 910 (m), 860 (m), 805 (vs), 777 (vs), 704 (m), 672 (m).

Synthesis of 2´-Fluor-5-formyl-cytidine (1)

1
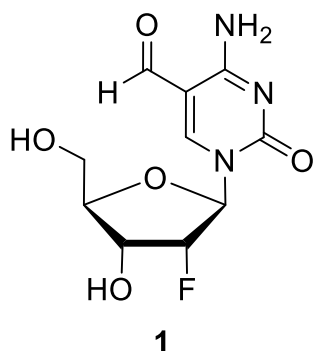
Nucleoside **8** (480 mg, 0.87 mmol, 1.0 equiv.) and was dissolved in dry THF (30 mL) aq. HCl (1 M, 3.5 mL, 3.45 mmol, 4.0 equiv.) were added. Reaction mixture was stirred at rt for 18 h. Reaction solution was diluted with H_2O (50 mL) and extracted with ethyl acetate (3 x 100 mL). Combined organic layers were washed with brine (50 mL) and dried over Na_2SO_4. All volatiles were removed under reduced pressure. Isolated product **8a** was used for next reaction step without further purification.

In a polypropylene falcon tube Nucleoside **8a** (180 mg, 0.46 mmol, 1.0 equiv.) was dissolved in dry ethyl acetate (7 mL) and cooled to 0 °C. At 0 °C a solution of HF·pyridine (70% HF, 0.18 mL, 6.97 mmol, 15 equiv.) was added, mixture was allowed to warm to rt and was stirred at rt for 20 h. After complete conversion of starting material, reaction was stopped by addition of TMSOMe (2.0 mL, 14.7 mmol, 32 equiv.) at 0 °C, and stirred for 1 h at 0 °C. Subsequently all volatiles were removed under reduced pressure. Crude product was dissolved in ddH₂O and purified by preparative reversed phase HPLC (0% buffer B → 13% buffer B in 45 min; buffer A: ddH2O, buffer B: MeCN). Purified product was lyophilized to yield **1** as a colourless solid (88 mg, 0.32 mmol, 70 %).
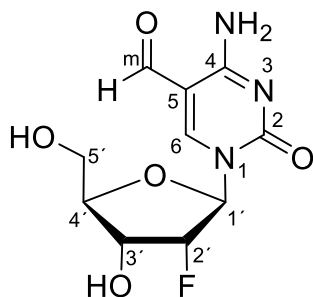
R_f = 0.13 (CH_2Cl_2/MeOH 5:1)

¹H-NMR (400 MHz, CDCl$_3$, ppm): δ = 9.51 (s, 1H, C$_m$-H), 8.88 (s, 1H, C6-H), 6.07 (d, $^3J_{H1'-F2'}$ = 17.7 Hz, 1H, C1'-H), 5.19 (dd, $^2J_{H2'-F2'}$ = 52.4 Hz, $^3J_{H2'-H1';H3'}$ = 4.3 Hz, 1H, C2'-H), 4.37 (ddd $^3J_{H3'-F2'}$ = 24.1 Hz, $^3J_{H3'-H2';H4'}$ = 9.3 Hz, $^3J_{H3'-H2';H4'}$ = 4.3 Hz, 1H, C3'-H), 4.27-4.19 (m, 1H, C4'-H) , 4.13 (dd, $^2J_{H5'a-H5'b}$ = 13.2 Hz, $^3J_{H5'a-H4'}$ = 2.3 Hz, 1H, C5'-H$_a$)^l, 3.91 (dd, $^2J_{H5'b-H5'a}$ = 13.2 Hz, $^3J_{H5'b-H4'}$ = 3.5 Hz, 1H, C$_{5'}$-H$_b$)^l.

¹³C-NMR (800 MHz, CDCl$_3$, ppm): δ = 190.2 (C$_m$), 162.8 (C4), 155.0 (C2), 154.9 (C6), 105.9 (C5), 93.7 (d, $^1J_{C2'-F2'}$ = 185.2 Hz, C2'), 89.9 (d, $^2J_{C1'-F2'}$ = 34.9 Hz, C1'), 82.4 (C4'), 67.2 (d, $^2J_{C3'-F2'}$ = 16.8 Hz, C3'), 59.1 (C5).

¹⁹F-NMR (400 MHz, CDCl$_3$, ppm): δ = -201.6 (ddd, $^2J_{F2'-H2'}$ = 52.4 Hz, $^3J_{F2'-H3'}$ = 24.3 Hz, $^3J_{F2'-H1'}$ = 17.8 Hz C2'-F).

HRMS (ESI+): calc. for C$_{10}$H$_{13}$N$_3$O$_5$F$^+$ [M+H]$^+$: 274.0834, found: 274.0835.

Supplementary tables

Table SI-1[2]: Compound-dependent LC-MS/MS-parameters used for the analysis of genomic DNA fed with fluorinated nucleosides. *CE: collision energy, CAV: collision cell accelerator voltage, EMV: electron multiplier voltage. The nucleosides were analysed in the positive ([M+H)]$^+$ species) as well as the negative ([M-H]$^-$ species) ion selected reaction monitoring mode (SRM).*

compound	Precursor ion (m/z)	MS1 Resolution	Product Ion (m/z)	MS2 Resolution	Dwell time [ms]	CE (V)	CAV (V)	Polarity
Time segment 1.5-3.0 min								
F-dC	246.09	Wide	112.06	Wide	70	15	3	Positive
[^{15}N$_2$]-F-dC	248.08	Wide	114.04	Wide	70	15	3	Positive
[^{15}N$_2$]-cadC	274.08	Wide	158.03	Wide	40	5	5	Positive
cadC	272.09	Wide	156.04	Wide	40	5	5	Positive
F-hmdC	276.10	Wide	142.06	Wide	50	10	3	Positive
[^{15}N$_2$]-F-hmdC	278.09	Wide	144.06	Wide	50	10	3	Positive
[^{15}N$_2$,D$_2$]-hmdC	262.12	Wide	146.07	Wide	25	27	1	Positive
hmdC	258.11	Wide	142.06	Wide	25	27	1	Positive
[D$_3$]-mdC	245.13	Wide	129.09	Wide	50	60	1	Positive
mdC	242.11	Wide	126.07	Wide	50	60	1	Positive
Time segment 3.0-4.7 min								
F-hmdU	275.07	Wide	255.06	Wide	80	3	7	Negative
F-dU	245.06	Wide	225.06	Wide	80	3	5	Negative
[D$_3$]-F-mdC	263.12	Wide	129.09	Wide	80	15	3	Positive
F-mdC	260.10	Wide	126.07	Wide	80	15	3	Positive
F-cadC	290.08	Wide	156.04	Wide	80	5	5	Positive
Time segment 4.7-10 min								
F-fdC	274.08	Wide	140.05	Wide	90	15	3	Positive
[^{15}N$_2$]-F-fdC	276.08	Wide	142.04	Wide	90	15	3	Positive
F-fdU	273.05	Wide	253.05	Wide	30	3	5	Negative

F-dT	259.07	Wide	239.07	Wide	70	3	5	Negative
dT	243.10	Wide	127.05	Wide	20	5	5	Negative
[^{15}N$_2$]-fdC	258.09	Wide	142.04	Wide	30	5	5	Positive
fdC	256.09	Wide	140.05	Wide	30	5	5	Positive
[^{15}N$_2$]-fdC	142.04	Wide	98.04	Wide	20	13	7	Positive
fdC	140.05	Wide	97.04	Wide	20	13	7	Positive

Table SI-2: Compound-dependent LC-MS/MS ranges of the corresponding linear equations.

compound	n (ULOQ)	n (LLOQ)	A/A* (ULOQ)	A/A* (LLOQ)
F-fdC	25.1 fmol	0.390 fmol	0.2267	0.003345
F-dC	800 fmol	3.13 fmol	4.074	0.01512
F-mdC	49.9 fmol	0.780 fmol	0.04174	0.0005833

Table SI-3: Compound-dependent LC-MS/MS-parameters used for the analysis of genomic DNA fed with fluorinated nucleosides and isotopically labelled dT. *CE: collision energy, CAV: collision cell accelerator voltage, EMV: electron multiplier voltage. The nucleosides were analysed in the positive ([M+H)]$^+$ species) as well as the negative ([M-H]$^-$ species) ion selected reaction monitoring mode (SRM).*

compound	Precursor ion (m/z)	MS1 Resolution	Product Ion (m/z)	MS2 Resolution	Dwell time [ms]	CE (V)	CAV (V)	Polarity
Time segment 1.5-3.7 min								
F-dC	246.09	Wide	112.06	Wide	70	15	3	Positive
[^{15}N$_2$]-F-dC	248.08	Wide	114.04	Wide	70	15	3	Positive
[^{15}N$_2$]-cadC	274.08	Wide	158.03	Wide	40	5	5	Positive
cadC	272.09	Wide	156.04	Wide	40	5	5	Positive
F-hmdC	276.10	Wide	142.06	Wide	50	10	3	Positive
[^{15}N$_2$]-F-hmdC	278.09	Wide	144.06	Wide	50	10	3	Positive
[^{15}N$_2$,D$_2$]-hmdC	262.12	Wide	146.07	Wide	25	27	1	Positive
hmdC	258.11	Wide	142.06	Wide	25	27	1	Positive
[D$_3$]-mdC	245.13	Wide	129.09	Wide	50	60	1	Positive
mdC	242.11	Wide	126.07	Wide	50	60	1	Positive
Time segment 3.7-4.7 min								

F-hmdU	275.07	Wide	255.06	Wide	80	3	7	Negative
F-dU	245.06	Wide	225.06	Wide	80	3	5	Negative
[D$_3$]-F-mdC	263.12	Wide	129.09	Wide	80	15	3	Positive
F-mdC	260.10	Wide	126.07	Wide	80	15	3	Positive
F-cadC	290.08	Wide	156.04	Wide	80	5	5	Positive
Time segment 4.7-12 min								
F-fdC	274.08	Wide	140.05	Wide	90	15	3	Positive
[^{15}N$_2$]-F-fdC	276.08	Wide	142.04	Wide	90	15	3	Positive
F-fdU	273.05	Wide	253.05	Wide	30	3	5	Negative
F-dT	259.07	Wide	239.07	Wide	70	3	5	Negative
dT	243.10	Wide	127.05	Wide	20	5	5	Negative
[^{15}N$_2$]-fdC	258.09	Wide	142.04	Wide	30	5	5	Positive
fdC	256.09	Wide	140.05	Wide	30	5	5	Positive
[^{15}N$_2$]-fdC	142.04	Wide	98.04	Wide	20	13	7	Positive
fdC	140.05	Wide	97.04	Wide	20	13	7	Positive
[^{13}C$_5$^{15}N$_2$]-dT	250.12	Wide	129.04	Wide	20	3	5	Positive
[^{13}C$_{10}$^{15}N$_2$]-dT	255.12	Wide	134.06	Wide	20	3	5	Positive

Appendix

NMR Spectra

Compound 11, ^{1}H-NMR, 400 MHz, CDCl$_3$
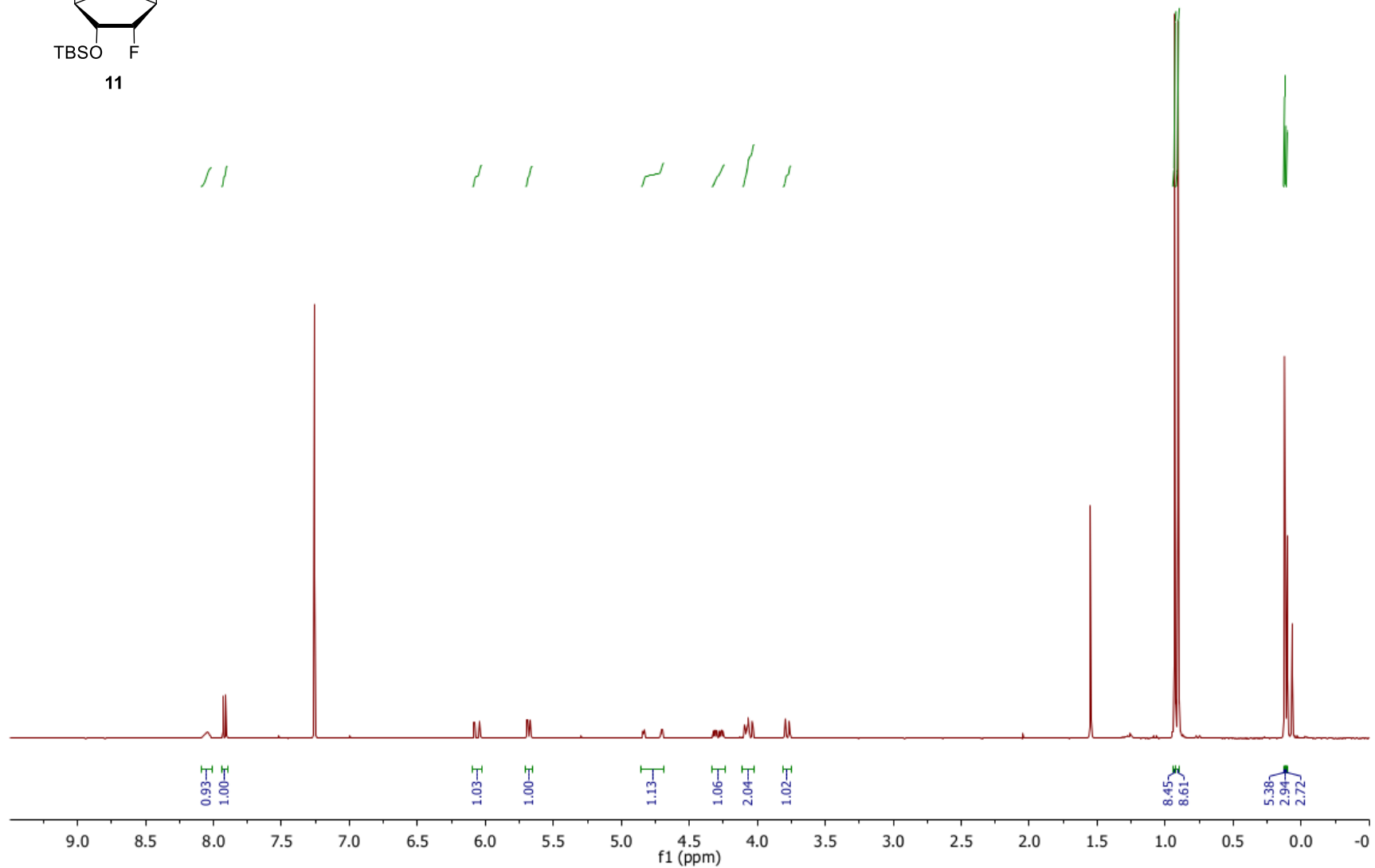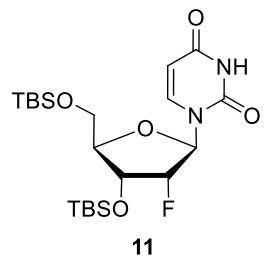
Compound 11, ^{13}C-NMR, 400 MHz, CDCl$_3$
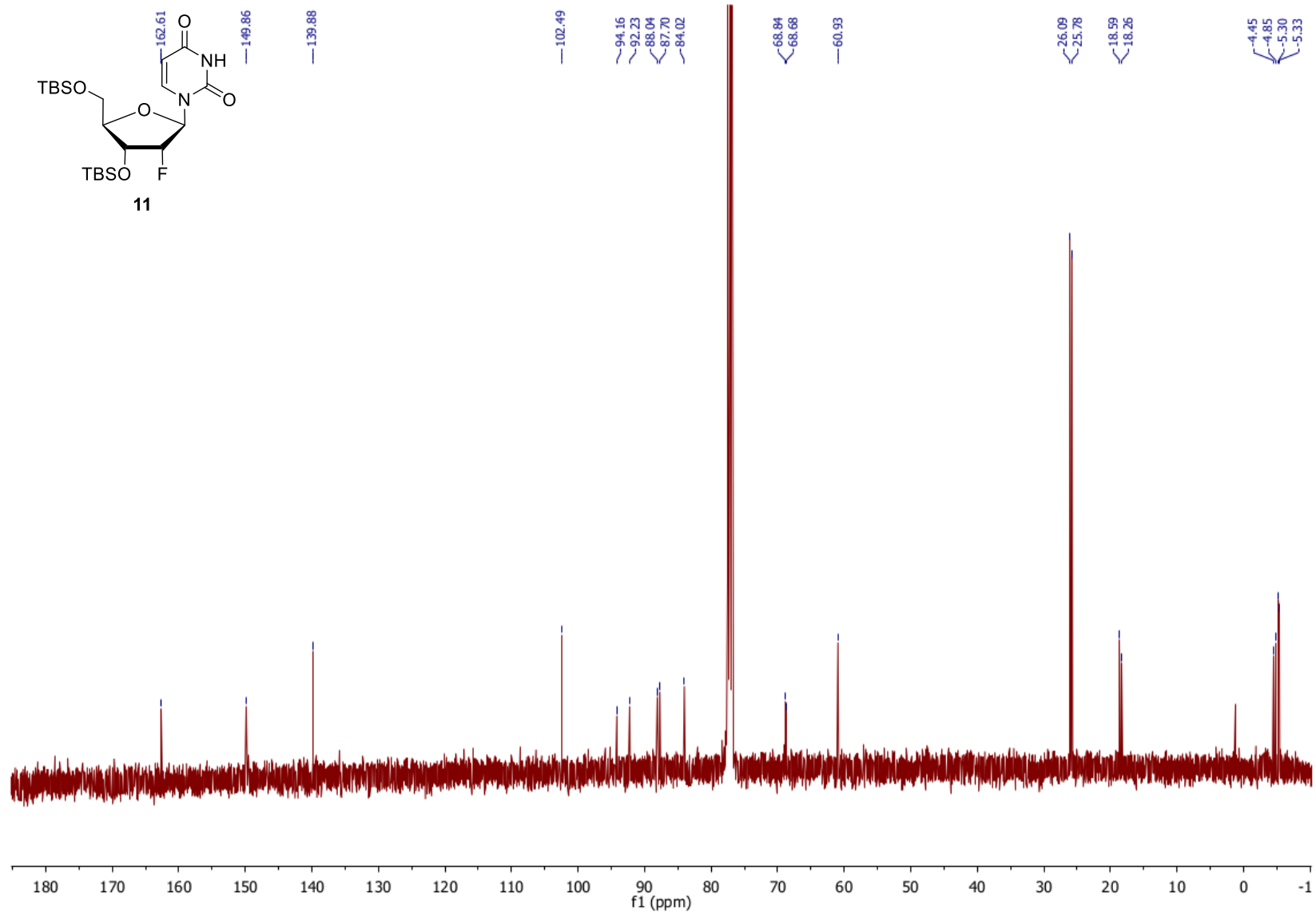
Compound 11, ^{19}F-NMR, 377 MHz, CDCl$_3$

-202.48
-202.52
-202.53
-202.57
-202.62
-202.66
-202.67
-202.71
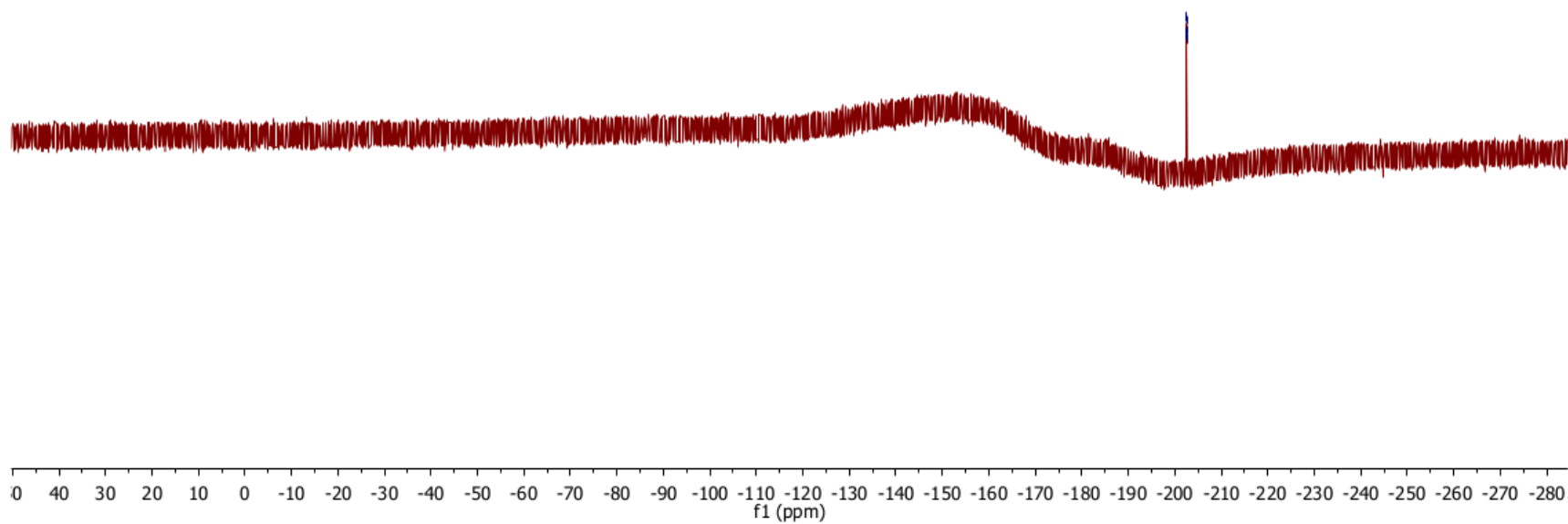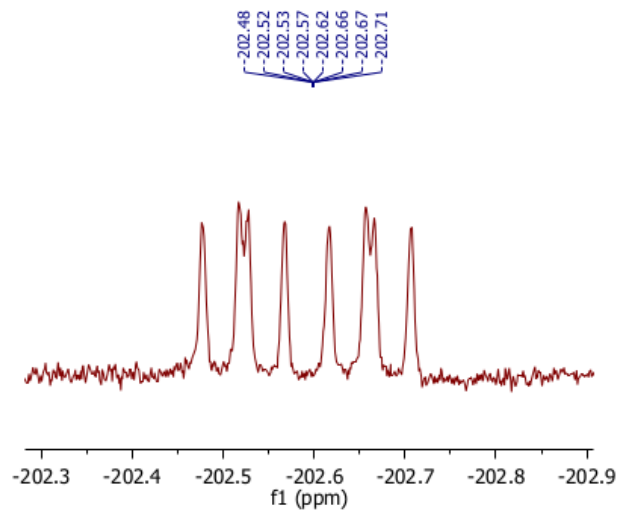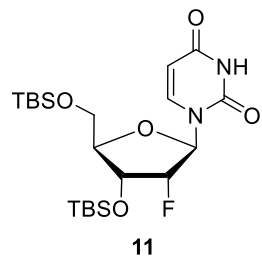
Compound 5, ¹H-NMR, 400 MHz, CDCl₃
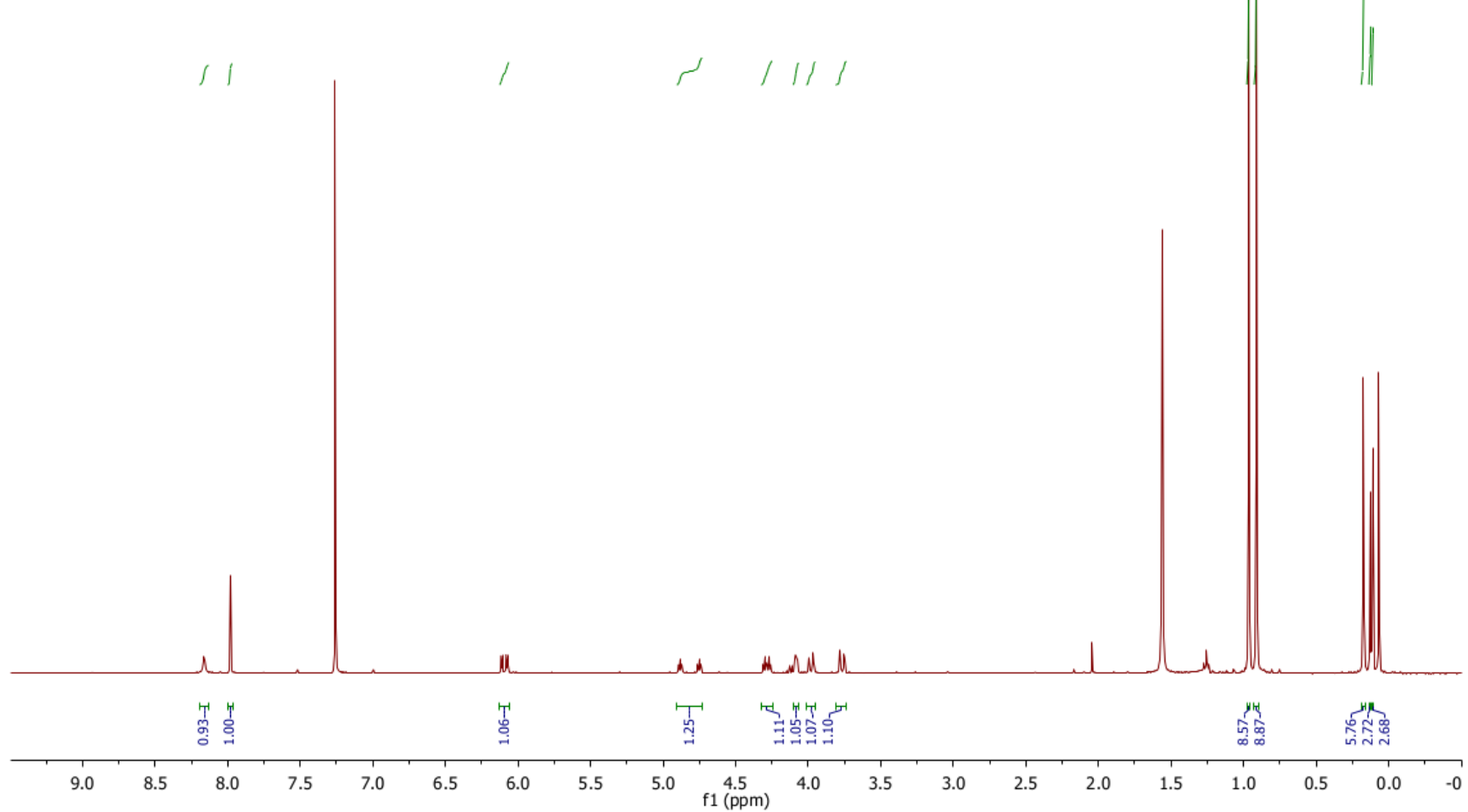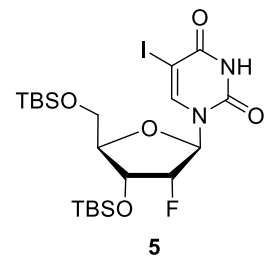
Compound 5, ^{13}C-NMR, 4800 MHz, CDCl$_3$
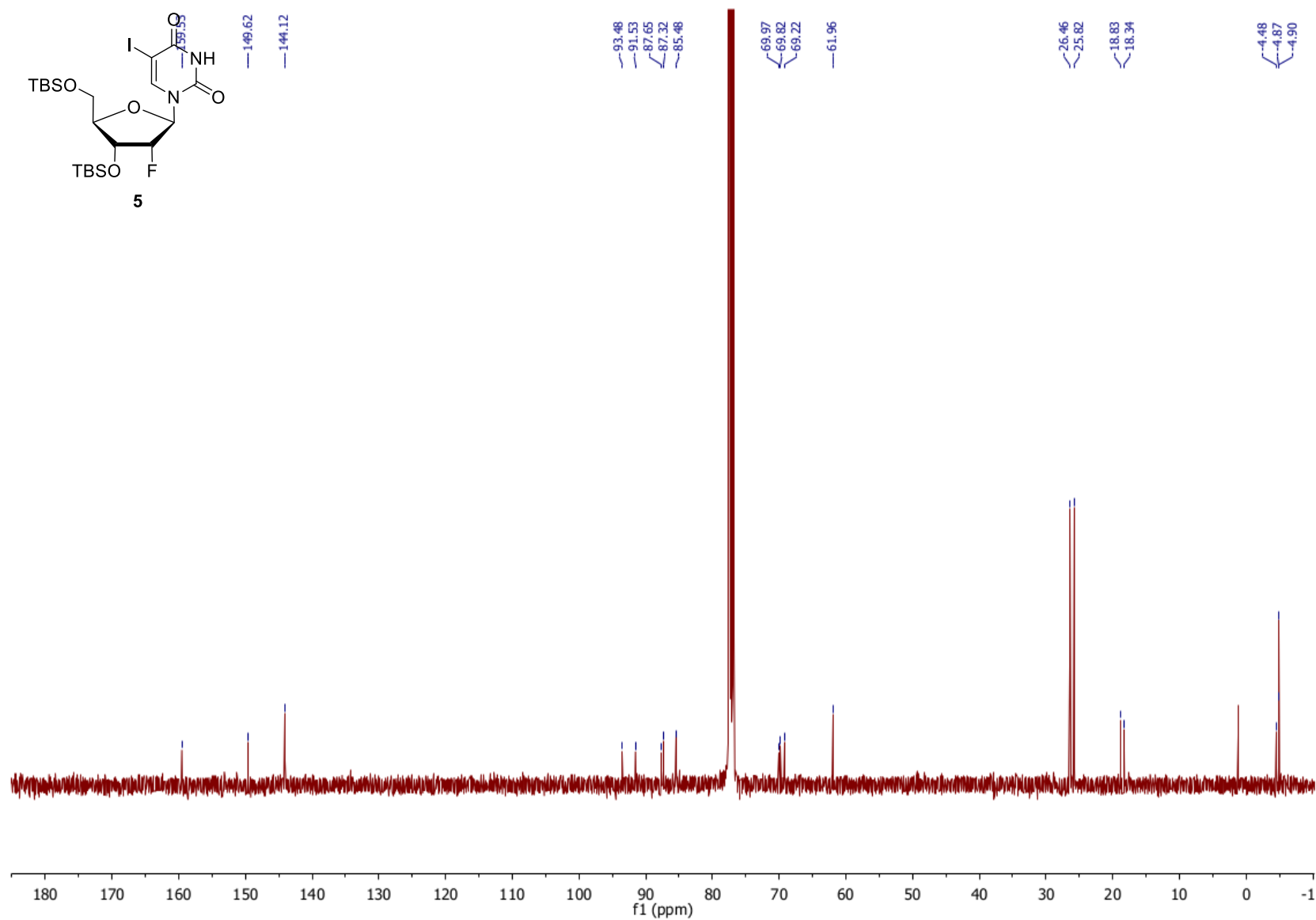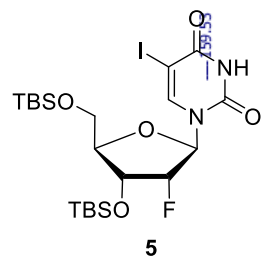
Compound 5, ^{19}F-NMR, 377 MHz, CDCl$_3$

-205.60
-205.64
-205.67
-205.74
-205.78
-205.81

-205.5   -205.6   -205.7   -205.8   -205.9
f1 (ppm)
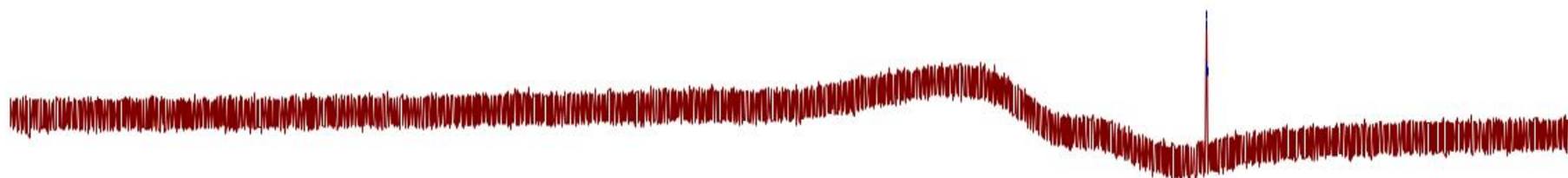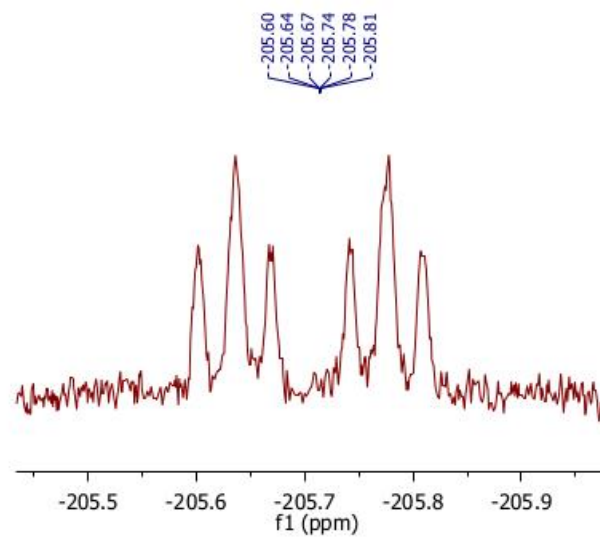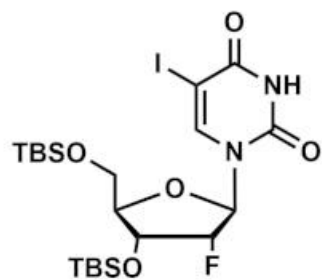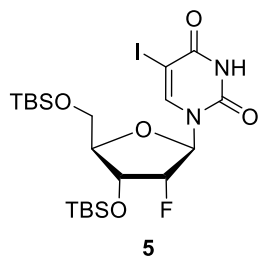
Compound 6, ¹H-NMR, 400 MHz, CDCl₃
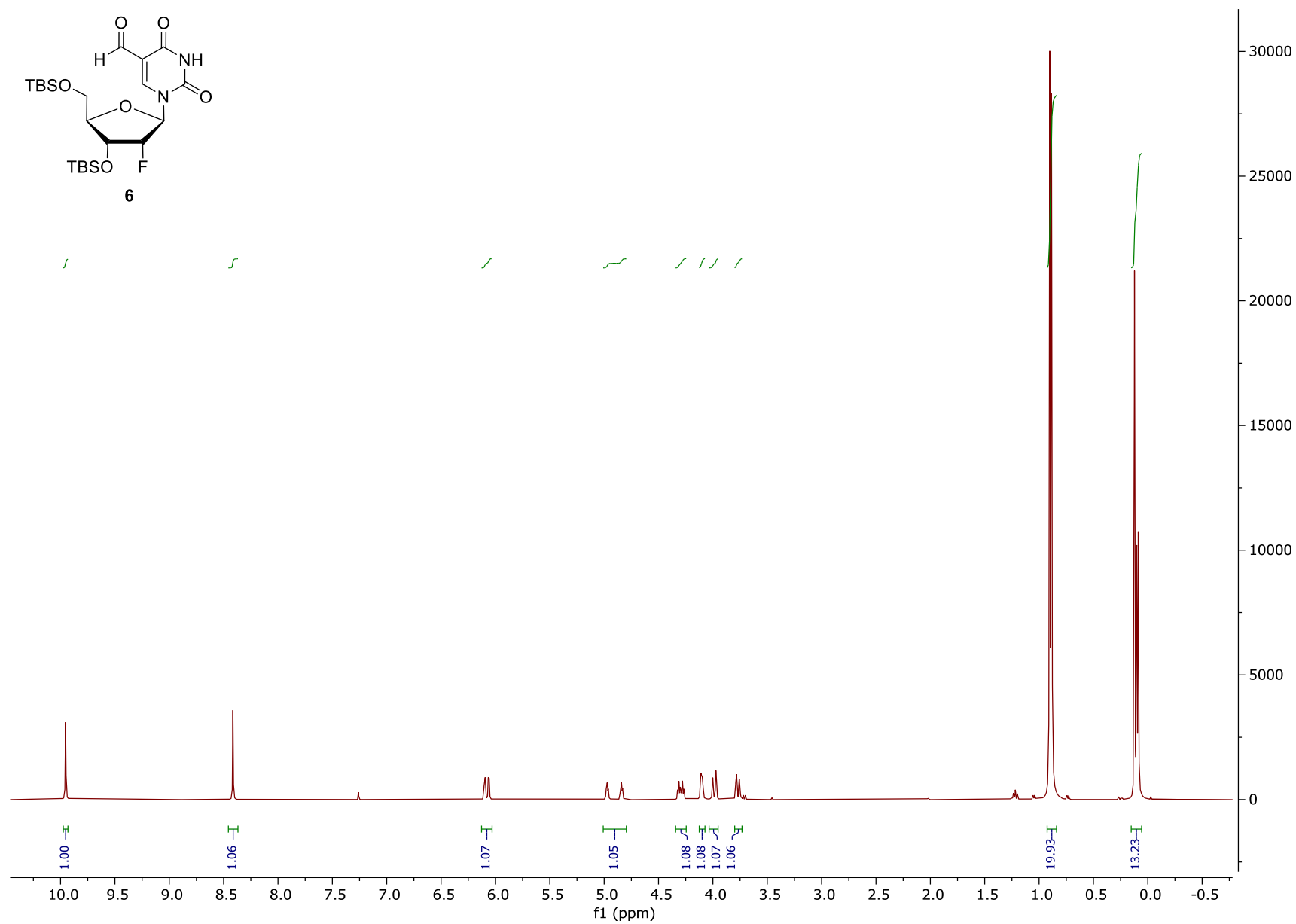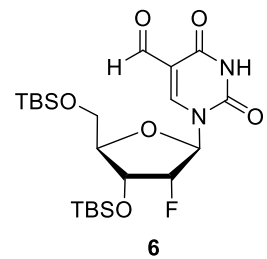
Compound 6, ^{13}C-NMR, 400 MHz, CDCl$_3$
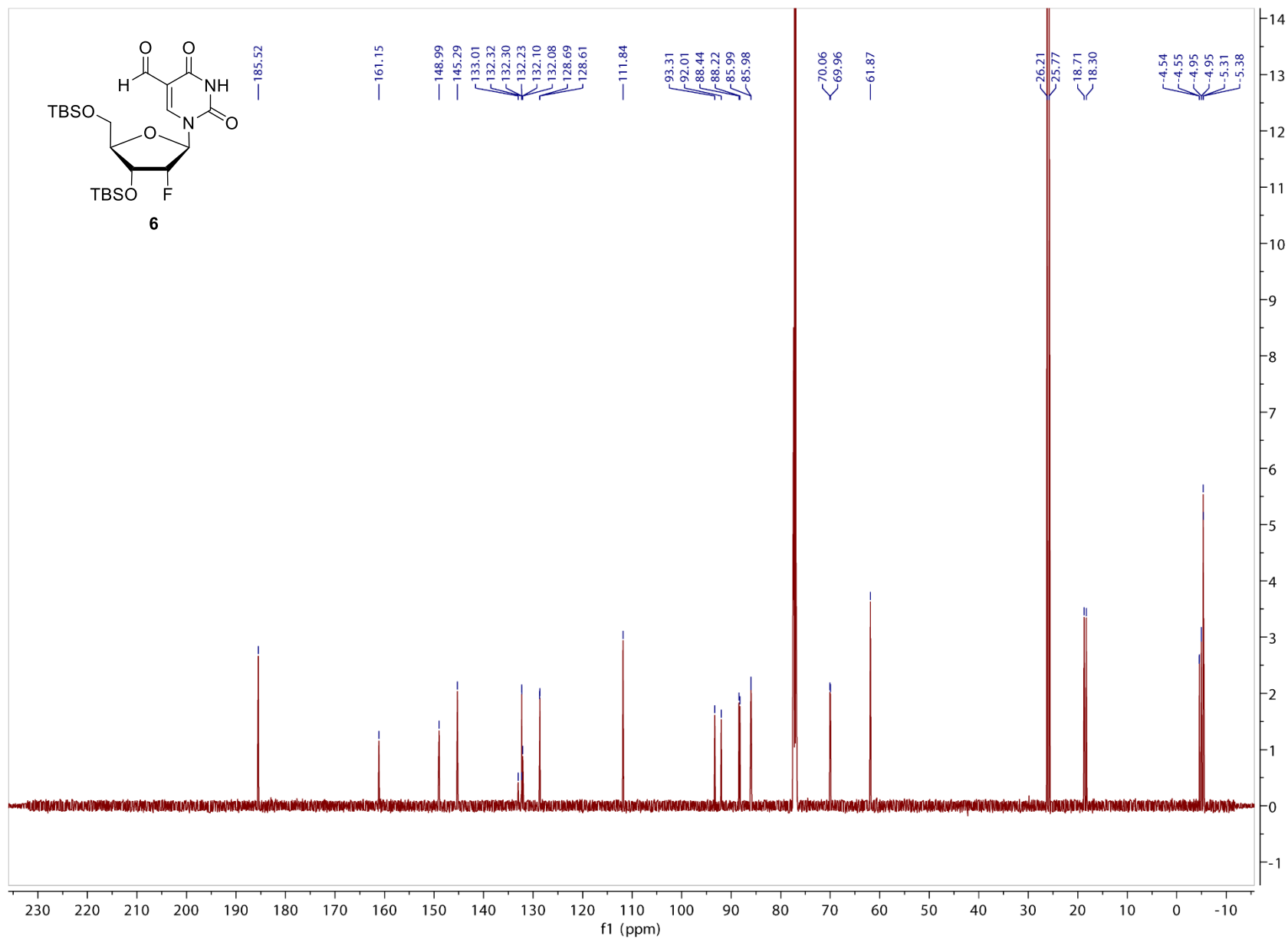
Compound 6, ^{19}F-NMR, 377 MHz, CDCl$_3$
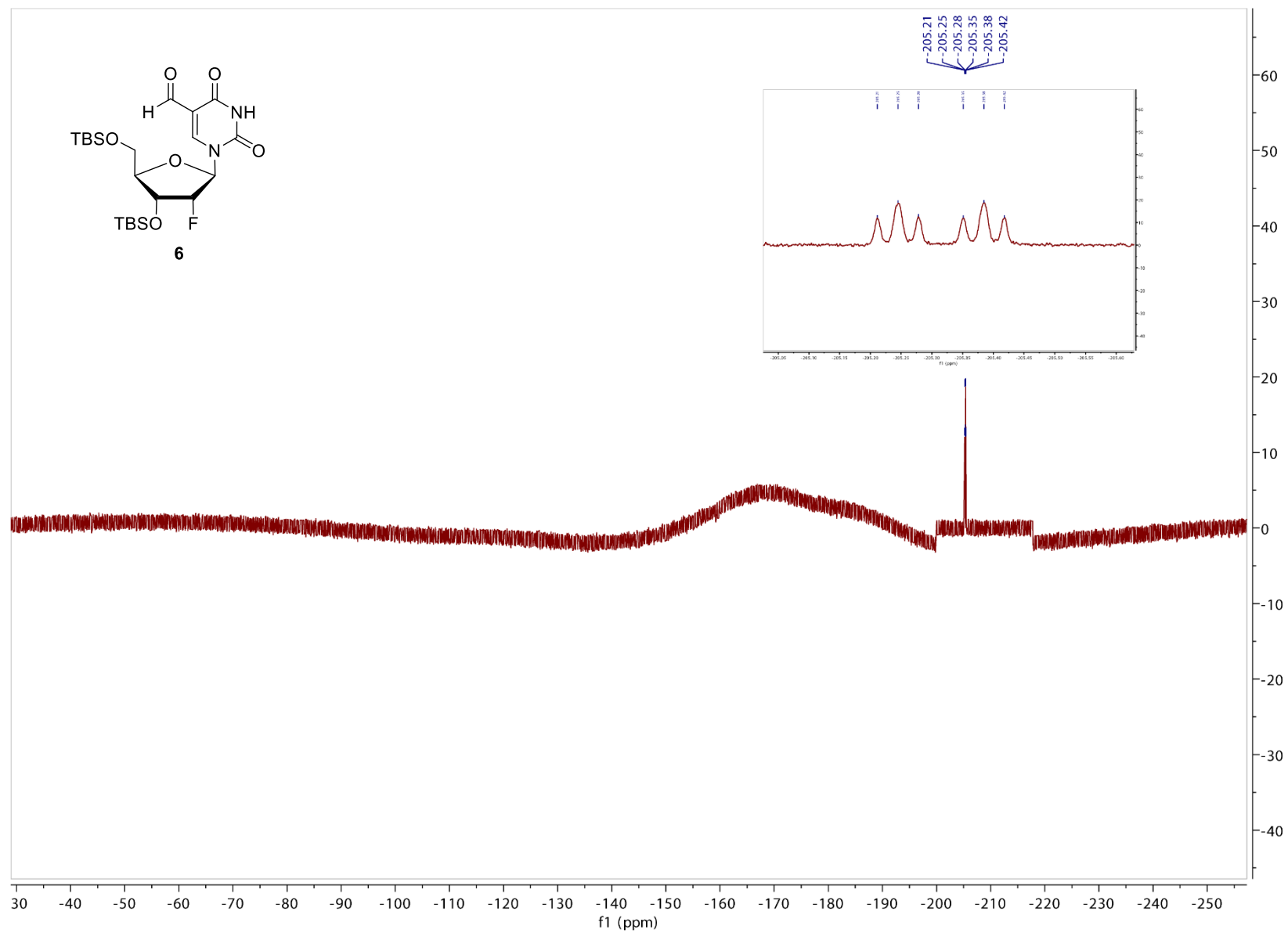
Compound 7, ^{1}H-NMR, 400 MHz, CDCl$_3$
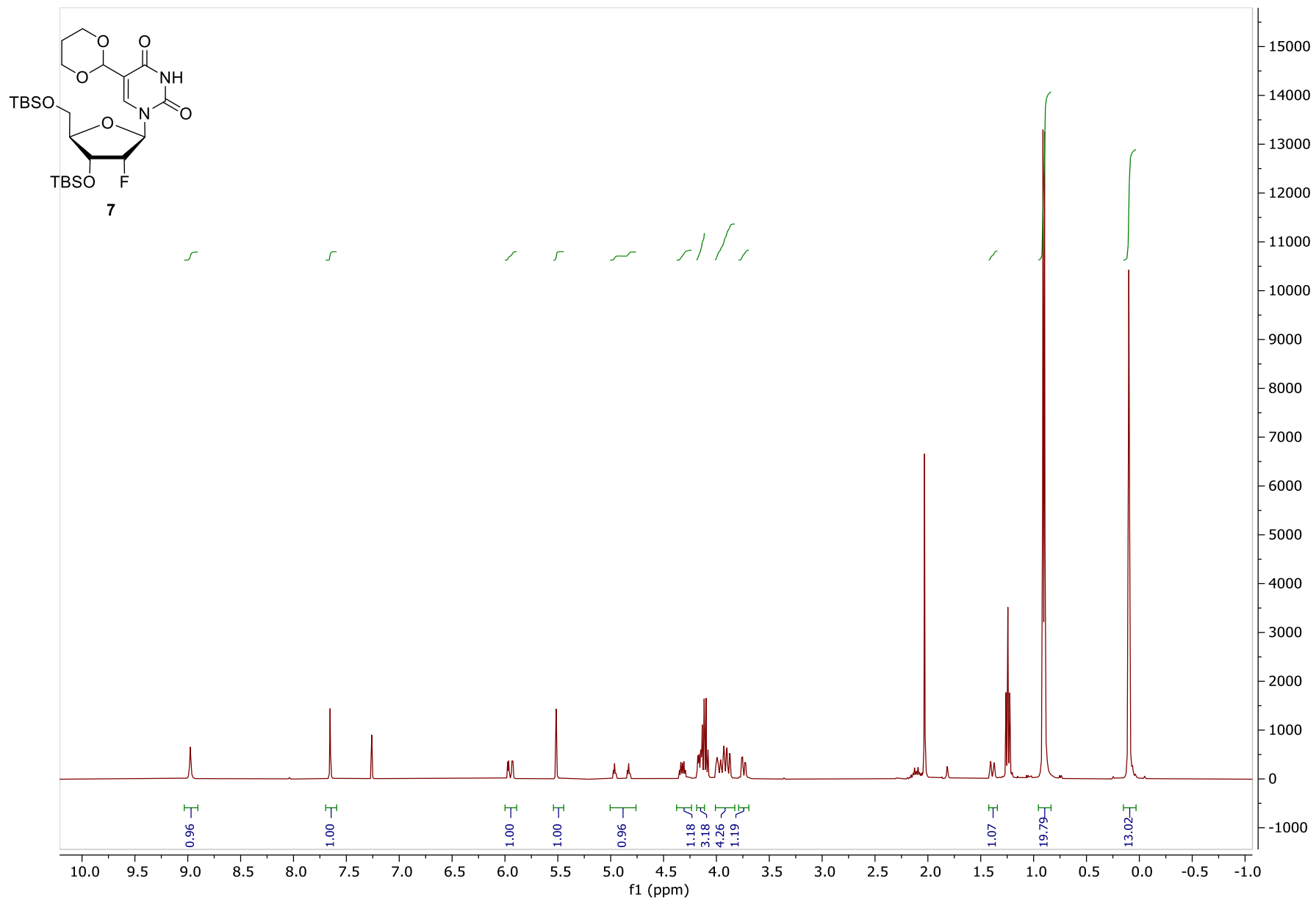
Compound 7, ^{13}C-NMR, 400 MHz, CDCl$_3$

161.08
149.59
139.66
113.24
95.48
92.44
91.15
88.81
88.59
85.01
69.98
69.87
67.68
67.67
62.26
26.24
25.85
25.69
18.68
18.35
4.55
4.92
5.13
5.44

7

f1 (ppm)
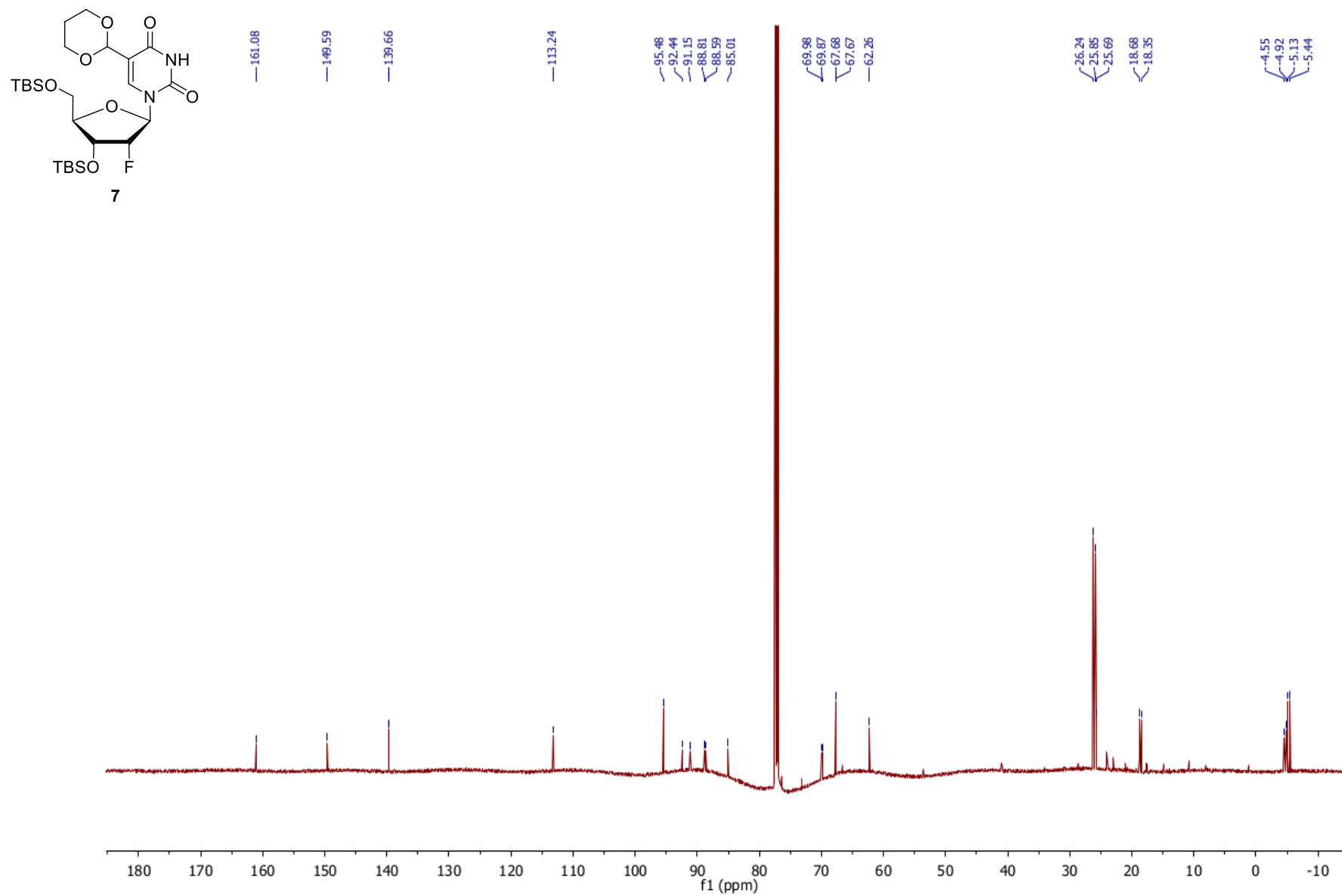
Compound 7, ^{19}F-NMR, 377 MHz, CDCl$_3$

-204.58
-204.61
-204.62
-204.65
-204.72
-204.75
-204.76
-204.79

43
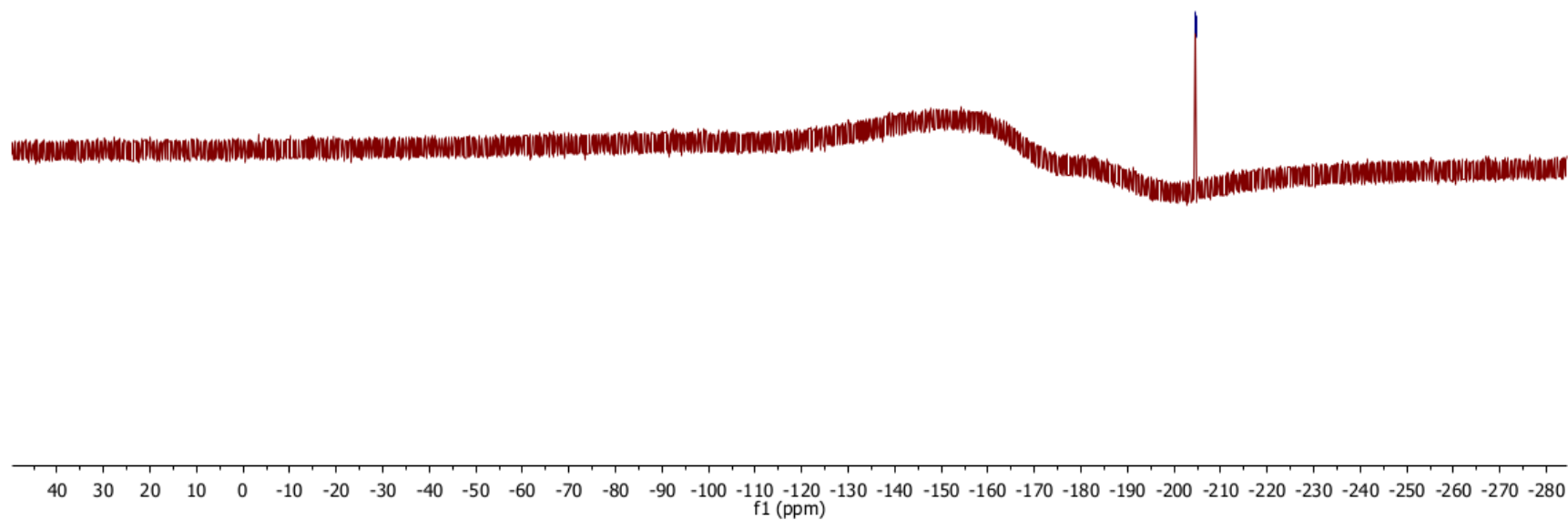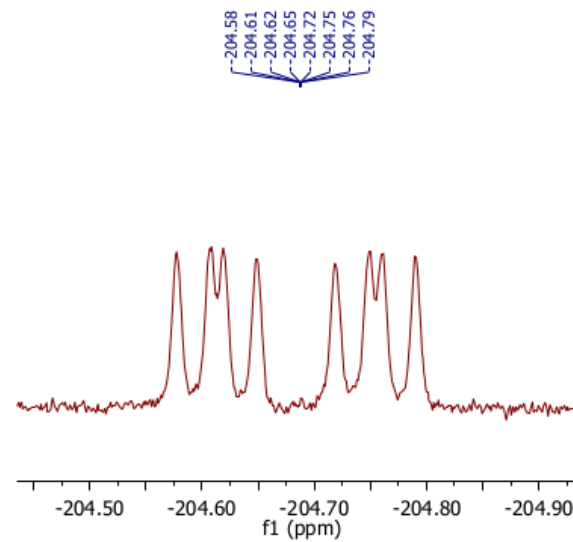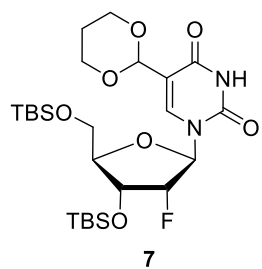
Compound 8, ^{1}H-NMR, 400 MHz, CDCl$_3$
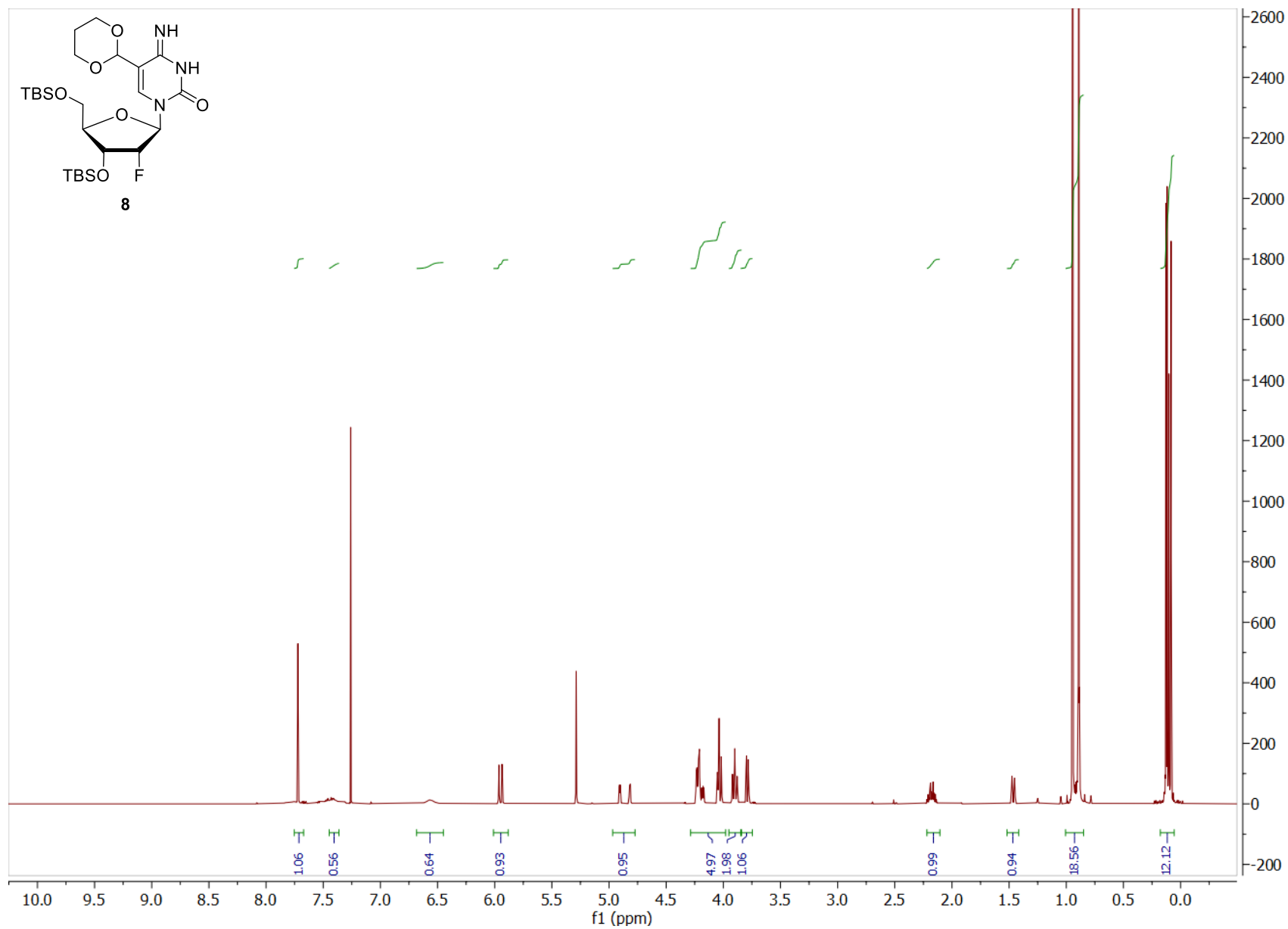
Compound 7, ^{13}C-NMR, 600 MHz, CDCl$_3$
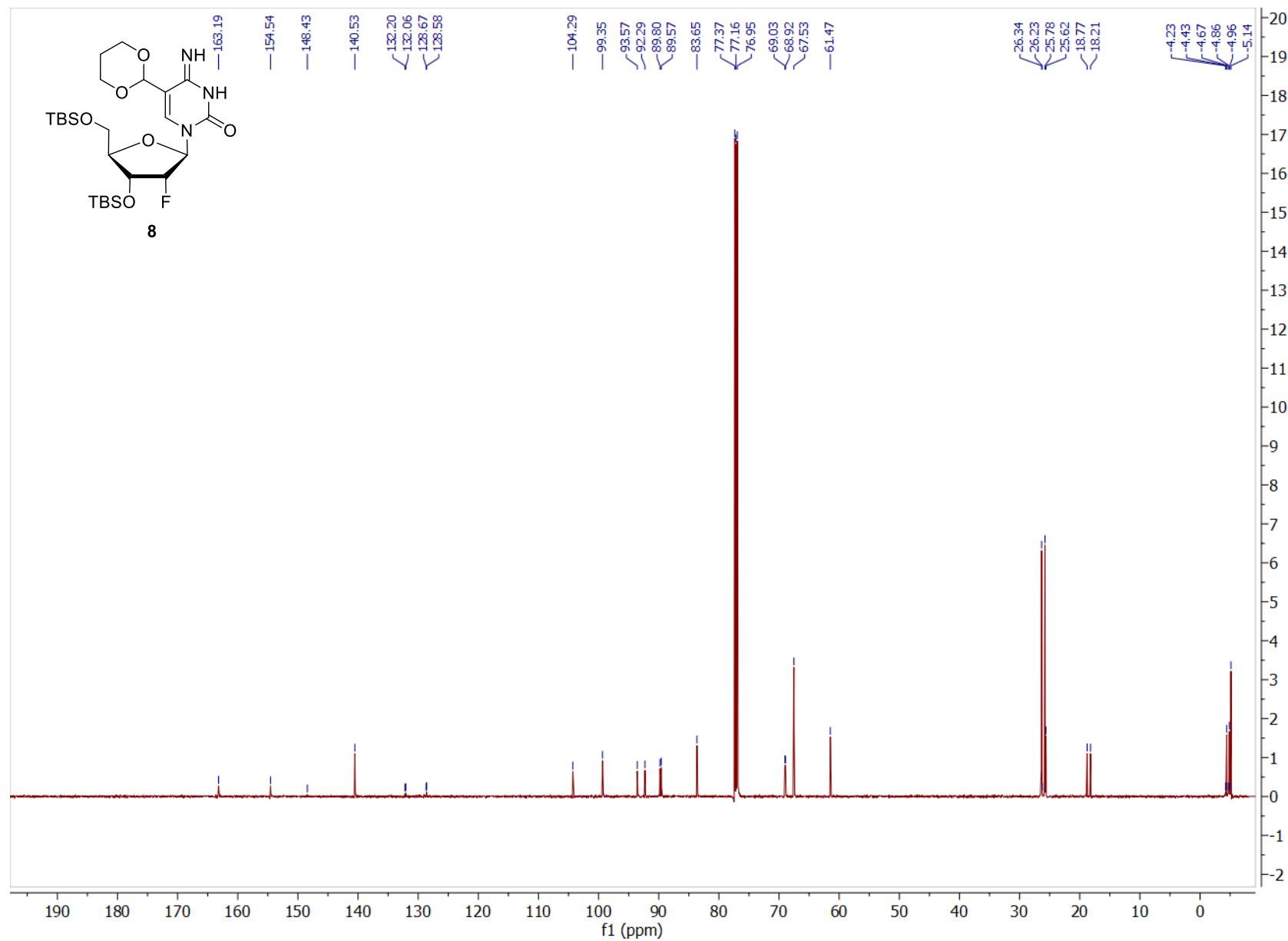
Compound 1, ¹H-NMR, 400 MHz, D₂O
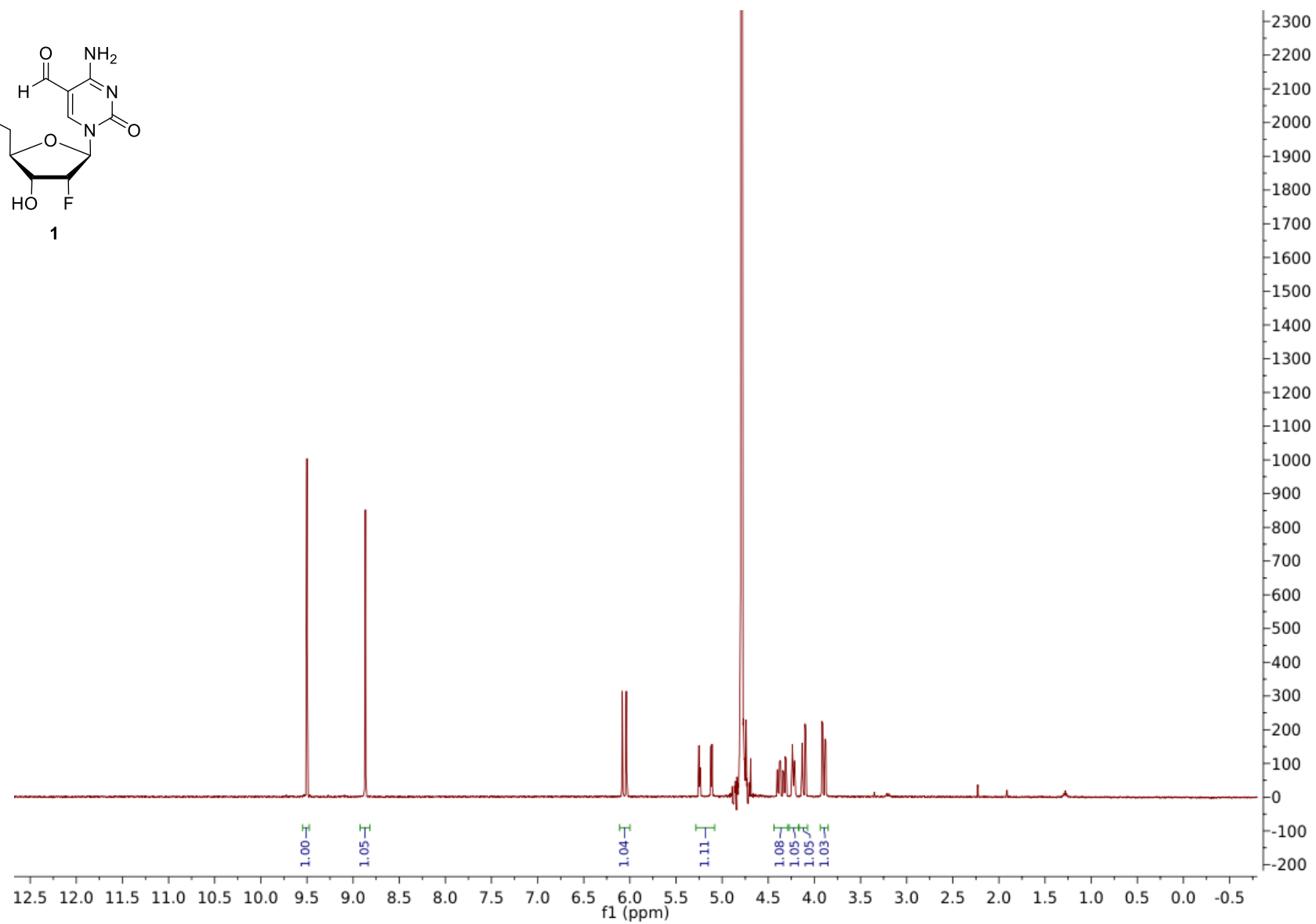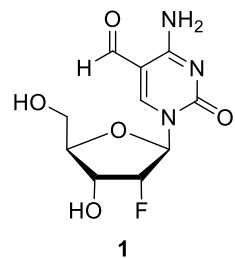
Compound 1, ^{13}C-NMR, 800 MHz, D$_2$O
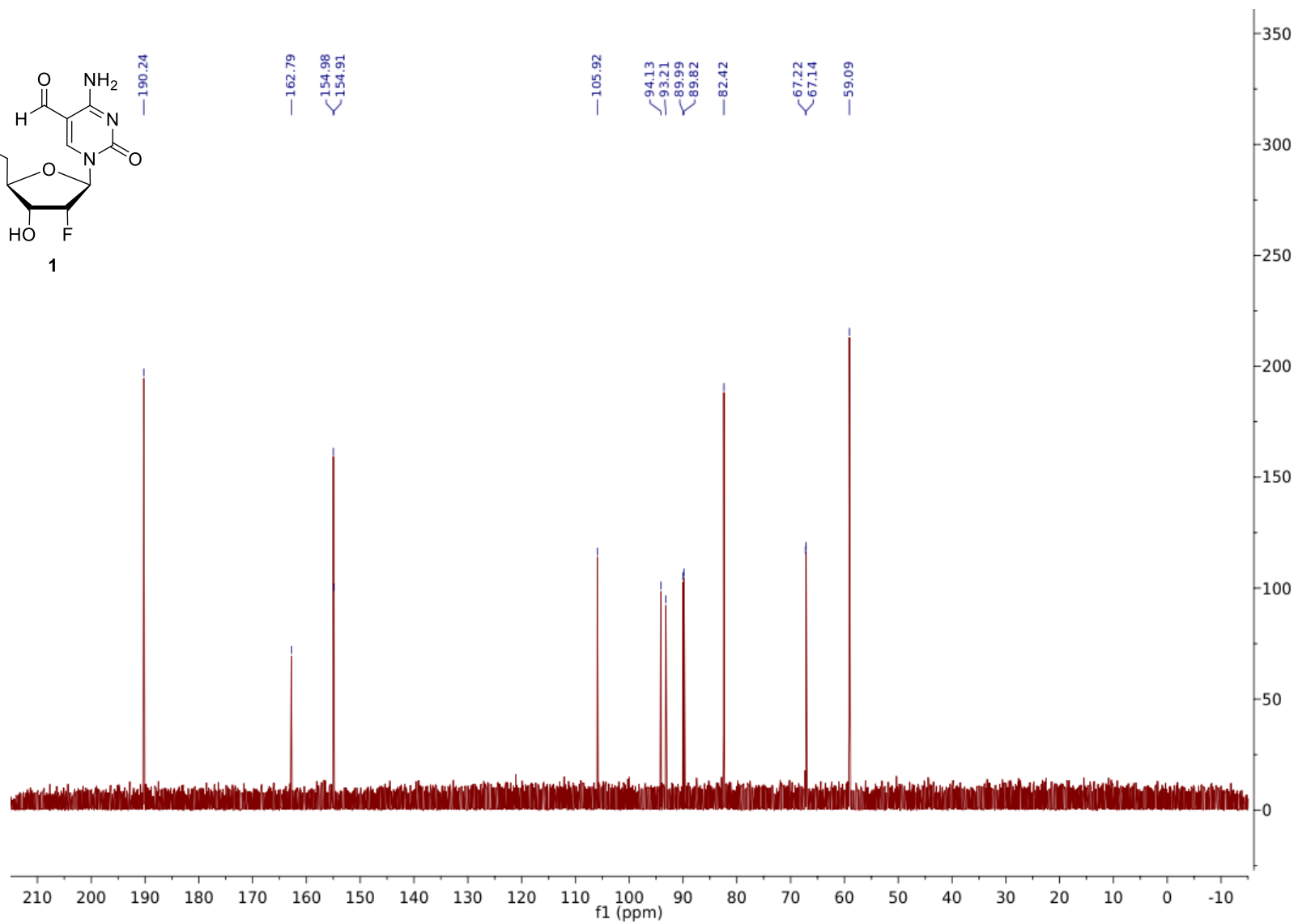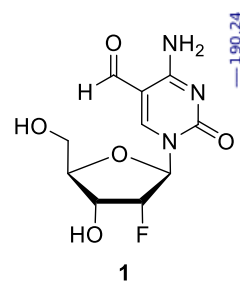
Compound 1, ^{19}F-NMR, 400 MHz, D$_2$O

201.49
201.53
201.55
201.60
201.63
201.67
201.69
201.74
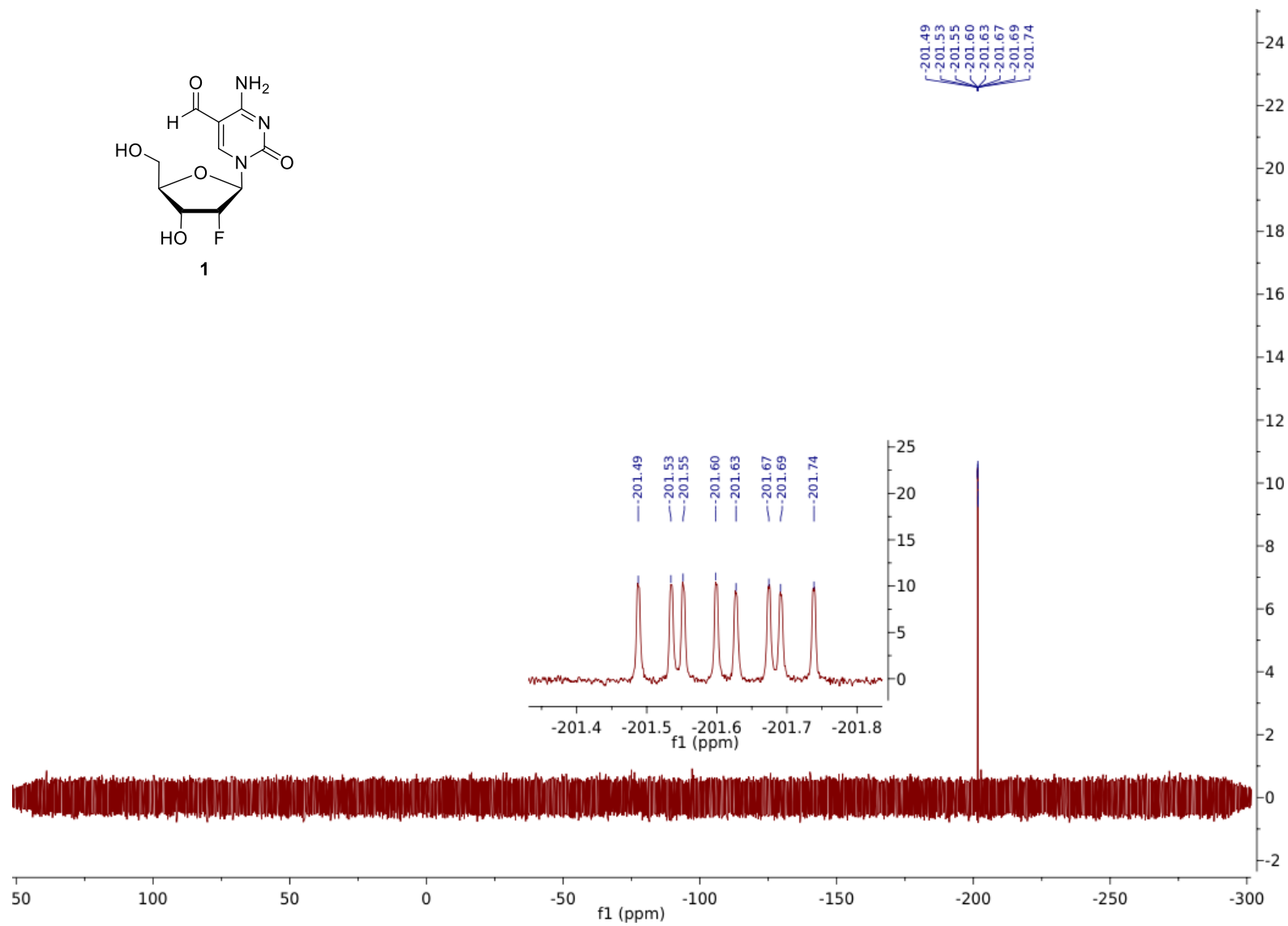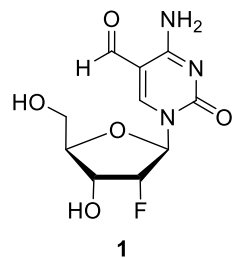
References

Uncategorized References

[1]   W. J. Lowary PT, *J Mol. Biol.* **1998**, *276*, 19-42.
[2]   K. Iwan, R. Rahimoff, A. Kirchner, F. Spada, A. S. Schröder, O. Kosmatchev, S. Ferizaj, J. Steinbacher, E. Parsa, M. Müller, T. Carell, *Nat. Chem. Biol.* **2018**, *14*, 72-78.
[3]   V. Busskamp, N. E. Lewis, P. Guye, A. H. Ng, S. L. Shipman, S. M. Byrne, N. E. Sanjana, J. Murn, Y. Li, S. Li, M. Stadler, R. Weiss, G. M. Church, *Mol. Syst. Biol.* **2014**, *10*, 760.
[4]   F. R. Traube, S. Schiffers, K. Iwan, S. Kellner, F. Spada, M. Müller, T. Carell, *Nat. Protoc.* **2019**, *14*, 283-312.
[5]   A. S. Schröder, O. Kotljarova, E. Parsa, K. Iwan, N. Raddaoui, T. Carell, *Org. Lett.* **2016**, *18*, 4368-4371.

Supporting Information

Intragenomic Decarboxylation of 5-Carboxy-2′-deoxycytidine

Ewelina Kamińska⁺, Eva Korytiaková⁺, Andreas Reichl, Markus Müller, and Thomas Carell*

anie_202109995_sm_miscellaneous_information.pdf

Table of Contents

Supplementary figures

Figure SI-1: Digestion efficiency of the DNA oligonucleotides containing F-cadC using degradase mix (Zymo Research) when compared to NEB digestion mix (New England BioLabs). The degradase in our digestion test showed the highest efficiency and was used for all the experiments. The result is compliant with the literature[1].
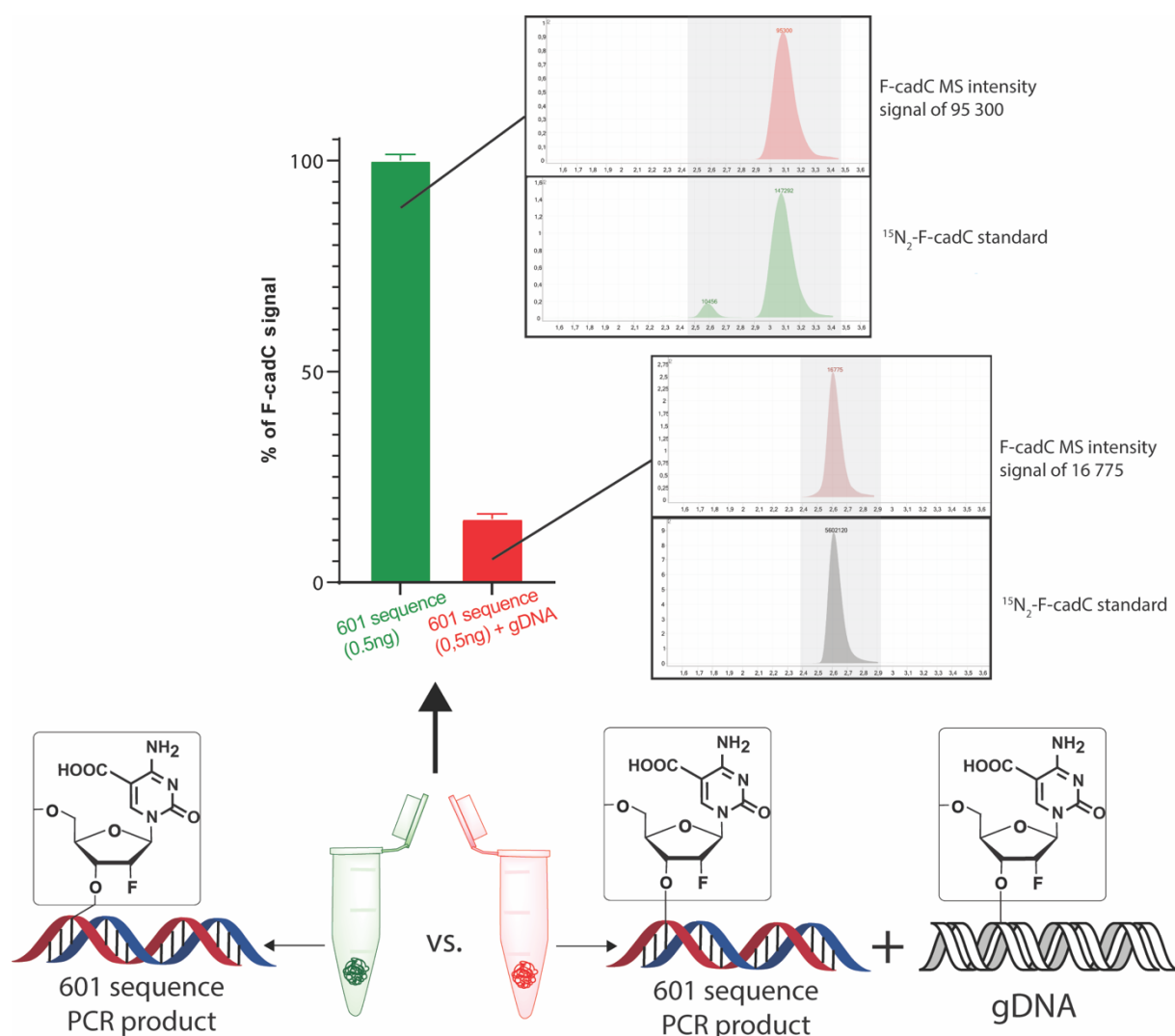
Figure SI-2: The graphical representation of the degree of signal suppression when F-cadC is present in the genomic DNA context. The green bar represents the signal intensity of the digested 0.5 ng of 601 Widom sequence[2] (containing 2'-F-cadC at 30 dC positions) on its own. The red bar shows 2'-F-cadC signal in the same sequence when digested together with 10μg of gDNA. The peaks are aligned with their isotopically labelled standards to confirm their identity. Any measurable quantity of the isotopically labelled standard when present within the sample also strongly supresses the signal, making it impossible to quantify the F-cadC amount using internal calibration and producing unreliable data when compared to external calibration standard curves.

601 sequence:

5´-ATCGATGTATATATCTGACACGTGCCTGGAGACTAGGGAGTAATCCCCTTGGCGGTTAAAA CGCGGGGGGACAG-3´
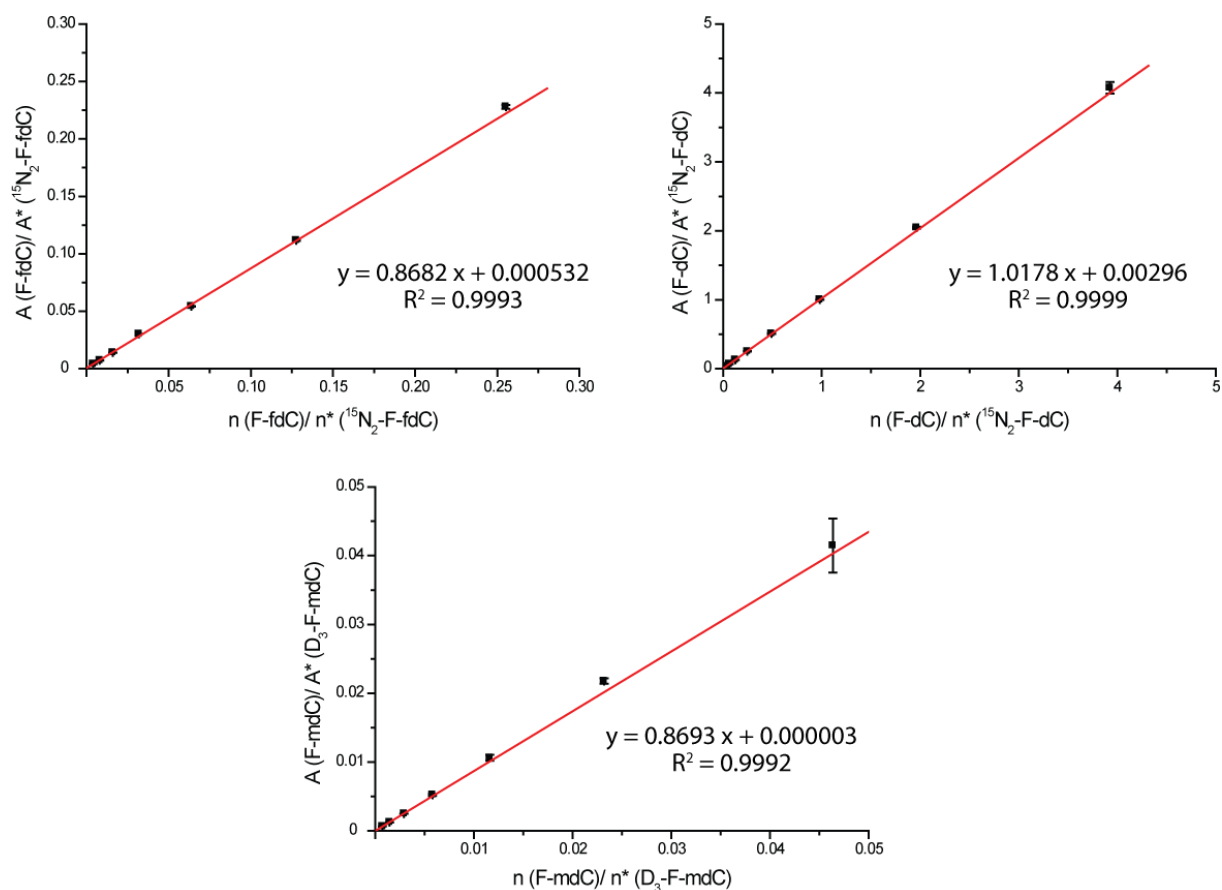
Figure SI-3: Internal calibration curves for the exact quantification of 2'-F-fdC (**5**), 2'-F-dC (**2**) and 2'-F-mdC (**7**) with the corresponding linear equation and coefficient of determination. [3]

0.2 fmol - 12.7 fmol

6.3 fmol - 202.5 fmol
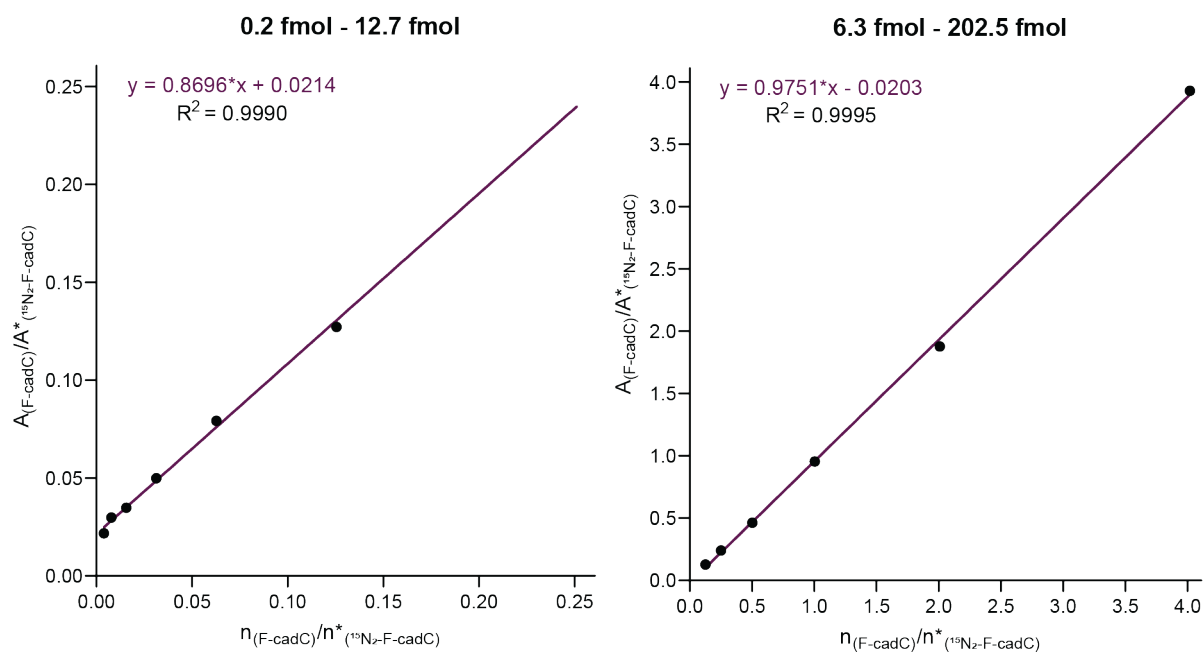
Figure SI-4: Internal calibration curves for the exact quantification of 2'-F-cadC (**1**), with the corresponding linear equation and coefficient of determination.
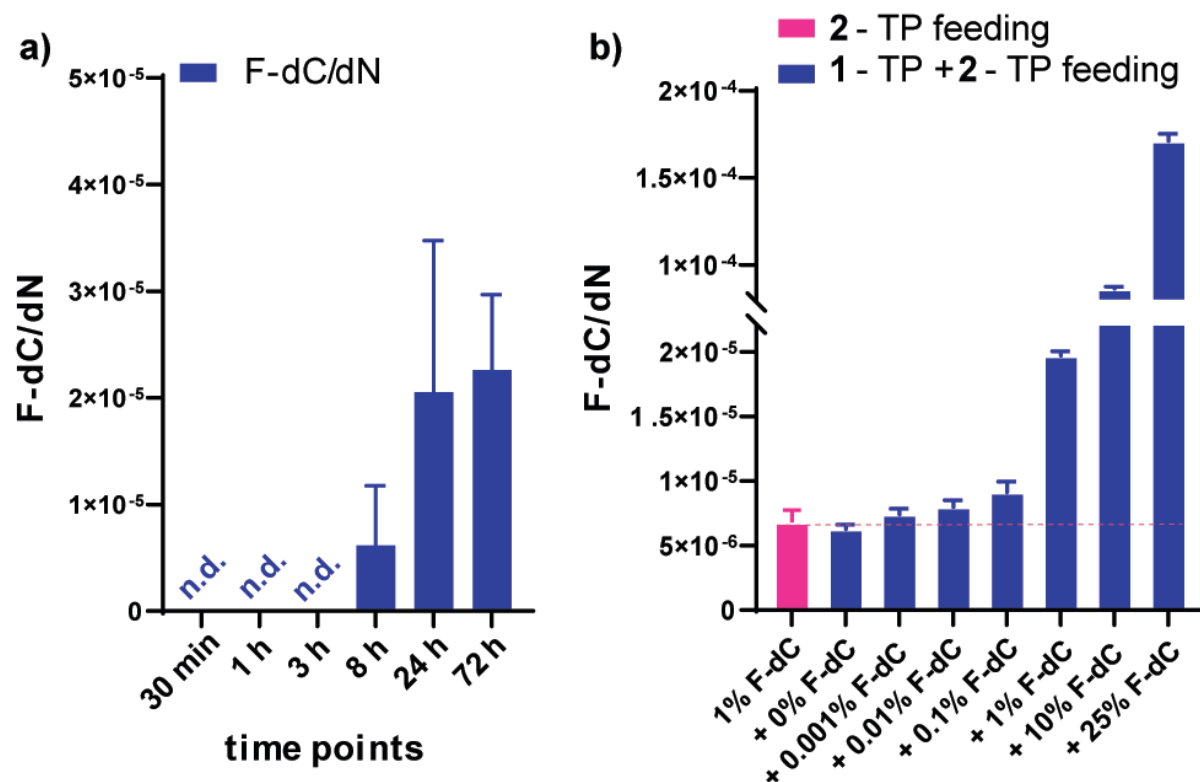
Figure SI-5: a) The pulse-chase study of the decarboxylation product 2'F-dC **2** formation upon 100μM of **1**-TP feeding to Neuro-2a cells for 30mins. Quantifiable amount of **2** are visible only after 8h from the start of the feeding. The error bars represent the propagated standard deviation of 3 biological replicates. b) The results of the control 24-hour experiment when different concentrations of "impurity" of **2**-TP were co-fed with 100μM of **1**-TP. The pink bar represents the cells fed only with 1% (1μM) **2**-TP. We can see that using 1μM of **2**-TP only ("1% F-dC") we get the incorporation rate marginally higher than in the feeding of 100μM of **1**-TP alone ("+0% F-dC"). In turn, when we combine both samples 1μM of **2**-TP and 100μM of **2**-TP ("+1% F-dC"), the intragenomic levels of **2** are about three times higher than in the 100μM **2**-TP ("+0% F-dC") feeding. The error bars show the standard deviation of a technical triplicate.

If decarboxylation was happening in the cytosol, not in the genome, we would have expected the turnover to happen to a higher extent than 1%. This result, together with the soluble pool investigation (Figure 3a in the main text) that shows no detectable F-dC in the cytosol hints at the observed decarboxylation as an intragenomic process.

TDG-dependent cadC levels
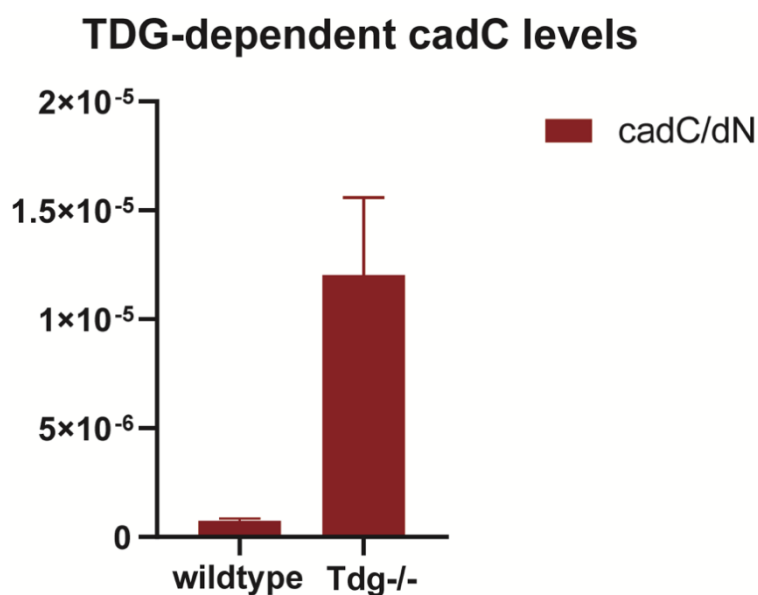
Figure SI-6: The natural cadC levels in the genome are usually very low in the presence of TDG. This was also the case when we measured cadC in the wt E14 mESCs. TDG knockout of the same cell line shows cadC levels two orders of magnitude higer than in the wildtype cells. This points out at the efficiency of the BER and assures us that in order to investigate the direct decarboxylation, we need a model system that excludes Tdg function. Using fluorine-modified nucleosides allows us to achieve this goal. From our previous studies, we know already that the levels of intragenomic 2'F-dC (**2**) formed upon 2'F-fdC (**5**) feeding shows the same result for both: wt and Tdg-/- cells. This confirms the lack of 2'F-fdC recognition by Tdg. The error bars represent the propagated error of the standard deviation.
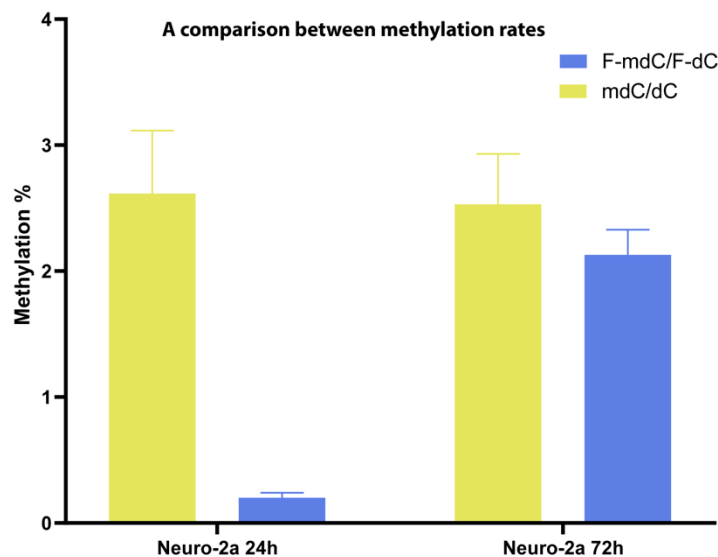
Figure SI-7: The methylation levels of natural dC and F-dC in the Neuro-2a cell line upon 24h and 72h feeding of 100μM F-cadCTP. Since the natural methylation levels show values ~2.5%, confirmed also in the unfed control, the F-mdC/F-dC reach about 0.2% after 24h and 2.1% after 72h. This data confirms that shorter feedings do not allow us to see full methylation, however, after three days we reach the values close to the natural methylation, thus decarboxylation happens in the CpG and non-CpG context. The error bars represent standard deviation of three technical replicates.
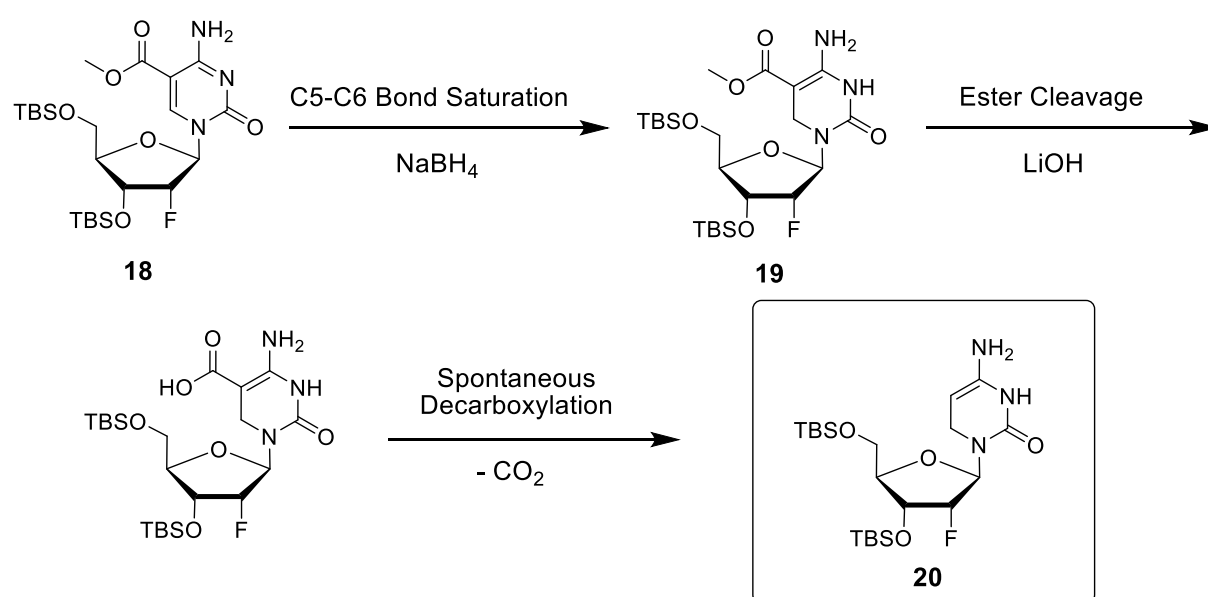
Figure SI-8: Spontaneous Decarboxylation of TBS-protected, 5-6 saturated F-cadC (Synthesis below). F-dC is first TBS protected with TBSCl and imidazole, before being iodinated with I_2 and oxidiser CAN. Afterwards, CO-insertion yields TBS protected 5-methoxycarbonyl-FdC, which is then saturated at the C5-C6 bond with $NaBH_4$. Ester cleavage with LiOH then leads to spontaneous decarboxylation towards TBS-protected 5,6-saturated FdC **20**.

Methods and Materials

Chemical Synthesis

Unless noted otherwise, all chemical transformations were performed using flame- or oven dried glassware under an atmosphere of argon. Chemicals were purchased from *Sigma-Aldrich*, *TCI*, *Fluka*, *ABCR*, *Carbosynth* or *Acros Organics* and used without further purification. Technical grade solvents were distilled prior to extraction or chromatography. Reactions and column chromatography fractions were monitored by thin-layer chromatography (TLC) on silica gel F_{254}TLC plates from Merck KGaA. Visualization of the developed TLC plates was achieved through UV-absorption or through staining with Hanessian's stain or $KMnO_4$ stain. Flash column chromatography was performed on Geduran®Si60 (40-63 µm) silica gel from Merck KGaA applying slight nitrogen pressure. NMR spectra were recorded on *Bruker Avance III HD 400 (*400 MHz), *Varian NMR-System 600* (600 MHz) and *Bruker Avance III HD* with *Cryo-Kopf 800* (800 MHz) spectrometers. ^{1}H-NMR shifts were calibrated to the residual solvent resonances: $CDCl_3$ (7.26 ppm), DMSO-d_6 (2.50ppm) and D_2O (4.79 ppm). ^{13}C-NMR shifts were calibrated to the residual solvent: $CDCl_3$ (77.16 ppm), DMSO-d_6 (39.52 ppm). All NMR spectra were analysed using the program MestReNova 10.0.1 from Mestrelab NMR data are reported as follows: chemical shift (multiplicity, coupling constants where applicable, number of hydrogens, assignment). Splitting is reported with the following symbols: s = singlet, d = doublet, t = triplet, dd = doublet of doublets, dt = doublet of triplets, ddd = doublet of doublets of doublets, m = multiplet. Whenever a signal is not clearly distinguishable from related nuclides, asterisks (e.g. *, **) are attached to indicate interchangeability. Research S. L. Low resolution mass spectra were measured on a LTQ FT-ICR by Thermo Finnigan GmbH. High resolution mass spectra were measured by the analytical section of the Department of Chemistry of Ludwigs-Maximilians-Universität München on aMAT 90 (ESI) from Thermo Finnigan GmbH. IR spectra were recorded on a PerkinElmer Spectrum BX II FT-IR system. Substances were applied as a film or directly as solids on the ATR unit. Analytical RP-HPLC was performed on an analytical HPLC Waters Alliance (2695 Separation Module, 2996 Photodiode Array Detector) equipped with the column Nucleosil120-2 C18 from Macherey Nagel applying an eluent flow of 0.5 mL/min. Preparative RP-HPLC was performed on HPLC Waters Breeze (2487 Dual λ Array Detector, 1525 Binary HPLC Pump) equipped with the column VP 250/32 C18 from Macherey Nagel. A flowrate of 5mL/min was applied. HPLC grade MeCN was purchased from VWR.

Cell culture

DMEM high glucose (Sigma Aldrich) containing 10% heat-inactivated FBS (Gibco), 100 U/mL Penicillin-Streptomycin (Gibco), 2 mM L-glutamine were used to culture rat neuroblastoma Neuro-2a cells.

DMEM high glucose and Ham's Nutrient Mixture F12 (Sigma Aldrich) mixture 50/50 containing 100 U/mL Penicillin-Streptomycin (Gibco), containing 10% heat-inactivated FBS (Gibco), 2 mM L-glutamine were used to culture CHO-K1 cells.

mES cells (J1, E14, R1) were cultured in three different media – slitting, maintenance (a2i) and priming (C/R) media. The maintenance medium keeps stem cells in naïve pre-implantation state. It is consisted of DMEM high glucose (Sigma Aldrich), 10% heat-inactivated FBS (Gibco), 0.1mM β-mercaptoethanol (Sigma Aldrich), 2mM L-Alanyl-L-Glutamine 200mM (Gibco), 1x Nonessential amino acids 100x (Sigma Aldrich) and 100 U/mL Penicillin-Streptomycin (Gibco). The culture medium consists of the splitting medium supplemented with LIF 1000U/mL (ORF Genetics), 3.0 µM CGP 77675 (Sigma Aldrich) and 3.0 µM GSK3 inhibitor CHIR-99021 (Sigma Aldrich). The priming media consisted of the splitting media supplemented with LIF 1000U/mL (ORF Genetics), 2.5 µM tankyrase inhibitor IWR-1 endo (Sigma Aldrich) and 3.0 µM CHIR-99021 (Sigma Aldrich). The cells were primed for 3 days via plating in C/R medium and fed with the **1**-TP during the last 24 hours of priming.

Metabolic-labelling experiments

Triphosphate feedings were conducted by plating cells in their dedicated medium for 24h in 6 cm petri dishes to obtain a confluency of 70-80% prior to adding the compound. The NTP transporter was purchased from Merck and resuspended to a stock solution of 5mM with autoclaved ddH$_2$O. The feeding was performed in a ratio of 1:1 transporter to triphosphate concentration in tricine buffer (prepared according to manufacturer's protocol) in the total volume of 450 µL. The medium was first removed from the dish, the cells were washed once with PBS (Sigma Aldrich) and the transporter + PPP mixture was applied for 30 minutes for somatic cells and 15mins for the stem cells at 37°C, 5% CO$_2$. Afterwards the cells were again washed with PBS and fresh medium was applied.

The study of the

Cell Harvesting and Isolation of genomic DNA

All samples were washed with PBS (Sigma Aldrich) once before harvesting and lysed directly in the plates by adding RLT buffer (Qiagen) supplemented with 400µM 2,6-di-tert-butyl-4-methylphenol (BHT) and desferoxamine mesylate (DM) both from Sigma Aldrich. Next, DNA was sheered in MM400 bead mill (Retsch) at 30 Hz for 1 min in 2mL microcentrifuge tubes (Eppendorf) with 5mm diameter stainless steel beads (one per tube) and centrifuged at 21000rcf for 5 minutes. Genomic DNA (gDNA) was extracted using Zymo Quick gDNA mini-prep® kit according to the protocol with an addition of RNAse A treatment (Qiagen) at 0.2mg/mL in Genomic Lysis Buffer – 400µL of the solution transferred directly on the column and incubated for 15 minutes. All samples were eluted in DNAse-free ddH$_2$O with 20µM BHT, the concentration of gDNA was measured on Nanodrop.

DNA digestion

While investigating the decarboxylation *in vivo* using 2´-F-cadC-TP (**1**-TP) as a metabolic label, the following conditions were used for the gDNA digestion procedure: 1 – 10 µg of genomic DNA or oligonucleotide in 35 µL H$_2$O were digested as follows: 2 µL of Degradase plus (*Zymo research*), 5 µL of Degradase plus reaction buffer (*Zymo research*), 0.5 µL of Benzonase nuclease (*Merck, Serratia marcescens*) and a specific amount of isotopically labelled internal standards were added. Before adding the digestion mixture, the gDNA was heated to 95ºC for 5 minutes and placed directly on ice. The mixture was incubated for 4 h at 37 °C, diluted with 450 µL of water and then extracted with chloroform (500 µL) twice. The samples were lyophilised and resuspended in 50 µL ddH$_2$O. Prior to UHPLC-MS/MS analysis, samples were filtered by using an AcroPrep Advance 9 filter plate 0.2 µm Supor (*Pall Life Science*).

For digestion of DNA samples using the Nucleoside Digestion Mix from New England BioLabs, a solution of 1 - 10 µg genomic DNA or oligonucleotide in 35 µL of water was incubated with 5 µL of the Nucleoside Digestion Mix Reaction Buffer (10x), 2.5 µL of the Nucleoside Digestion Mix and a specific amount of isotopically labelled internal standards for 2 h at 37 °C. The sample was filtered using an AcroPrep Advance 96 filter plate 0.2 µm Supor from Pall Life Sciences and subsequently analyzed by UHPLC-MS/MS.

LC/MS-MS analysis of DNA samples[4]

Quantitative UHPLC-MS/MS analysis of digested DNA samples was performed using an Agilent 1290 UHPLC system equipped with a UV detector and an Agilent 6490 triple quadrupole mass spectrometer. Nucleosides of interest were quantified using the stable isotope dilution technique.[3] The source-dependent parameters were as follows: gas temperature 80 °C, gas flow 15 L/min (N_2), nebulizer 30 psi, sheath gas heater 275 °C, sheath gas flow 11 L/min (N2), capillary voltage 2.500 V in the positive ion mode, capillary voltage −2.250 V in the negative ion mode and nozzle voltage 500 V. The fragmentor voltage was 380 V/ 250 V. Delta EMV was set to 500 (positive mode) and 800 (negative mode). Compound-dependent parameters are summarized in Supplementary Tables 1-3. Chromatography was performed by a Poroshell 120 SB-C8 column (Agilent, 2.7 µm, 2.1 mm × 150 mm) at 35 °C using a gradient of water and MeCN, each containing 0.0085% (v/v) formic acid, at a flow rate of 0.35 mL/min: 0 →4 min; 0 →3.5% (v/v) MeCN; 4 →7.9 min; 3.5 →5% MeCN; 7.9 →8.2 min; 5 →80% MeCN; 8.2 →11.5 min; 80% MeCN; 11.5 →12 min; 80 →0% MeCN; 12 →14 min; 0% MeCN. The autosampler was cooled to 4 °C. The injection volume was amounted to 39 µL.

Quantification of nucleosides in DNA samples

Nucleosides were quantified using internal calibration curves and the stable isotope dilution technique as described in the literature. [4] Used internal calibration curves for F-nucleosides are shown in Supplementary Figures 3 and 4.

Soluble nucleoside pool extraction and purification

The soluble pool extraction protocol started with a cell harvest using typical cell passaging trypsinisation technique using TrypLE reagent, Gibco. The cells were counted with trypan blue via Countess II FL Automated Cell Counter, Life Technologies (ThermoFisher Scienitfic) and pelleted via centrifugation at 260g for 3min. The pellet was resuspended with the ice-cold mixture of acetonitrile and water (1:1 ratio), 1mL per $1x10^6$ cells and gently vortexed. The cell suspension was incubated on ice for 15min and centrifuged at 10 000g for 10min at 4°C. The supernatant containing the cytosolic fraction was collected and lyophilised, while the cell pellet was treated with the lysis buffer for subsequent gDNA isolation.

Samples were dissolved in H_2O (100 µL H_2O for soluble pool from 400 000 cells) and treated with Antarctic Phosphatase (New England BioLabs) for 30 min. Supel-Select HBL 60 mg columns were equilibrated with 1.5 mL MeOH and 3 mL H_2O (pH = 4). Samples were loaded

the surface of the column, washed with 4 mL H_2O (pH = 4). Cartridges were dried under high vacuum for 15 min. Sample was eluted with 1.5 mL MeOH/MeCN = 1:1. Eluent was lyophilized, dissolved in H_2O and analysed by UHPLC-MS/MS as described for genomic DNA samples.[3]

Synthetic procedures

The nucleoside triphosphates were synthesised according to literature. [5-6]
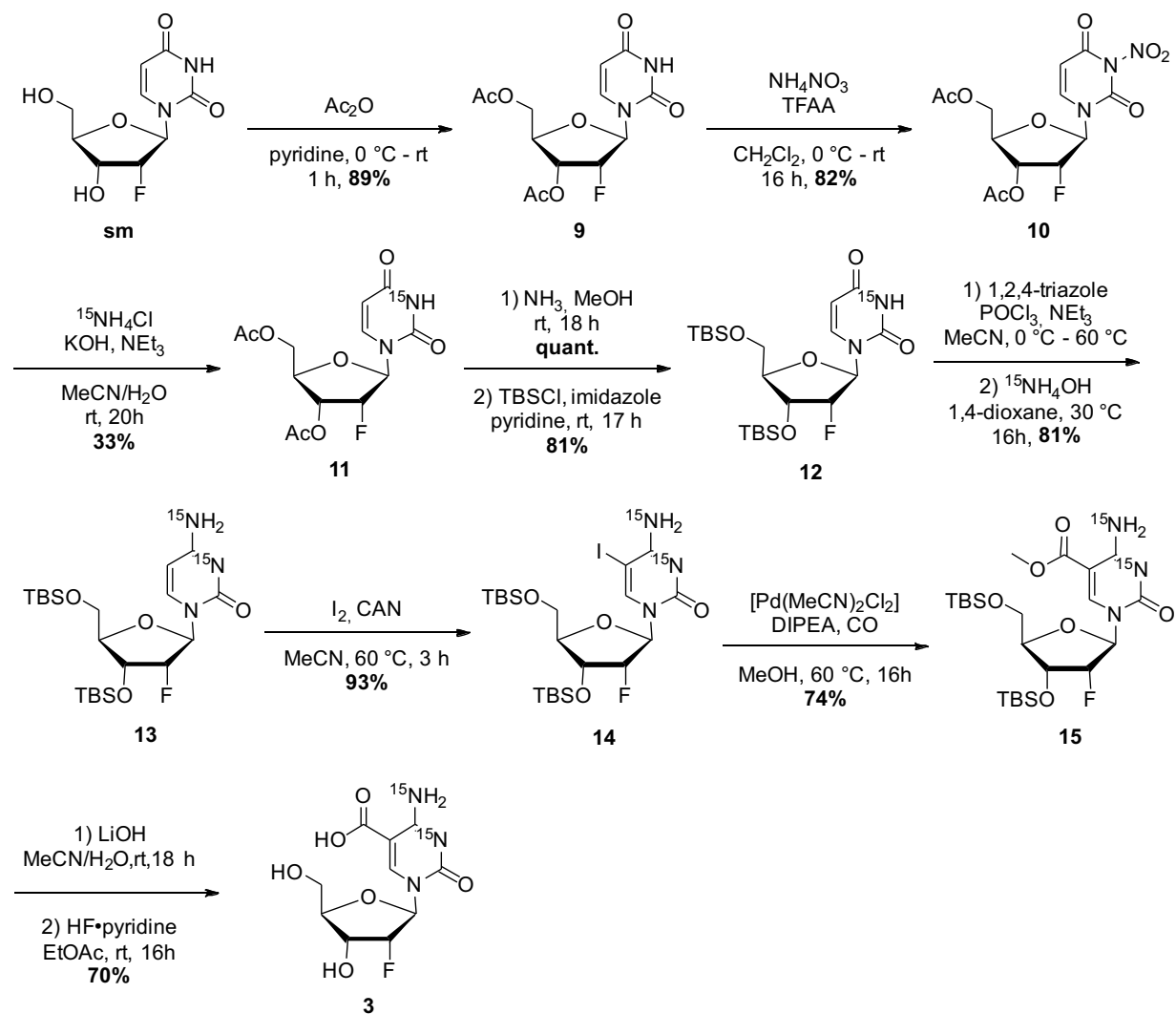
Scheme SI-1: Synthesis pathway towards $^{15}N_2$-2'-F-cadC (**3**).

3',5'-di-O-Acetyl-2'-deoxy-2'-(R)-fluorouridine (9)

9
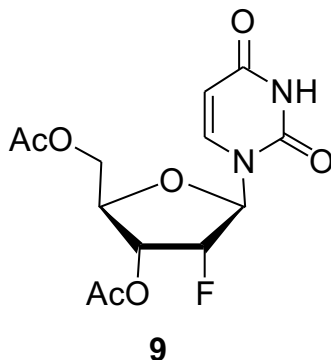
2'-Deoxy-2'-fluorouridine (2.40 g, 9.75 mmol, 1.0 equiv.) was dissolved in pyridine (100 mL), cooled to 0 °C and acetic anhydride (19.2 mL, 68.3 mmol, 7.0 equiv.) was added. The reaction mixture was stirred for 1 h at 0 °C, warmed to room temperature and stirred for additional 16 h. Afterwards, pyridine was removed under reduced pressure, the resulting yellow oil was coevaporated with toluene (3 x 200 mL) and taken up in an aqueous, saturated NaHCO$_3$-solution (100 mL). The solution was extracted with CH$_2$Cl$_2$ (3 x 150 mL), the combined organic layers were washed with an aqueous, saturated NaCl solution (300 mL) and dried over Na$_2$SO$_4$. The mixture was concentrated *in vacuo*. The resulting crude was purified by flash chromatography (SiO$_2$, CH$_2$Cl$_2$ → CH$_2$Cl$_2$/MeOH = 99:1) to yield compound **9** as a colourless solid (2.85 g, 8.63 mmol, 89%).
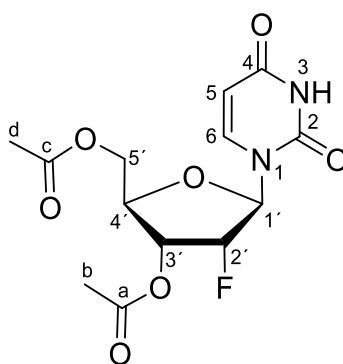
R$_f$ = 0.72 (SiO$_2$, CH$_2$Cl$_2$/MeOH = 20:1).

^{1}H-NMR (600 MHz, CDCl$_3$): δ/ppm = 8.24 (s, 1H, C3-H), 7.39 (d, ^3J = 8.2 Hz, 1H, C6-H), 5.88 (dd, $^3J_{H-F}$ = 19.4 Hz, ^3J = 1.5 Hz 1H, C1'-H), 5.76 (dd, ^3J = 8,2 Hz, ^3J = 2.2 Hz, 1H, C5-H), 5.37 (ddd, $^2J_{H-F}$ = 52.2 Hz, ^3J = 5.0 Hz, ^3J = 1.5 Hz, 1H, C2'-H), 5.15 (ddd, $^3J_{H-F}$ = 17.9 Hz, ^3J = 8.2 Hz, ^3J = 5.0 Hz, 1H, C3'-H), 4.45 (dd, ^2J = 12.4 Hz, ^3J = 2.5 Hz, 1H, C5'-H), 4.43–4.40 (m, 1H, C4'-H), 4.31 (dd, ^2J = 12.4 Hz, ^3J = 4.5 Hz, 1H, C5'-H), 2.17 (s, 3H, C$_d$-H/C$_b$-H)*, 2.11 (s, 3H, C$_d$-H/C$_b$-H)*.

^{13}C-NMR (101 MHz, CDCl$_3$): δ/ppm = 170.6 (C$_c$), 170.2 (C$_a$), 162.6 (C4), 149.8 (C2), 140.9 (C6), 103.4 (C5), 91.7 (d, $^2J_{C-F}$ = 35.6 Hz, C1'), 91.3 (d, $^1J_{C-F}$ = 191.9 Hz, C2'), 79.1 (C4'), 70.0 (d, $^2J_{C-F}$ = 15.6 Hz, C3'), 62.6 (C5'), 21.2 (C$_b$/C$_d$)*, 20.8 (C$_b$/C$_d$)*.

^{19}F-NMR (126 MHz, CDCl$_3$): δ/ppm = -199.1 (dt, $^2J_{H-F}$ = 52.3 Hz, $^3J_{H-F}$ = 18.7 Hz, C2'-*F*).

HRMS (ESI): calc. for C$_{13}$H$_{16}$FN$_2$O$_7^+$ [M+H]$^+$: 331.09361, found: 331.09363.

The data are consistent with literature.[3]

3',5'-di-*O*-Acetyl-2'-deoxy-2'-(*R*)-fluoro-3-nitrouridine (10)

10
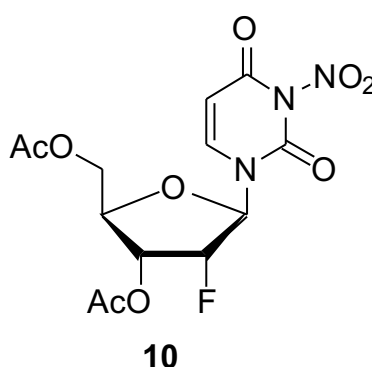
NH$_4$NO$_3$ (1.84 g, 23.0 mmol, 2.0 equiv.) was ground with a mortar, suspended in CH$_2$Cl$_2$ (65 mL) and cooled to 0 °C for 20 min before trifluoroacetic anhydride (6.49 mL, 46.0 mmol, 4.0 equiv.) was added dropwise. The reaction mixture was stirred for 3 h at room temperature until the remaining NH$_4$NO$_3$ was dissolved. After the resulting yellow solution was cooled to 0 °C, compound **9** (3.80 g, 11.5 mmol, 1.0 equiv.) was added in portions. The reaction mixture was stirred for 3 h at 0 °C, warmed to room temperature and stirred for additional 16 h. After complete conversion of the nucleoside, the mixture was neutralised by addition of PBS-buffer (130 mL) and the aqueous solution was extracted with CH$_2$Cl$_2$ (3 x 100 mL). The combined organic layers were dried over Na$_2$SO$_4$ and solvents were removed *in vacuo*. The resulting yellowish oil was purified by flash chromatography (SiO$_2$, CH$_2$Cl$_2$/MeOH = 50:1) to yield nucleoside **10** as a colourless solid (3.52 g, 9.38 mmol, 82%).
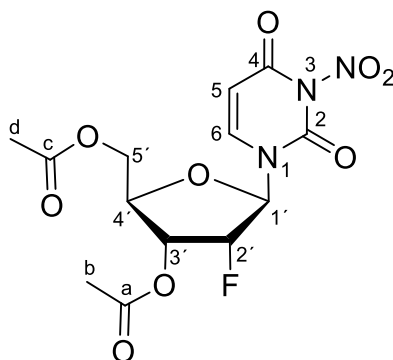
R$_f$ = 0.44 (SiO$_2$, CH$_2$Cl /MeOH = 9:1).

^{1}H-NMR (400 MHz, CDCl$_3$): δ/ppm = 7.46 (d,^3J = 8.4 Hz, 1H, C6-H), 5.90 (d,^3J = 8.4 Hz, 1H,C5-H), 5.81 (dd, $^3J_{H-F}$ = 18.7 Hz, ^3J = 1.6 Hz, 1H, C1'-H), 5.42 (ddd, $^2J_{H-F}$ = 51.8 Hz, ^3J = 4.9 Hz, ^3J = 1.6 Hz, 1H, C2'-H), 5.14 (ddd, $^3J_{H-F}$ = 18.0 Hz, ^3J = 8.4 Hz, ^3J = 4.9 Hz, 1H, C3'-H), 4.49 – 4.42 (m, 2H, C5'-H, C4'-H), 4.36 – 4.29 (m, 1H, C5'-H), 2.17 (s, 3H, C$_b$-H/C$_d$-H)*, 2.11 (s, 3H, C$_b$-H/C$_d$-H)*.

^{13}C-NMR (101 MHz, CDCl$_3$): δ/ppm = 170.2 (C$_c$), 169.8 (C$_a$), 154.8 (C4), 145.1 (C2), 140.1 (C6), 102.1 (C5), 92.1 (d, $^2J_{C-F}$ = 36.2 Hz, C1'), 90.8 (d, $^1J_{C-F}$ = 193.2 Hz, C2'), 79.3 (C4'), 69.3 (d, $^2J_{C-F}$ = 15.6 Hz, C3'), 61.8 (C5'), 20.8 (C$_b$/C$_d$)*, 20.5 (C$_b$/C$_d$)*.

^{19}F-NMR (376 MHz, CDCl$_3$): δ/ppm = 199.3 (dt, $^2J_{H-F}$ = 51.8 Hz, $^3J_{H-F}$ = 18.4 Hz, C2'-F).

HRMS (ESI): calc. for C$_{13}$H$_{18}$FN$_4$O$_9^+$ [M+NH$_4$]$^+$: 393.10523, found: 393.10527.

The data are consistent with literature. [3]

3',5'-di-*O*-Acetyl-2'-deoxy-2'-(*R*)-fluoro-(*N^3*-^{15}N)-uridine (11)

11
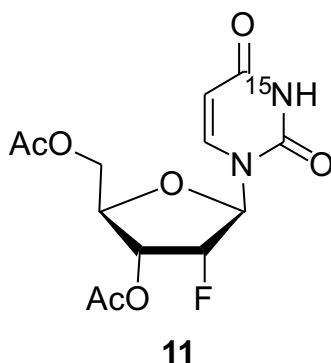
^{15}NH$_4$Cl (0.261 g, 4.79 mmol, 1.3 equiv.) was dissolved in milli-Q-H$_2$O (10 mL), KOH (0.269 g, 4.79 mmol, 1.3 equiv.) and NEt$_3$ (0.89 mL, 6.39 mmol, 1.8 equiv.) were added. A solution of compound **10** (1.36 g, 3.62 mmol, 1.0 equiv.) in MeCN (10 mL) was added to the reaction

mixture and stirred for 20 h at room temperature. After complete conversion, MeCN was removed *in vacuo*. The resulting solution was neutralised with aqueous, saturated NH$_4$Cl (15 mL) and extracted with CH$_2$Cl$_2$ (3 x 100 mL). The combined organic layers were dried over Na$_2$SO$_4$ and concentrated under reduced pressure. The resulting colourless crude was purified by flash chromatography (SiO$_2$, CH$_2$Cl$_2$ → CH$_2$Cl$_2$/MeOH = 97:3). Compound **11** was obtained as a colourless solid (0.402 g, 1.21 mmol, 33%).
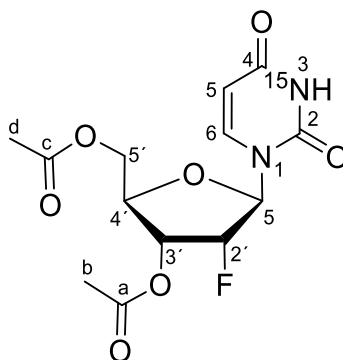
R$_f$ = 0.38 (SiO$_2$, CH$_2$Cl$_2$/MeOH = 99:1).

^{1}H-NMR (400 MHz, CDCl$_3$): δ/ppm = 8.29 (dd, $^2J_{H^{-15}N}$ = 91.2 Hz, ^4J = 2.2 Hz, 1H, N3-H), 7.39 (d, ^3J = 8.2 Hz, 1H, C6-H), 5.83 – 5.74 (m, 1H, C1'-H), 5.77 (d, ^3J = 8.2 Hz, 1H, C5-H), 5.37 (ddd, $^2J_{H-F}$ = 52.2 Hz, ^3J = 4.8 Hz, ^3J = 1.6 Hz, 1H, C2'-H), 5.15 (ddd, $^3J_{H-F}$ = 17.9 Hz, ^3J = 5.3 Hz, ^3J = 2.8 Hz, 1H, C3'-H), 4.47 – 4.43 (m, 1H, C5'-H), 4.41 (dd, ^3J = 8.0 Hz, ^3J = 2.6 Hz, 1H, C4'-H), 4.31 (dd, ^2J = 12.1 Hz, ^3J = 4.2 Hz, 1H, C5'-H), 2.16 (s, 3H, C$_b$-H/C$_d$-H)*, 2.12 (s, 3H, C$_b$-H/C$_d$-H)*.

^{13}C-NMR (101 MHz, CDCl$_3$): δ/ppm = 170.3 (C$_c$), 169.9 (C$_a$), 162.3 (sd, $^1J_{C-N}$ = 9.2 Hz, C4), 149.5 (sd, $^1J_{C-N}$ = 18.3 Hz, C2), 140.6 (C6), 103.1 (sd, $^1J_{C-N}$ = 7.2 Hz, C5), 91.8 (d, $^2J_{C-F}$ = 38.3 Hz, C1'), 90.7 (d, $^1J_{C-F}$ = 118.0 Hz, C2'), 78.8 (C4'), 69.8 (d, $^2J_{C-F}$ = 15.6 Hz, C3'), 62.3 (C-5'), 20.9 (C$_b$/C$_d$)*, 20.5 (C$_b$/C$_d$)*.

^{19}F-NMR (376 MHz, CDCl$_3$): δ/ppm = -199.1 (dt, $^2J_{H-F}$ = 52.3 Hz, $^3J_{H-F}$ = 18.7 Hz, C2'-*F*).

^{15}N-NMR (41 MHz, CDCl$_3$): δ (ppm) = -224.6.

HRMS (ESI): calc. for C$_{13}$H$_{14}$FN^{15}NO$_7^-$ [M-H]$^-$: 330.07609, found: 330.07654.

The data are consistent with literature. [3]

3',5'-bis-*O*-(tert-Butyl(dimethyl)silyl)-2'-deoxy-2'-(*R*)-fluoro-(*N*³-¹⁵N)-uridine (12)

12
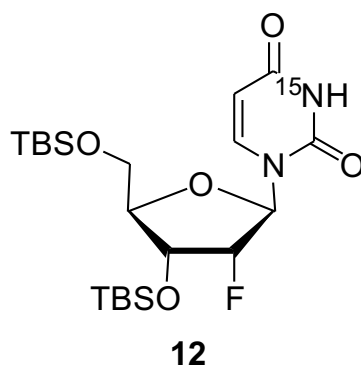
Compound **11** (0.875 g, 2.64 mmol, 1.0 equiv.) was dissolved in ammonia (7 N in methanol, 5.5 mL) and stirred for 18 h at room temperature. Solvents were removed *in vacuo* and the residue dried under high vacuum after which the compound was processed without further purfiication. The nucleoside was dissolved in pyridine (6 mL) and imidazole (0.809 g, 11.9 mmol, 4.5 equiv.) and TBSCl (1.19 g, 7.92 mmol, 3.0 equiv.) were added. The mixture was stirred for 17 h at room temperature. Pyridine was removed under reduced pressure; the resulting residue was taken up in an aqueous $NaHCO_3$ solution (100 mL) and extracted with CH_2Cl_2 (3 x 150 mL). The combined organic layers were washed with an aqueous, saturated NaCl solution (100 mL) and dried over Na_2SO_4. Solvents were removed *in vacuo* and the resulting yellow oil was coevaporated with toluene (3 x 50 mL). The TBS-protected nucleoside **12** was yielded as a colourless, waxy solid (1.02 g, 2.14 mmol, 81%).
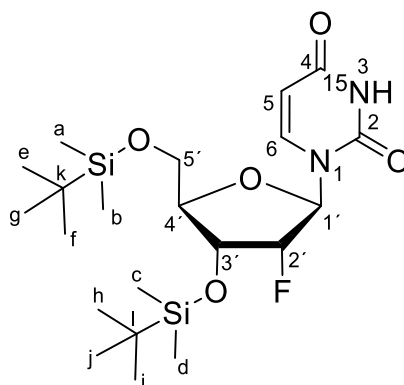
R$_f$ = 0.49 (SiO$_2$, CH$_2$Cl$_2$/MeOH = 99:1).

¹H-NMR (600 MHz, CDCl$_3$): δ/ppm = 8.10 (d, ^3J = 7.6 Hz, 1H, C6-H), 6.04 (d, $^3J_{H-F}$ = 15.5 Hz, 1H, C1'-H), 5.96 (d, ^3J = 7.6 Hz, 1H, C5-H), 4.79 (dd, $^2J_{H-F}$ = 52.2 Hz, ^3J = 4.0 Hz, 1H, C2'-H), 4.23 (ddd, $^3J_{H-F}$ = 21.6 Hz, ^3J = 8.0 Hz, ^3J = 4.0 Hz, 1H, C3'-H), 4.13–4.07 (m, 2H, C4'-H, C5'-H), 3.79 (d, ^2J = 10.8 Hz, 1H, C5'-H), 0.94 (s, 9H, C$_{e-g}$-H)*, 0.89 (s, 9H, C$_{h-j}$-H)*, 0.12 (s, 6H, C$_a$-H, C$_b$-H), 0.12 (s, 3H, C$_c$-H)*, 0.10 (s, 3H, C$_d$-H)*.

¹³C-NMR (201 MHz, CDCl₃): δ/ppm = 162.7 (C4), 149.9 (C2), 140.0 (C6), 102.6 (C5), 93.3 (d, $^1J_{C-F}$ = 196.6 Hz, C2'), 88.0 (d, $^2J_{C-F}$ = 33.8 Hz, C1'), 84.1 (C4'), 68.9 (d, $^2J_{C-F}$ = 16.6 Hz, C3'), 61.0 (C5'), 26.2 (C$_{e-g}$)*, 25.9 (C$_{h-j}$)*,18.7 (C$_l$)**, 18.4 (C$_k$)**, -4.3 (C$_a$)***, -4.7 (C$_b$)***, -5.1 (C$_c$)***, -5.2 (C$_d$)***.

¹⁹F-NMR (377 MHz, CDCl₃): δ/ppm = -202.7 – -202.4 (m, C2'-*F*).

¹⁵N-NMR (41 MHz, CDCl₃): *δ* (ppm) = -224.9.

HRMS (ESI): calc. for $C_{21}H_{40}FN^{15}NO_5Si_2^+$ [M+H]⁺: 476.24246, found: 476.24280.

The data are consistent with literature. [3]

3',5'-bis-*O*-[tert-Butyl(dimethyl)silyl]-2'-deoxy-2'-(*R*)-fluoro-(*N³*,*N⁴*-¹⁵N₂)-cytidine (13)

13
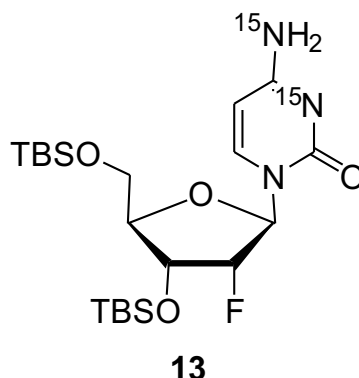
1,2,4-triazole (0.718 g, 10.4 mmol, 9.9 equiv.) was dissolved in MeCN (20 mL), cooled to 0 °C and stirred at 0 °C for 10 min. POCl₃ (0.20 mL, 2.15 mmol, 2.1 equiv.) was slowly added dropwise and the reaction mixture was stirred for additional 10 min at 0 °C. Afterwards, NEt₃ (1.4 mL, 10.0 mmol, 9.5 equiv.) was added dropwise and the mixture was stirred for 20 min at 0 °C. Next, a solution of compound **12** (0.500 g, 1.05 mmol, 1.0 equiv.) in MeCN (7.0 mL) was slowly added dropwise. The reaction mixture was stirred for 16 h at 60 °C. Afterwards, the reaction was quenched by addition of an aqueous, saturated NaHCO₃ solution (100 mL), MeCN was removed under reduced pressure and the solution was extracted with CH₂Cl₂ (3 x 100 mL). The combined organic layers were dried over Na₂SO₄ and solvents were removed *in vacuo*. The triazole derivative, which was obtained as a yellow oil, was dissolved in 1,4-dioxane (2.1 mL) and combined dropwise with a solution of ¹⁵NH₃ (3 N, 2.1 mL). The reaction mixture was stirred for 16 h at 30 °C and subsequently stopped by addition of an aqueous, saturated NH₄Cl solution (35 mL) and extracted with CH₂Cl₂ (3 x 100 mL). The

combined organic layers were washed with an aqueous, saturated NaCl solution (100 mL) and dried over Na_2SO_4. Solvents were removed *in vacuo* and the resulting yellowish oil was purified by flash chromatography (SiO_2, CH_2Cl_2 → CH_2Cl_2/MeOH = 99:1 → CH_2Cl_2/MeOH = 96:4) to yield compound **13** as a colourless foam (0.289 g, 0.608 mmol, 58%).
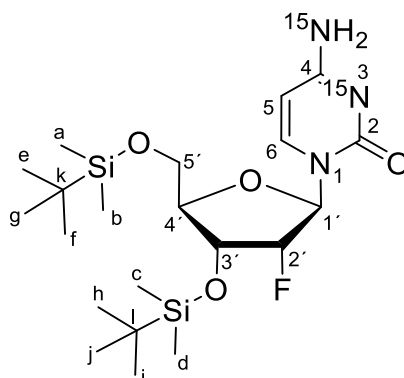
R_f = 0.28 (SiO_2, DCM/MeOH = 99:1).

¹H-NMR (500 MHz, $CDCl_3$): δ/ppm = 7.92 (d, ^3J = 8.2 Hz, 1H, C6-H), 6.06 (dd, $^3J_{H-F}$ = 15.2 Hz, ^3J = 1.9 Hz, 1H, C1'-H), 5.68 (d, ^3J = 8.2 Hz, 1H, C5-H), 4.77 (ddd, $^2J_{H-F}$ = 52.4 Hz, ^3J = 4.2 Hz, ^3J = 1.9 Hz, 1H, C2'-H), 4.29 (ddd, $^3J_{H-F}$ = 18.7 Hz, ^3J = 7.1 Hz, ^3J = 4.2 Hz, 1H, C3'-H), 4.11–4.02 (m, 2H, C4'-H, C5'-H), 3.78 (dd, ^2J = 11.7 Hz, ^3J = 1.6 Hz, 1H, C5'-H), 0.95–0.84 (m, 18H, C_{e-g}–H, C_{h-j}-H), 0.16–0.07 (m, 12H, C_{a-d}-H).

¹³C-NMR (126 MHz, $CDCl_3$): δ/ppm = 162.7 (d, $^1J_{C-N}$ = 9.1 Hz, C4), 143.8 (d, $^1J_{C-N}$ = 18.1 Hz, C2), 141.9 (C6), 101.7 (C5), 93.4 (d, $^1J_{C-F}$ = 193.5 Hz, C2'), 89.0 (d, $^2J_{C-F}$ = 33.8 Hz, C1'), 83.7 (C4'), 68.4 (d, $^2J_{C-F}$ = 16.2 Hz, C3'), 60.7 (C5'), 26.3 (C_{e-g})*, 25.9 (C_{h-j})*,18.8 (C_k)**, 18.4 (C_l)**, -4.3 (C_a)***, -4.7 (C C_b)***, -5.0 (C_c)***, -5.1 (C_d)***.

¹⁹F-NMR (126 MHz, $CDCl_3$): δ/ppm = -201.7 (dt, $^2J_{H-F}$ = 35.8 Hz, $^3J_{H-F}$ = 17.7 Hz, C2'-*F*).

¹⁵N-NMR (41 MHz, $CDCl_3$): δ (ppm) = -247.3, -283.6.

HRMS (ESI): calc. for $C_{21}H_{41}FN^{15}N_2O_4Si_2^+$ [M+H]⁺: 476.25548, found: 476.25568.

3',5'-bis-*O*-(tert-Butyl(dimethyl)silyl)-5-iodo-2'-deoxy-2'-(*R*)-fluoro-(*N*³,*N*⁴-¹⁵N₂)-cytidine (14)

14
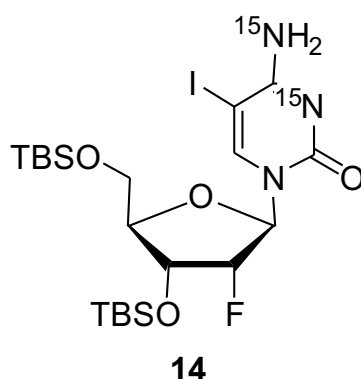
Compound **13** was dissolved in MeCN (35 mL) and CAN (0.540 g, 0.985 mmol, 2.2 equiv.) and iodine (0.250 g, 0.985 mmol, 2.2 equiv.) were added. The reaction mixture was stirred for 3 h at 60 °C and subsequently stopped by addition of an aqueous, saturated NaHCO₃ solution (100 mL) and an aqueous, saturated Na₂S₂O₃ solution (50 mL). MeCN was removed *in vacuo* and the mixture was extracted with CH₂Cl₂ (3 x 150 mL). The combined organic layers were dried over Na₂SO₄ and solvents were removed *in vacuo*. The resulting yellowish oil was purified by flash chromatography (SiO₂, CH₂Cl₂ → CH₂Cl₂/MeOH = 99:1 → CH₂Cl₂/ MeOH = 98:2) and compound **14** was obtained as a yellow solid (0.251 g, 0.417 mmol, 93%).
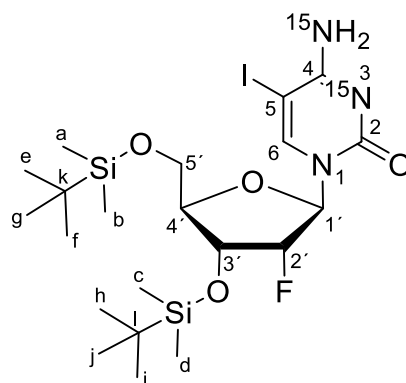
R*f* = 0.50 (SiO₂, CH₂Cl₂/MeOH = 99:1).

¹H-NMR (500 MHz, CDCl₃): δ/ppm = 8.15 (s, 1H, C6-H), 6.03 (dd, ³*J*_{H-F} = 15.0 Hz, ³*J* = 2.5 Hz, 1H, C1'-H), 4.88 (dt, 1H, C2'-H), 4.24 (ddd, ³*J*_{H-F} = 16.4 Hz, ³*J* = 6.6 Hz, ³*J* = 4.5 Hz, 1H, C3'-H), 4.12 (d, ³*J* = 7.0 Hz, 1H, C5'-H), 4.05 (dd, ³*J*_{H-F} = 11.9 Hz, ³*J* = 1.7 Hz, 1H, C4'-H), 3.79 (dd, ²*J* = 12.0 Hz, ³*J* = 2.2 Hz, 1H, C5'-H), 0.96 (s, 9H, C_{e-g} -H), 0.91 (s, 9H, C_{h-j} -H), 0.17 (2 x s, 6H, C_a-H, C_b-H)*, 0.13 (s, 3H, C_c-H)*, 0.11 (s, 3H, C_d-H)*.

13C-NMR (126 MHz, CDCl$_3$): δ/ppm = 165.0 (C4), 150.8 (C2), 147.1 (C6), 92.0 (d, $^1J_{C\text{-}F}$ = 200.8 Hz, C2'), 88.3 (d, $^2J_{C\text{-}F}$ = 34.3 Hz, C1'), 84.2 (C4'), 68.2 (d, $^2J_{C\text{-}F}$ = 15.1 Hz, C3'), 60.5 (C5'), 55.7 (C5), 25.7 (C$_{e\text{-}g}$)*, 25.0 (C$_{h\text{-}j}$)*, 18.0 (C$_k$)**, 17.4 (C$_l$)**, -5.26 (C$_a$)***, -5.38 (C$_b$)***, -5.51 (C$_c$)***, -5.75 (C$_d$)***.

19F-NMR (126 MHz, CDCl$_3$): δ/ppm = -203.1 (d, $^2J_{H\text{-}F}$ = 47.9 Hz, C2'-*F*).

15N-NMR (41 MHz, CDCl$_3$): δ (ppm) = -248.6, -267.1.

HRMS (ESI): calc. for C$_{21}$H$_{40}$FlN^{15}N$_2$O$_4$Si$_2$$^+$ [M+H]$^+$: 602.15213, found: 602.15288.

3',5'-bis-*O*-(tert-Butyl(dimethyl)silyl)-5-(methylcarboxy)-2'-deoxy-2'-(*R*)-fluoro-(*N*^{3},*N*^{4}-^{15}N$_2$)-cytidine (15)

15

The reaction vessel of a high-pressure CO-autoclave was charged with **14** (50.0 mg, 83.1 µmol, 1.0 equiv.) and [Pd(MeCN)$_2$Cl$_2$] (11.0 mg, 4.24 µmol, 5.0 Mol%). The solids were taken up in MeOH (1 mL). The reactor was purged with CO (3.5 bar) to test for leakage. Subsequently, the reactor was purged two more times with CO (3.5 bar). DIPEA (28.0 µL, 0.16 mmol, 2.0 equiv.) was added and the reaction mixture was stirred for 20 h at 65 °C at 3.5 bar CO. The reaction mixture was filtered to remove residual solids and the filtrate was concentrated *in vacuo*. Purification by flash chromatography (SiO$_2$, CH$_2$Cl$_2$ → CH$_2$Cl$_2$/MeOH = 99:1 → CH$_2$Cl$_2$/MeOH = 98:2) yielded carboxylated compound **15** (33.0 mg, 61.8 µmol, 74%) as a colourless foam.

R$_f$ = 0.57 (SiO$_2$, CH$_2$Cl$_2$/MeOH = 10:1).

^{1}H-NMR (599 MHz, CDCl$_3$): δ/ppm = 9.15 (s, 1H, C6-H), 8.71 (d, ^2J = 93.6 Hz 2H, C4-NH$_2$), 5.91 (dd, $^3J_{H-F}$ = 16.0 Hz, 1H, C1'-H), 5.20 (ddd, $^2J_{H-F}$ = 51.9 Hz, ^3J = 4.4 Hz, 1H, C2'-H), 4.45 (ddd, $^3J_{H-F}$ = 20.1 Hz, ^3J = 7.9 Hz, ^3J = 4.2 Hz 1H, C3'-H), 4.16 (dd, ^3J = 20.2 Hz, ^3J = 7.7 Hz,1H, C4'-H), .3.90 (s, 3H, C$_n$-H$_3$), 3.71-3.63 (m 1H, C5'-H$_a$)*, 3.14-3.07 (m, 1H, C5'-H$_b$)*, 0.91 (s, 9H, C$_{e-g}$-H$_3$), 0.15 (s, 3H, C$_a$)**, 0.13 (s, 3H, C$_b$)**.

^{13}C-NMR, HSQC, HMBC (126 MHz, CDCl$_3$): δ/ppm = 163.0, 159.0, 162.4, 150.1, 90.2, 95.6, 92.0, 89.3, 84.7, 68.3, 59.8, 53.6, 64.2, 42.6, 25.7, 18.0, -5.26, -5.38.

^{15}N-NMR (41 MHz, CDCl$_3$): δ (ppm) = -234.9, -274.3.

HRMS (ESI): calc. for C$_{17}$H$_{28}$FN^{15}N$_2$O$_6$Si [M+Na]$^+$ 442.1564, found: 442.1561.

IR (ATR): v/cm^{-1} = 3359 (w), 3255 (w), 3104 (w), 2953 (w), 2930 (w), 2892 (w), 2858 (w), 1760 (m), 1713 (s), 1694 (s), 1642 (w), 1614 (m), 1453 (s), 1406 (w), 1391 (w), 1328 (w), 1298 (w), 1254 (s), 1230 (w), 1193 (w), 1162 (w), 1122 (m), 1084 (s), 1069 (s), 1005 (w), 983 (w), 938 (w), 910 (w), 863 (s), 840 (vs), 815 (s), 780 (vs), 758 (s), 672 (m).

5-(Carboxy)-2'-deoxy-2'-(R)-fluoro-(N^3,N^4-$^{15}N_2$)-cytidine (3)

3

Compound **15** (30.0 mg, 71.5 μmol, 1.0 equiv.) was dissolved in H_2O/MeCN = 3:5 (3 mL/5 mL), LiOH (21.0 mg, 0.86 mmol, 12.0 equiv.) was added and the reaction mixture was stirred for 20 h at room temperature. HCl (2 M, 10 mL) was then added until the solution showed pH = 3 and the mixture was concentrated *in vacuo*. The obtained nucleoside was used without further processing for the final deprotection.

For this, the nucleoside (71.5 μmol, 1.0 equiv.) was dissolved in EtOAc (1.0 mL) and HF • Pyridin (70%, 21.0 mg, 28.0 μL, 1.07 mmol, 15.0 equiv.) was added. The reaction mixture was stirred for 20 h at room temperature, then cooled to 0 °C, combined with TMS • OMe (0.23 g, 0.30 mL, 2.29 mmol, 32.0 equiv.) and stirred for 1 h at 0 °C. The obtained nucleoside was purified by semi-preparative HPLC (MeCN/H_2O, MeCN: 0 → 20%, 45 min, retention time: 10 min) yielded nucleoside **3** (2.44 mg, 8.38 μmol, 12%) after lyophilisation as a colourless powder.

Retention time = 10 min (C_{18}-SiO_2, MeCN/H_2O, MeCN: 0 → 20%, 45 min).

^{1}H-NMR (800 MHz, D_2O): δ/ppm = 8.71 (s, 1H, C6-H), 6.10 (d, $^3J_{H-F}$ = 18.5 Hz, 1H, C1'-H), 5.21 (dd, $^2J_{H-F}$ = 52.5 Hz, ^3J = 4.6 Hz, 1H, C2'-H), 4.41 (ddd, $^3J_{H-F}$ = 22.6 Hz, ^3J = 8.8 Hz,

^3J = 4.4 Hz 1H, C3'-H), 4.22 (d, ^3J = 9.0 Hz 1H, C4'-H), 4.09 (dd, ^2J = 13.2 Hz, ^3J = 2.4 Hz, 1H, C5'-H$_a$)*, 3.90 (dd, ^2J = 13.0 Hz, ^3J = 4.1 Hz, 1H, C5'-H$_b$)*.

^{13}C-NMR, HSQC, HMBC (101 MHz, D$_2$O): δ/ppm = 168.8 (C$_m$), 162.2 (C-4), 152.6 (C-2), 147.6 (C-6), 93.3, (C2'-H), 89.7, (C1'-H), 82.3 (C4'-H), 67.3, (C3'-H), 59.5 (C5'-H).

^{19}F-NMR (377 MHz, D$_2$O): δ/ppm = −201.1 - −201.5 (m, C2'-*F*).

HRMS (ESI): calc. C$_{10}$H$_{13}$FN^{15}N$_2$O$_6$$^+$ [M+H]$^+$ 292.0724, found: 292.0721.
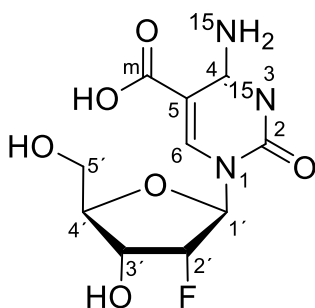
IR (ATR): ν/cm^{-1} = 3334 (m), 3206 (m), 2928 (w), 2892 (w), 2857 (w), 2227 (w), 1647 (vs), 1493 (s), 1408 (m), 1361 (m), 1295 (m), 1252 (m), 1097 (s), 1060 (s), 986 (w), 950 (w), 917 (w), 833 (m), 813 (w), 781 (m), 703 (w), 672 (w).

In vitro decarboxylation experiment
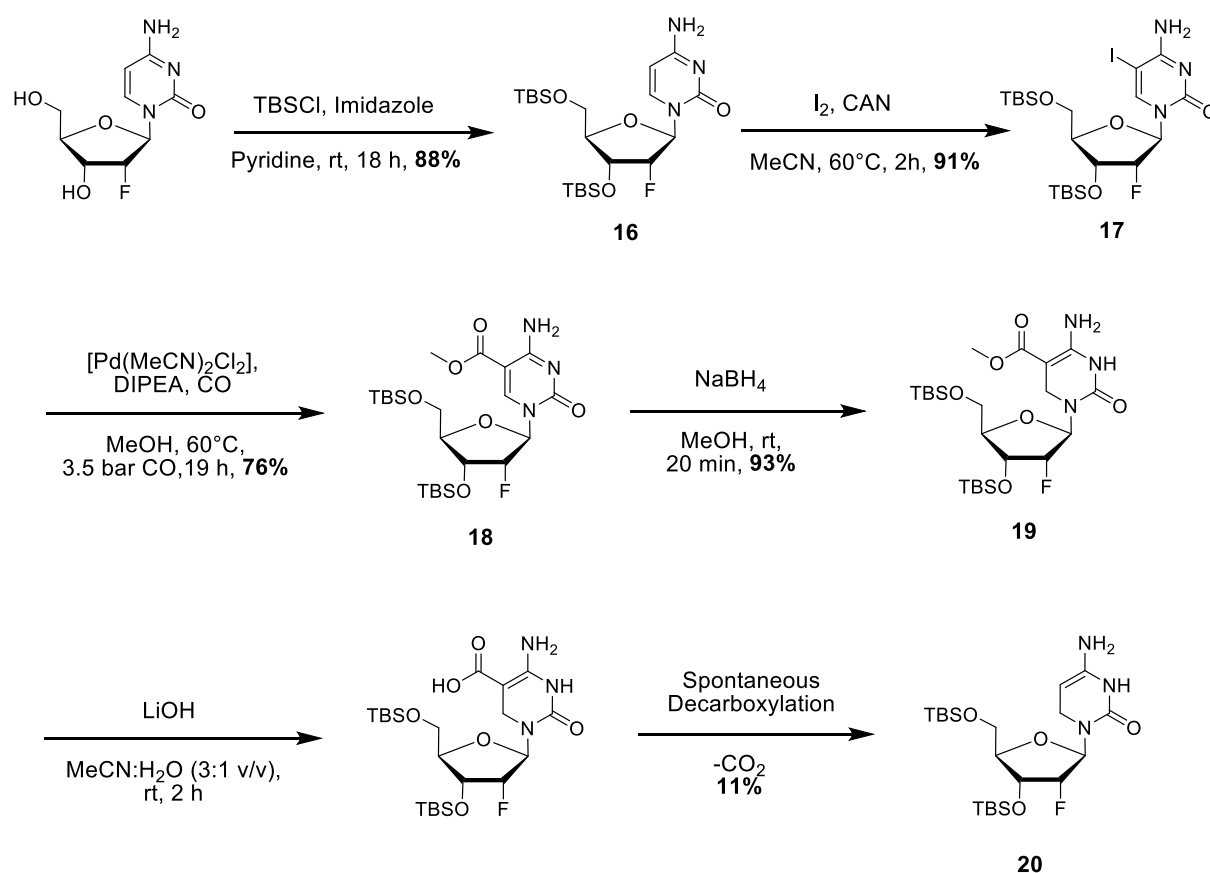
Scheme SI-2: Synthesis pathway towards 2'-F-5,6-dihydro-dC (**20**).

_O_-3',5'-bis(_tert_-Butyldimethylsilyl)-2'-deoxy-2'-(R)-fluoro-cytidine (16)
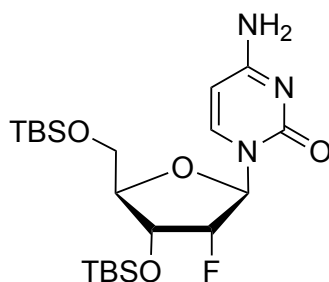
16

2'-deoxy-2'-fluoro-cytidine (1 g, 4.1 mmol, 1 equiv.) was dissolved in pyridine (0.5 M) and imidazole (1.25 g, 18.4 mmol, 4.5 equiv.) and TBSCl (1.85 g, 12.3 mmol, 3.0 equiv.) were added. The mixture was stirred for 17 h at room temperature. Pyridine was removed under reduced pressure; the resulting residue was taken up in an aqueous $NaHCO_3$ solution (100 mL) and extracted with CH_2Cl_2 (3 x 150 mL). The combined organic layers were washed with an aqueous, saturated NaCl solution (100 mL) and dried over Na_2SO_4. Solvents were removed _in vacuo_ and the resulting yellow oil was coevaporated with toluene (3 x 50 mL). The TBS-protected nucleoside **16** was yielded as a colourless, waxy solid (1.7 g, 3.6 mmol, 88%).
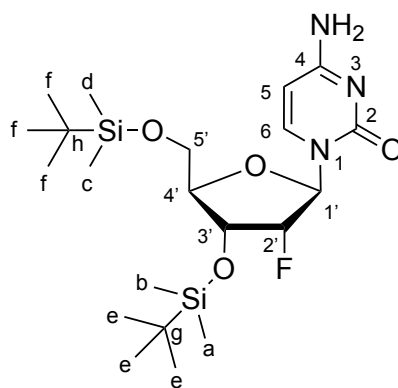
TLC: R_f = 0.60 (DCM/MeOH 10:1)

^{1}H-NMR (400 MHz, $CDCl_3$, ppm): δ = 9.20 (s, 1H, NH_2), 7.97 (s, 1H, 6-H), 5.96 (dd, $^3J_{H-F}$ = 16,5 Hz, ^3J = 1.8 Hz, 1H, 1'-H), 5.69 (s, 1H, NH_2), 4.79 (ddd, $^2J_{H-F}$ = 52,4 Hz, ^3J = 4.3 Hz, ^3J = 1.8 Hz, 1H, 2'-H), 4.17 (ddd, $^3J_{H-F}$ = 18.8 Hz, ^3J = 7.3 Hz, ^3J = 4.2 Hz, 1H, 3'-H), 4.05 - 4.02 (m, 2H, 4'-H, 5'-H), 3.76 (dd, ^2J = 11.9 Hz, ^3J = 2.4 Hz, 1H, 5'-H)*, 0.93 (s, 9H, Ce-H)*, 0.86 (s, 9H, Cf-H)**, 0.14 (s, 3H, Ca-H)***, 0.13 (s, 3H, Cb-H)***, 0.08 (s, 3H, Cc-H)***, 0.06 (s, 3H, Cd-H)***.

¹³C-NMR (75 MHz, CDCl₃, ppm): δ = 164.1, 154.6, 146.0, 93.5 (d, J = 192.8 Hz), 89.0 (d, J = 34.1 Hz), 83.7, 68.6 (d, J = 16.7 Hz), 60.9, 57.2, 26.4, 25.6, 18.7, 18.1, −4.6, −4.9, −5.0, −5.1.

¹⁹F-NMR (377 MHz, CDCl₃, ppm): -200.9 (ddd, $^2J_{H-F}$ = 52,5 Hz, $^3J_{H-F}$ = 24.1 Hz, $^3J_{H-F}$ = 16,3 Hz)

HRMS (ESI) calc. $C_{21}H_{41}FN_3O_4Si_2^+$ [M+H]⁺, 474.26141, found: 474.26160.
calc. $C_{21}H_{39}FN_3O_4Si_2^-$ [M-H]⁻, 472.24686, found: 472.24715

O-3',5'-bis(*tert*-Butyldimethylsilyl)-5-iodo-2'-deoxy-2'-(R)-fluoro-cytidine (17)

17
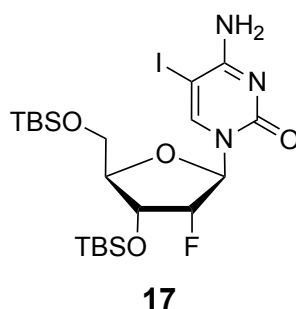
Compound **16** (1.1 g, 2.3 mmol, 1 equiv.) was dissolved in MeCN (100 mL) and CAN (2.8 g, 5.1 mmol, 2.2 equiv.) and iodine (1.3 g, 5.1 mmol, 2.2 equiv.) were added. The reaction mixture was stirred for 3 h at 60 °C and subsequently stopped by addition of an aqueous, saturated NaHCO₃ solution (100 mL) and an aqueous, saturated Na₂S₂O₃ solution (50 mL). MeCN was removed *in vacuo* and the mixture was extracted with CH₂Cl₂ (3 x 150 mL). The combined organic layers were dried over Na₂SO₄ and solvents were removed *in vacuo*. The resulting yellowish oil was purified by flash chromatography (SiO₂, CH₂Cl₂ → CH₂Cl₂/MeOH = 99:1 → CH₂Cl₂/ MeOH = 98:2) and compound **17** was obtained as a pale yellow solid (0.99 g, 2.1 mmol, 91%).
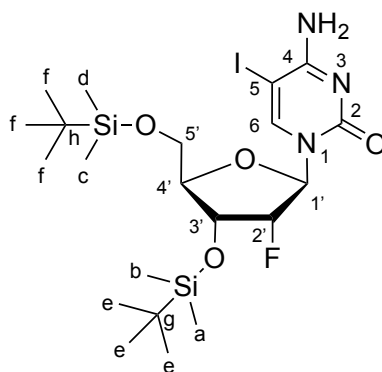
TLC: R_f = 0.57 (DCM/MeOH 10:1)

¹H-NMR (400 MHz, CDCl₃, ppm): δ = 8.36 (s, 1H, NH₂), 7.97 (d, ^3J = 7.4 Hz, 1H, 6-H), 6.19 (s, 1H, NH₂), 6.01 (d, $^3J_{H\text{-}F}$ = 16.5 Hz, 3 J = 1.8 Hz, 1H, C1'-H), 5,78 (d, ^3J = 7.5 Hz, 1H, C5-H), 4,76 (dd, $^2J_{H\text{-}F}$ = 52,2 Hz, ^3J = 3,8 Hz, 1H, C2'-H), 4.21 (ddd, $^3J_{H\text{-}F}$ = 23.3 Hz, ^3J = 8.5 Hz, ^3J = 3.9 Hz, 1H, C3'-H), 4,08 (dd, ^3J = 11.9 Hz, ^3J = 1.8 Hz, 1H, C4'-H), 4,05 (dd, ^2J = 8,8 Hz, ^3J = 2,0 Hz, 1H, C5'-H)*, 3,78 (dd, ^2J = 11.8 Hz, ^3J = 1.6 Hz, 1H, C5'-H)*, 0.92 (s, 9H, Ce-H)**, 0.87 (s, 9H, Cf-H)**, 0.11 (s, 3H, Ca-H)***, 0.10 (s, 3H, Cb-H)***, 0.08 (s, 3H, Cc-H)***, 0.06 (s, 3H, Cd-H)***

¹³C-NMR (75 MHz, CDCl₃, ppm): δ = 166.3, 155,7, 140.4, 94,9, 93.5 (d, J = 191.1 Hz), 88.7 (d, J = 33.4 Hz), 82,6, 67.8 (d, J = 16.7 Hz), 60.2, 57.2, 26.0, 25.6, 18.4, 18.0, −4.6, −5.0, −5.5, −5.5.

¹⁹F-NMR (377 MHz, CDCl₃, ppm): δ = -201.7 (dt, $^2J_{H\text{-}F}$ = 52.2 Hz, $^3J_{H\text{-}F}$ = 17.2 Hz)

HRMS (ESI) calc. $C_{21}H_{40}FIN_3O_4Si_2^+$ [M+H]⁺, 600.15806, found: 600.15848.
calc. $C_{21}H_{38}FIN_3O_4Si_2^-$ [M-H]⁻, 598.14350, found: 598.14399

3'-5'-*O*-bis(*tert*-Butyldimethylsilyl)-5-methoxycarbonyl-2'-desoxy-2'-(R)-fluoro-cytidine (18)
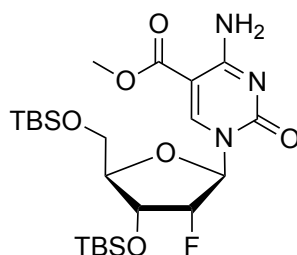
18

2.00 g of compound **17** (3.03 mmol, 1.0 equiv.) and 43 mg [Pd(MeCN)$_2$Cl$_2$] (165 µmol, 0.05 equiv.) were placed in a high-pressure glass autoclave. To test its tightness, the autoclave was filled with CO gas to 3.5 bar and left for 10 min. Subsequently, the CO gas was released, 1.10 mL DIPEA (6.60 mmol, 2.0 equiv.) and 36 mL methanol (0.9 M) were added. The autoclave was flushed twice with CO gas and finally filled with CO gas to 3.5 bar. The suspension was heated to 60 °C for 18 h. After release of the CO gas, removal of the solvent *in vacuo* and purification by column chromatography (dryload, DCM/MeOH 1% → 5%), 1.3 g of compound **22** (2.3 mmol, **76%**) were obtained as a pale yellow solid.
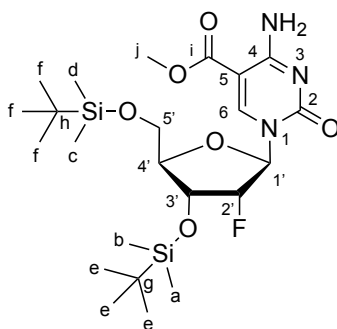
TLC   R$_f$ = 0.33 (5 % MeOH in DCM)

^{1}H NMR   (400 MHz, CDCl$_3$) δ = **8.49** (s, 1H, C6-H) **8.17** (s, 1H, 1·NH$_2$), **7.65** (s, 1H, 1·NH$_2$), **5.92** (dd, 1H, $^3J_{H-F}$ = 17.0 Hz, ^3J = 2.2 Hz, C1'-H), **4.97** (ddd, 1H, $^2J_{H-F}$ = 52.1 Hz, ^3J = 4.1 Hz, ^3J = 2.1 Hz, C2'-H), **4.21** (ddd, 1H, $^3J_{H-F}$ = 18.1 Hz, ^3J = 7.3 Hz, ^3J =3.7 Hz, C3'-H), **4.12-4.08** (m, 1H, C4'-H), **4.01** (dd, 1H, ^2J = 12.1 Hz, ^3J = 2.9 Hz, C5'-H)*, **3.83** (s, 3H, Cj-H), **3.79** (dd, 1H, ^2J = 12.0 Hz, ^3J = 3.8 Hz,

C5'-H)*, **0.90** (s, 9H, Ce-H)** **0.89** (s, 9H, Cf-H)**, **0.1** (s, 3H, Ca-H)***, **0.1** (s, 3H, Cb-H)***, **0.09** (s, 3H, Cc-H)***, **0.08**(s, 3H, Cd-H)***.

¹³C NMR　　　(75 MHz, CDCl₃) δ = 165.1, 163.6, 153.6, 148.6, 95.5, 92.9 (d, $^1J_{C-F}$ = 193.3 Hz), 90.8 (d, J = 34.4 Hz), 84.7, 69.5 (d, J = 16.4 Hz), 61.9, 52.5, 26.3 (3C), 26.0 (3C), 18.8, 18.4, −4.3, −4.5, −5.0, −5.2.

¹⁹F-NMR　　　(377 MHz, CDCl₃): δ/ppm = -201.2 (dt, $^2J_{F2'-H2'}$ = 51.5 Hz, $^3J_{F-H}$ = 18.5 Hz)

HRMS (ESI): calc. $C_{23}H_{43}FN_3O_6Si_2^+$ [M+H]⁺, 532.26689, found: 532.26712.

3'-5'-*O*-bis(*tert*-Butyldimethylsilyl)-5-methoxycarbonyl-2'-desoxy-2'-(R)-fluoro-5,6-dihydrocytidine (19)
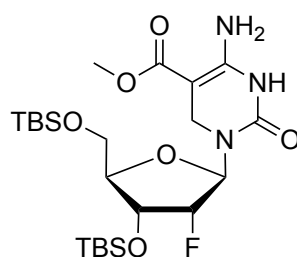
19

To 3´,5´-(tert-Butyl-dimethylsilyl)-5-carboxymethyl-2´-fluoro-2'-deoxy-cytidine (52.0 mg, 97.32 µmol, 1.0 equiv.) dissolved in methanol (5.0 mL) was added sodium borohydride (3.6 mg, 95.16 µmol, 0.98 equiv.). After stirring for 20 min at room temperature a saturated NH₄Cl solution (33 mL) was added and extracted with EtOAc (33 mL). The organic phase was washed with a saturated NH₄Cl solution (2 · 30 mL), dried over MgSO₄ and evaporated to dryness to give the target as an off white solid (50 mg, 93.4 µmol, **96%**).
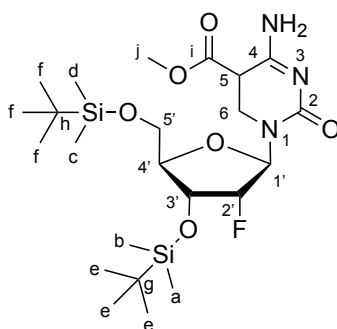
^{1}H NMR (400 MHz, CDCl$_3$) δ = **9.40** (s, 1H, N3-H), **6.44** (s, br, 2H, 2·NH$_2$), **6.06** (dd, 1H, ^{3}J$_{H-F}$ = 18.3 Hz, ^{3}J = 4.0 Hz, C1'-H), **4.85** (dt, 1H, ^{2}J$_{H-F}$ = 53.6 Hz, ^{3}J = 4.6 Hz, C2'-H), **4.29** (dt, 1H, ^{3}J$_{H-F}$ = 12.8 Hz, ^{3}J = 5.2 Hz, C3'-H), **4.09** (d, 1H, J = 12.3 Hz, 1·C6-H), **3.91-3.81** (m, 3H, 1·C6-H, C4'-H, 1·C'5-H), **3.71** (dd, 1H, ^{2}J = 11.6 Hz, ^{3}J = 2.7 Hz, 1·C5'-H), **3.65** (s, 3H, Cg-H), **0.91** (s, 9H, Ce-H)* **0.90** (s, 9H, Cf-H)*, **0.12** (s, 3H, Ca-H)**, **0.11** (s, 3H, Cb-H)**, **0.07** (s, 3H, Cc-H)**, **0.06** (s, 3H, Cd-H)**.

^{13}C NMR (75 MHz, CDCl$_3$) δ = 168.1, 154.9, 90.4 (d, ^{1}J$_{C-F}$ = 191.6 Hz), 86.5 (d, J = 32.9 Hz), 83.3, 70.2 (J = 15.4), 62.1, 39.7, 26.3 (3C), 26.0 (3C), 18.5, 18.7, -4.4, -4.7, -5.1, -5.3.

^{19}F NMR (377 MHz, CDCl$_3$) δ = -206.3 (ddd, ^{2}J$_{F2'-H2'}$ = 52.8 Hz, ^{3}J$_{F-H'}$ = 18.3 Hz, ^{3}J$_{F-H'}$ = 13.4 Hz)

HRMS (ESI) calc. C$_{23}$H$_{45}$FN$_3$O$_6$Si$_2$$^+$ [M+H]$^+$, 534.28254, found: 534.28264.
calc. C$_{23}$H$_{43}$FN$_3$O$_6$Si$_2$$^-$ [M-H]$^-$, 532.26799, found: 532.26838

3'-5'-*O*-bis(*tert*-Butyldimethylsilyl)-2'-desoxy-2'-(R)-fluoro -5,6-dihydrocytidine (20)
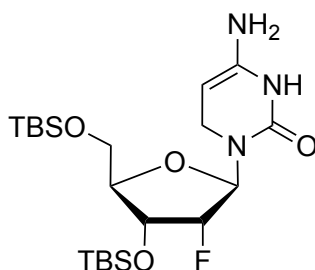
20

To a solution of **18** (200 mg, 0.4 mmol, 1.0 equiv.) in an acetonitrile/water mixture (17 mL/5 mL) LiOH (117 mg, 4.9 mmol, 12.2 equiv.) was added and the resulting solution was stirred at room temperature for 2 h. After addition of a saturated NaHCO$_3$-solution (40 mL) and acidification with 2 M hydrochloric acid (20 mL) the aqueous phase was extracted with diethyl ether (3 · 40 mL), dried over MgSO$_4$ and evaporated to dryness. After column chromatography (DCM/MeOH 50:1 → 5:1) the product was obtained as a colorless, waxy solid (22 mg, 0.04 mmol, **11%**).
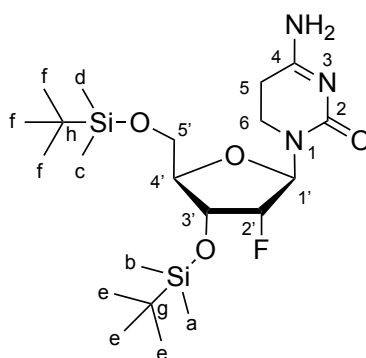
¹H NMR   (400 MHz, CDCl$_3$) δ = **5.99** (dd, 1H, $^3J_{H\,F}$ = 19.0 Hz, ^3J = 3.4 Hz, C1'-H), **4.75** (dt, 1H, $^2J_{H\text{-}F}$ = 53.7 Hz, ^3J = 4.1 Hz, C2'-H), **4.23** (dt, 1H, $^3J_{H\text{-}F}$ = 15.1 Hz, ^3J = 5.5 Hz, C3'-H), **3.89-3.83** (m, 2H, 1·C5-H, C4'-H), **3.70** (dd, 1H, ^2J = 12.2 Hz, ^3J = 3.4 Hz, 1·C5'-H), **3.56** (dt, 1H, ^3J = 12.1 Hz, ^3J = 6.0 Hz, 1·C6-H), **3.28-3.18** (m, 1H, 1·C6-H), **2.66-2.50** (m, 1H, 2·C5-H), **0.90** (s, 9H, Ce-H)* **0.88** (s, 9H, Cf-H)*, **0.10** (s, 3H, Ca-H)**, **0.09** (s, 3H, Cb-H)**, **0.06** (s, 3H, Cc-H)**, **0.05**(s, 3H, Cd-H)**.

¹³C NMR   (75 MHz, $CDCl_3$) δ = 166.6, 157.6, 91.8 (d, $^1J_{C-F}$ = 191.0 Hz), 88.0 (d, J = 33.2 Hz), 82.9, 70.1 (J = 17.1), 62.1, 37.9, 26.2 (3C), 26.0 (3C), 18.5, 18.7 -4.4, -4.7, -5.1, -5.2.

¹⁹F NMR   (377 MHz, $CDCl_3$) δ = -205.7 (dt, $^2J_{F2'-H2'}$ = 52.4 Hz, $^3J_{F-H'}$ = 12.5 Hz)

HRMS (ESI)   calc. $C_{21}H_{43}FN_3O_4Si_2^+$ [M+H]$^+$, 476.27706, found: 476.27754.
                 calc. $C_{21}H_{41}FN_3O_4Si_2^-$ [M-H]$^-$, 474.26251, found: 474.26331.

Supplementary tables

Table SI-1: Compound-dependent LC-MS/MS-parameters used for the analysis of genomic DNA fed with fluorinated nucleosides. *CE: collision energy, CAV: collision cell accelerator voltage, EMV: electron multiplier voltage. The nucleosides were analysed in the positive ([M+H)]⁺ species) as well as the negative ([M-H]⁻ species) ion selected reaction monitoring mode (SRM).*

compound	Precursor ion (m/z)	MS1 Resolution	Product Ion (m/z)	MS2 Resolution	Dwell time [ms]	CE (V)	CAV (V)	Polarity
Time segment 1.5-3.7 min								
F-dC	246.09	Wide	112.06	Wide	70	15	3	Positive
[¹⁵N₂]-F-dC	248.08	Wide	114.04	Wide	70	15	3	Positive
[¹⁵N₂]-cadC	274.08	Wide	158.03	Wide	40	5	5	Positive
cadC	272.09	Wide	156.04	Wide	40	5	5	Positive
F-hmdC	276.10	Wide	142.06	Wide	50	10	3	Positive
[¹⁵N₂]-F-hmdC	278.09	Wide	144.06	Wide	50	10	3	Positive
[¹⁵N₂,D₂]-hmdC	262.12	Wide	146.07	Wide	25	27	1	Positive
hmdC	258.11	Wide	142.06	Wide	25	27	1	Positive
[D₃]-mdC	245.13	Wide	129.09	Wide	50	60	1	Positive
mdC	242.11	Wide	126.07	Wide	50	60	1	Positive
F-cadC	290.08	Wide	156.04	Wide	80	5	5	Positive
[¹⁵N₂]-F-cadC	292.08	Wide	158.04	Wide	80	5	5	Positive
Time segment 3.7-4.7 min								

[D$_3$]-F-mdC	263.12	Wide	129.09	Wide	80	15	3	Positive
F-mdC	260.10	Wide	126.07	Wide	80	15	3	Positive
F-cadC	290.08	Wide	156.04	Wide	80	5	5	Positive
[^{15}N$_2$]-F-cadC	292.08	Wide	158.04	Wide	80	5	5	Positive
Time segment 4.7-10 min								
F-fdC	274.08	Wide	140.05	Wide	90	15	3	Positive
[^{15}N$_2$]-F-fdC	276.08	Wide	142.04	Wide	90	15	3	Positive
F-fdU	273.05	Wide	253.05	Wide	30	3	5	Negative
F-dT	259.07	Wide	239.07	Wide	70	3	5	Negative
dT	243.10	Wide	127.05	Wide	20	5	5	Negative
[^{15}N$_2$]-fdC	258.09	Wide	142.04	Wide	30	5	5	Positive
fdC	256.09	Wide	140.05	Wide	30	5	5	Positive
[^{15}N$_2$]-fdC	142.04	Wide	98.04	Wide	20	13	7	Positive
fdC	140.05	Wide	97.04	Wide	20	13	7	Positive

Table SI-2: Compound-dependent LC-MS/MS ranges of the corresponding linear equations.

compound	n (ULOQ)	n (LLOQ)	A/A* (ULOQ)	A/A* (LLOQ)
F-fdC	25.1 fmol	0.390 fmol	0.2267	0.003345
F-dC	800 fmol	3.13 fmol	4.074	0.01512
F-mdC	49.9 fmol	0.780 fmol	0.04174	0.0005833
F-cadC	12.7	0.2	0.240576	0.0217614
F-cadC	202.5	6.3	3.92816	0.1271398

Appendix

NMR Spectra

3',5'-di-*O*-Acetyl-2'-deoxy-2'-(*R*)-fluorouridine (9)
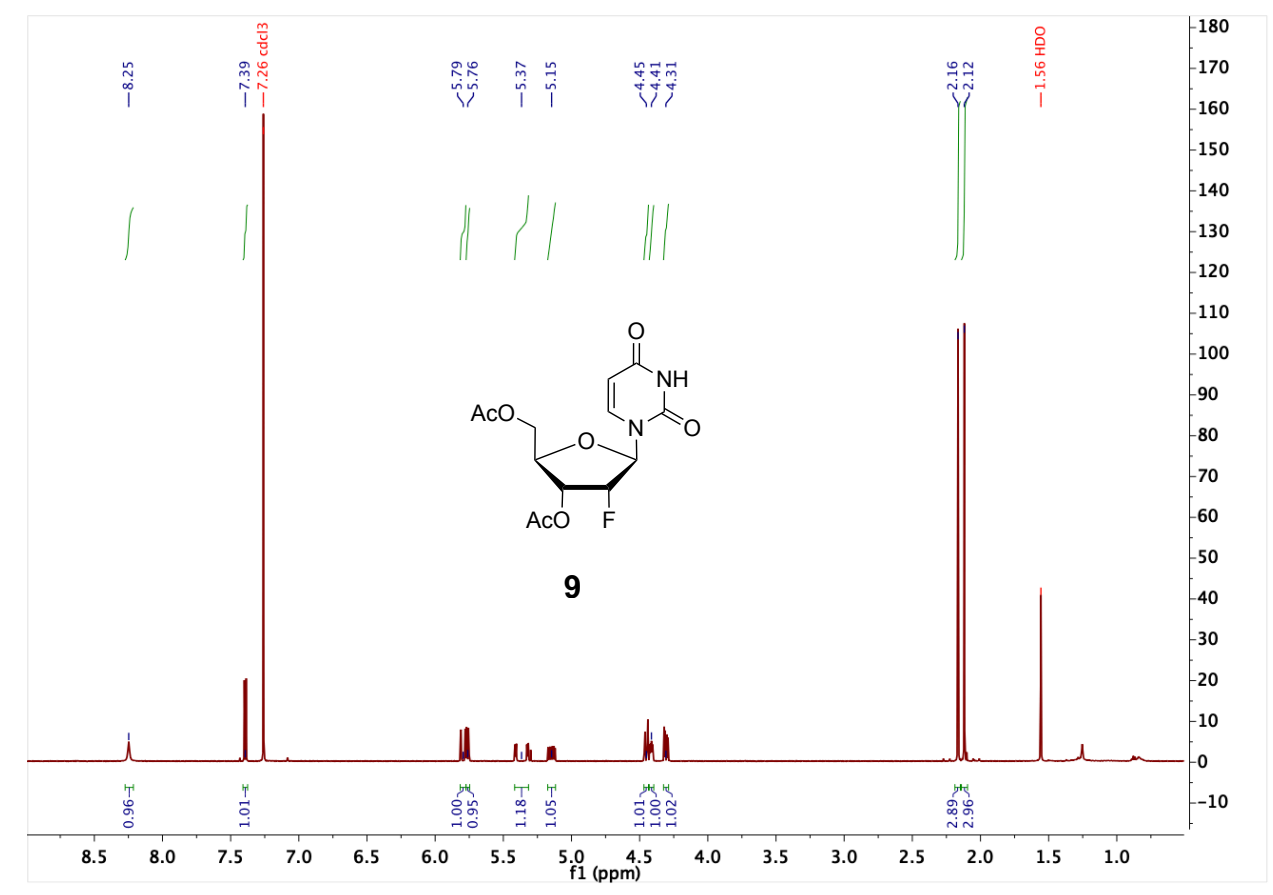
Figure SI-9: ^{1}H-NMR spectrum of compound **9**.
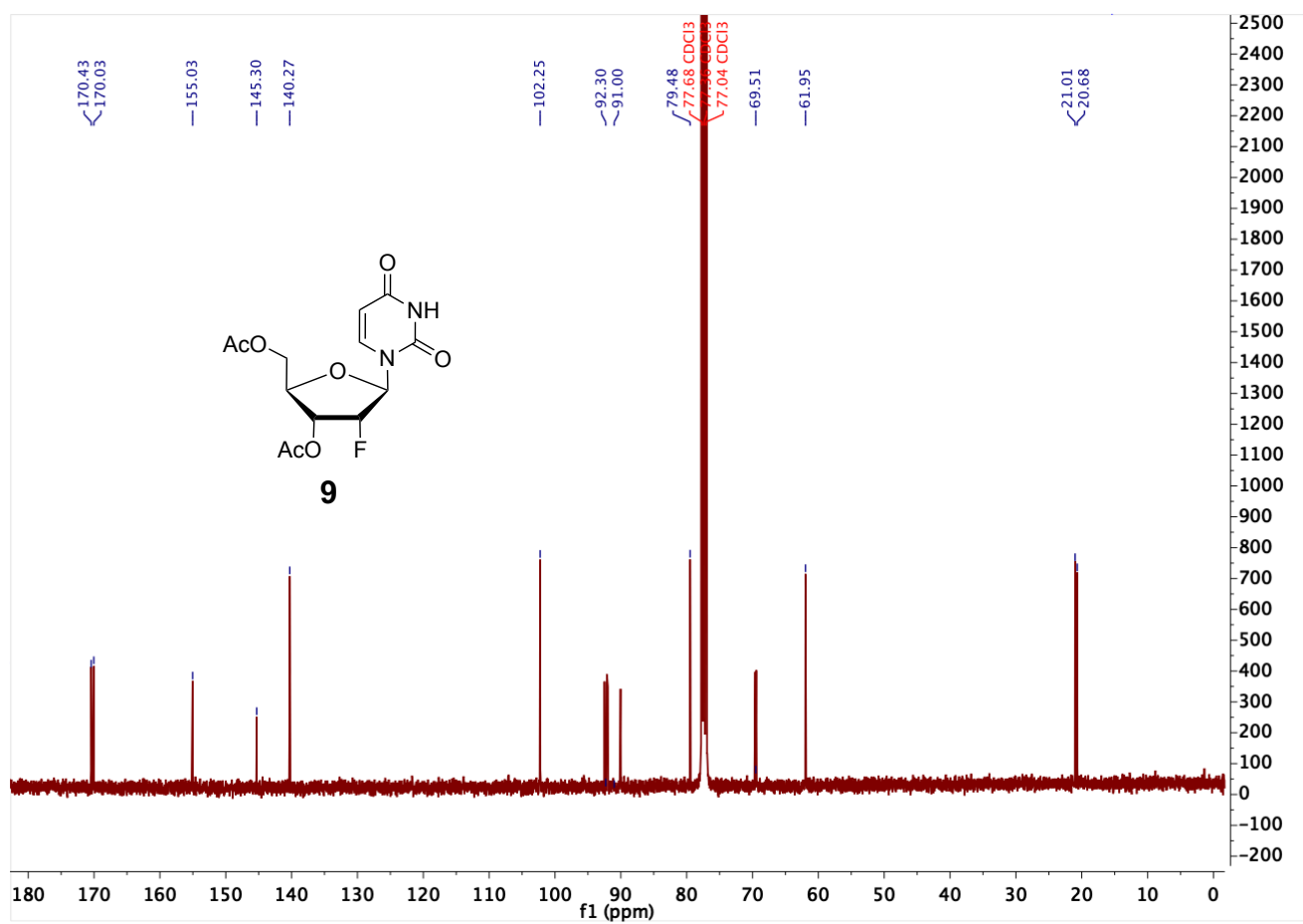
Figure SI-10: ^{13}C-NMR spectrum of compound **9**.
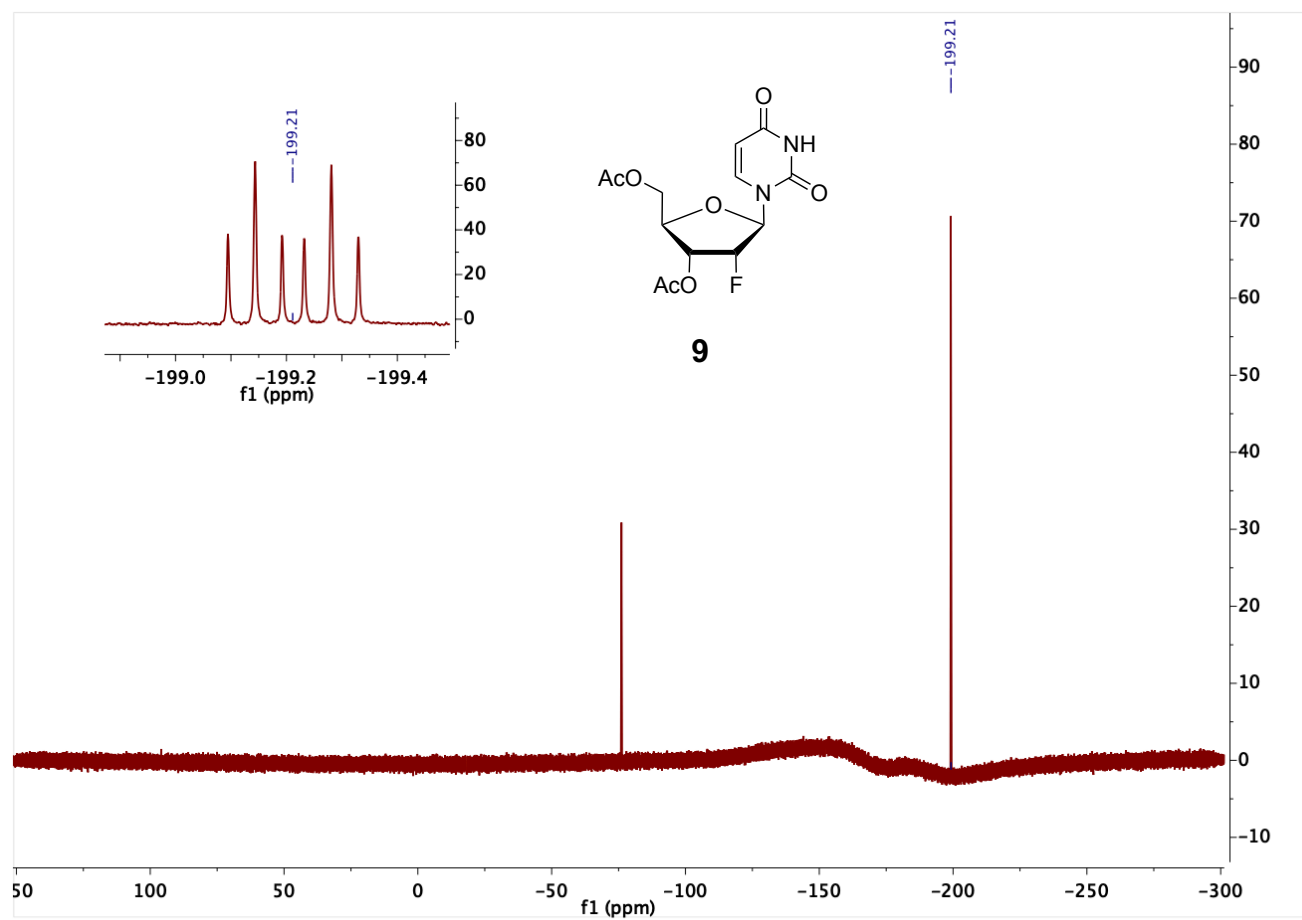
Figure SI-11: ^{19}F-NMR spectrum of compound **9**.

3',5'-di-*O*-Acetyl-2'-deoxy-2'-(*R*)-fluoro-3-nitrouridine (10)
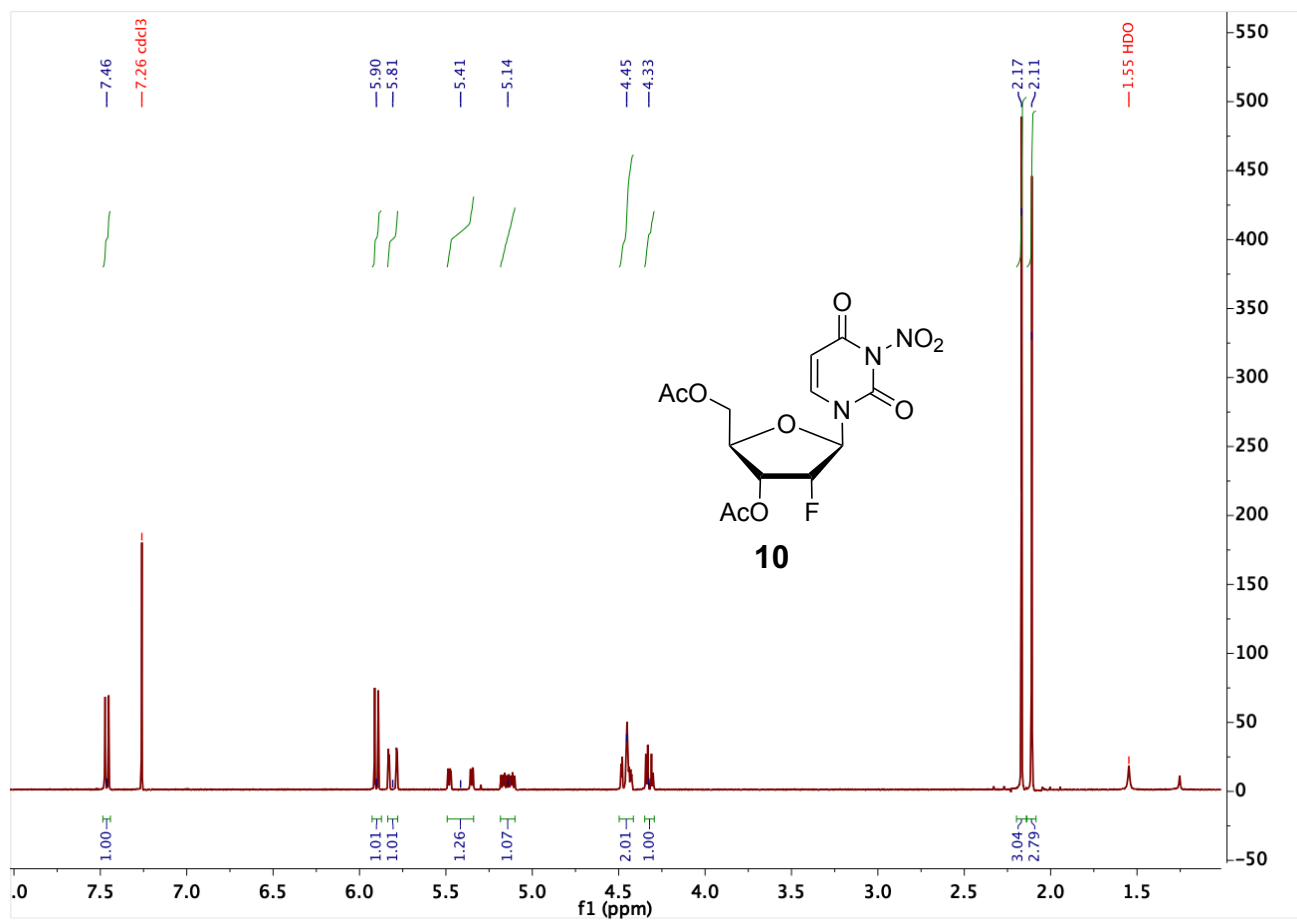
Figure SI-12: ^{1}H-NMR spectrum of compound **10**.
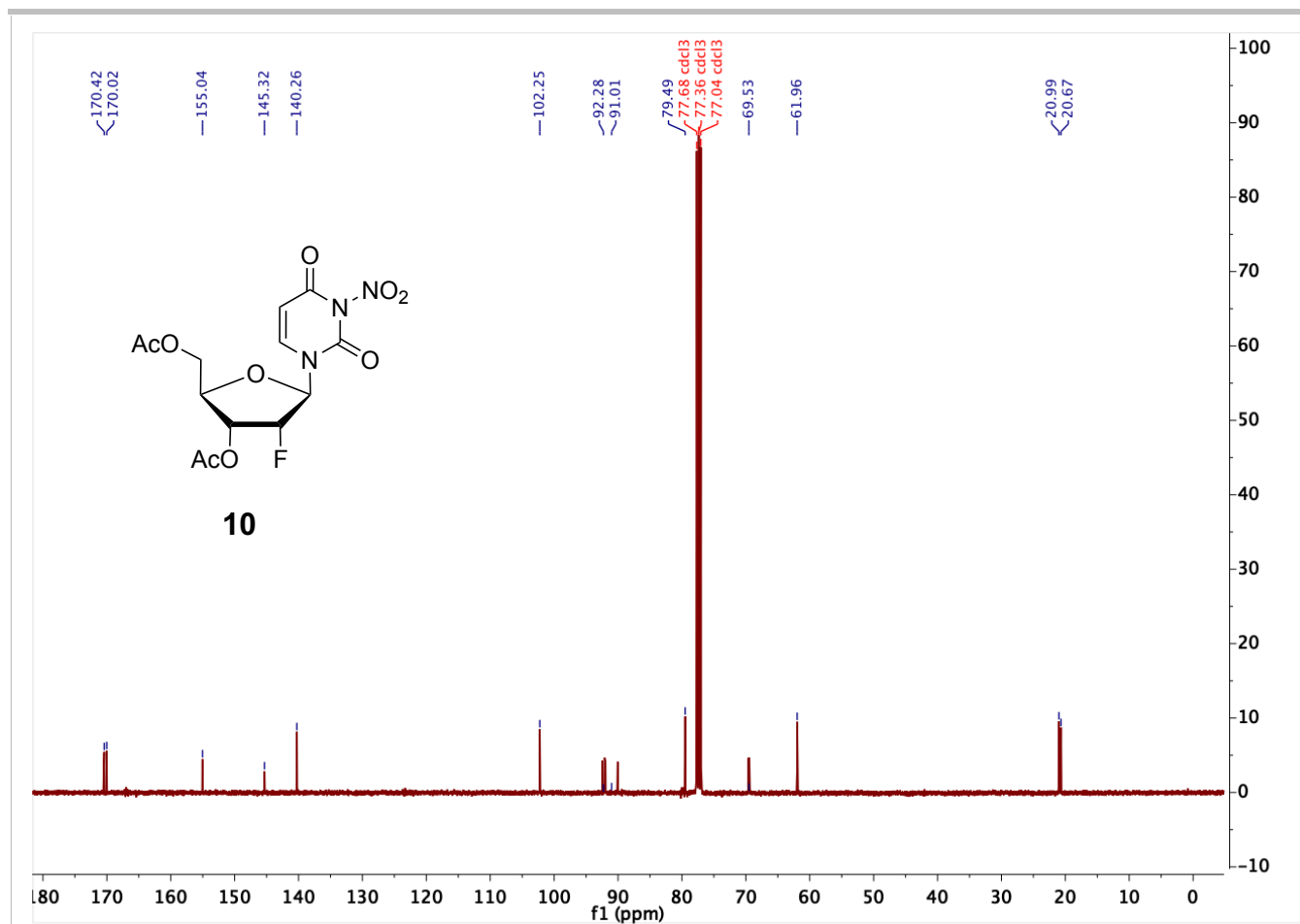
Figure SI-13: ^{13}C-NMR spectrum of compound **10**.
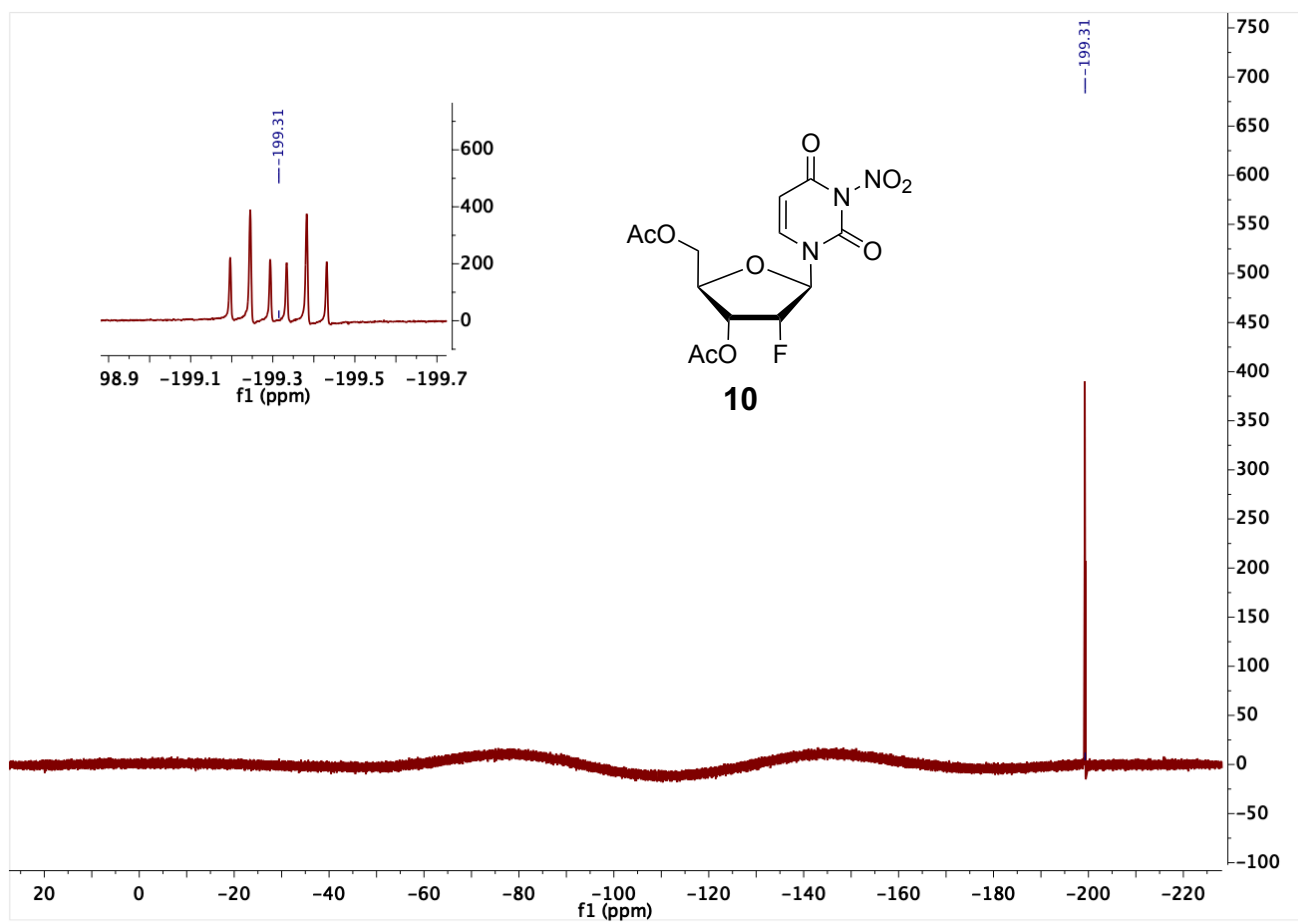
Figure SI-14: ^{19}F-NMR spectrum of compound **10**.

3',5'-di-*O*-Acetyl-2'-deoxy-2'-(*R*)-fluoro-(*N*[3]-^{15}N)-uridine (11)
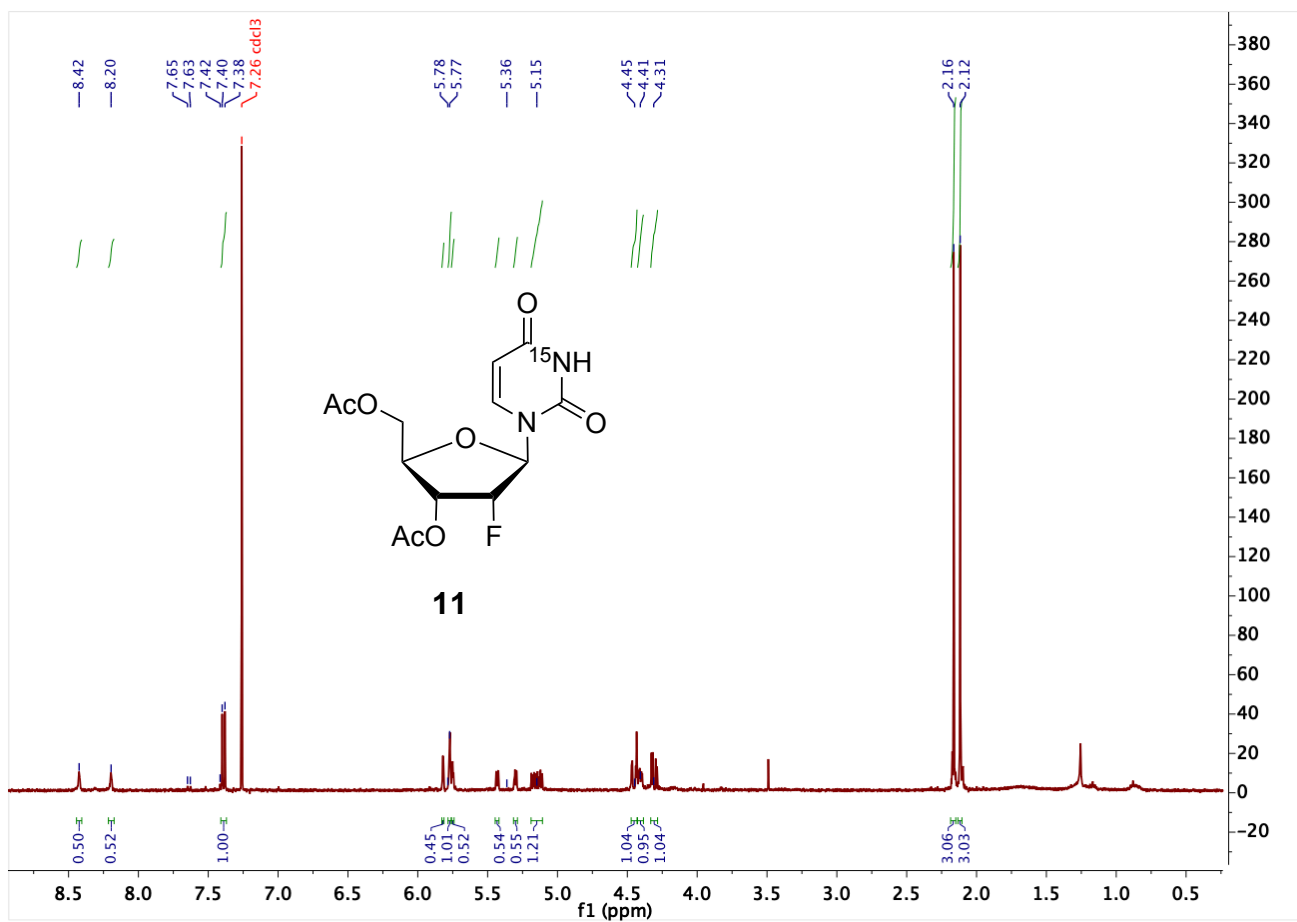
Figure SI-15: ^{1}H-NMR spectrum of compound **11**.
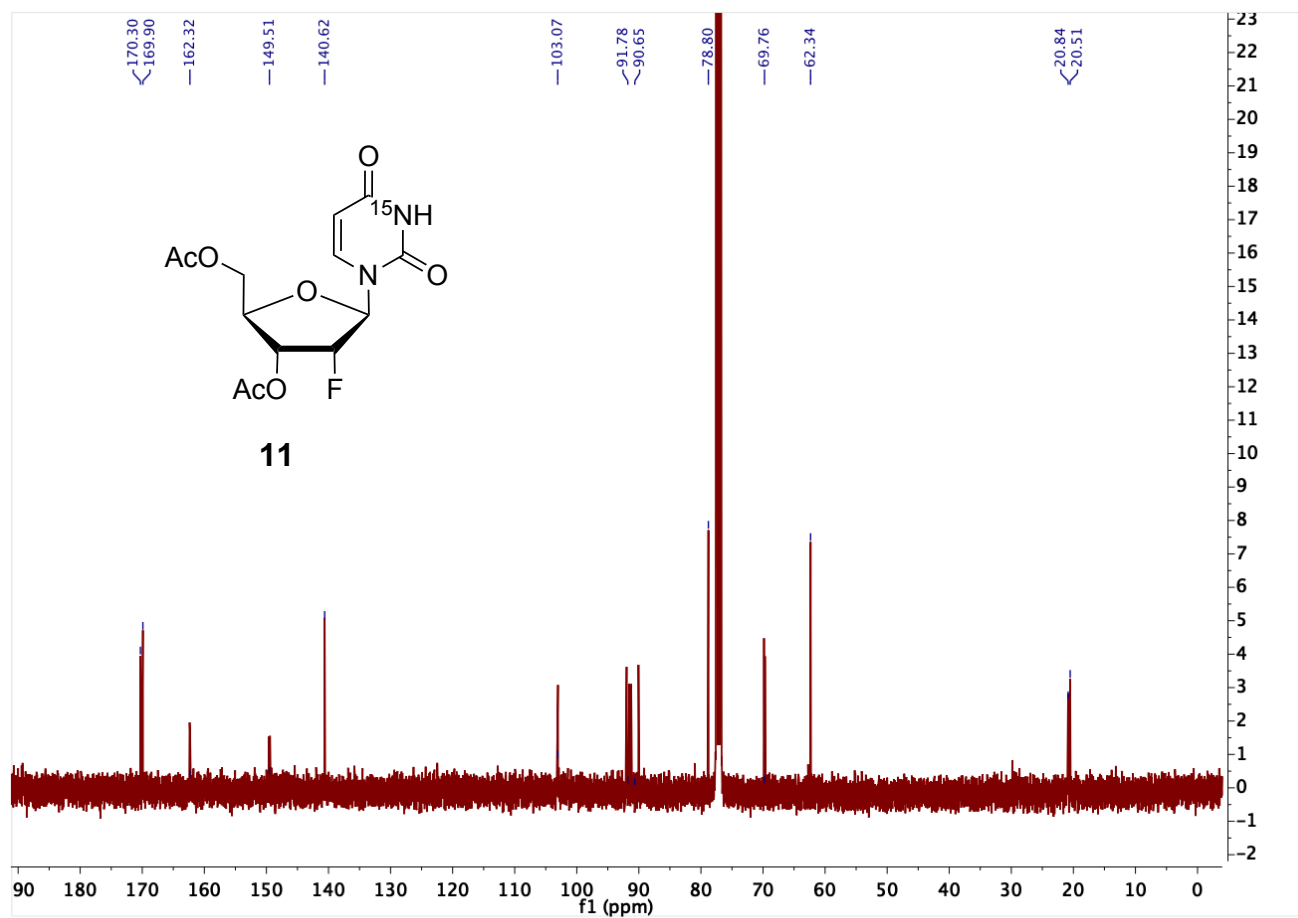
Figure SI-16: ^{13}C-NMR spectrum of compound **11**.
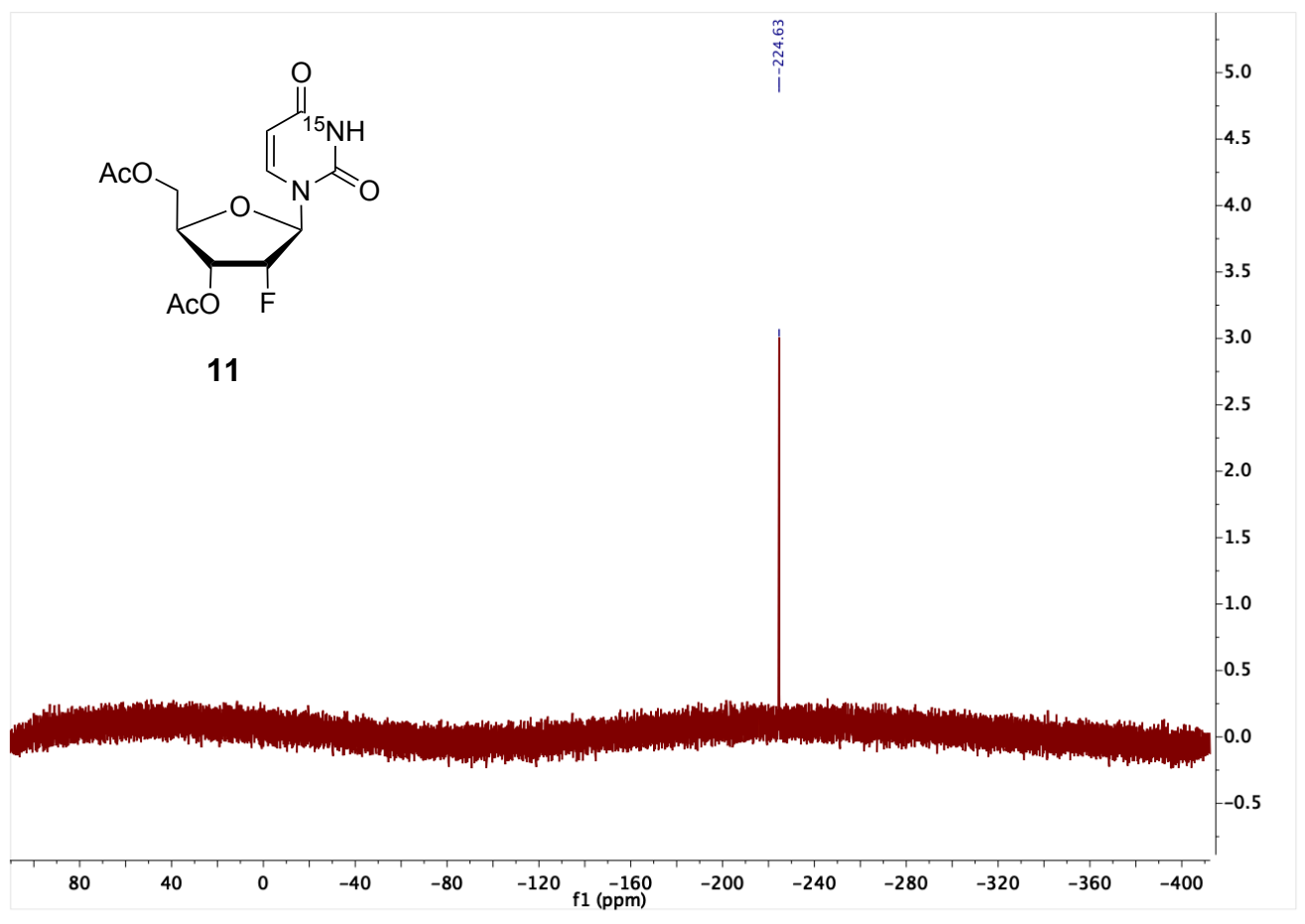
Figure SI-17: ^{15}N-NMR spectrum of compound **11**.
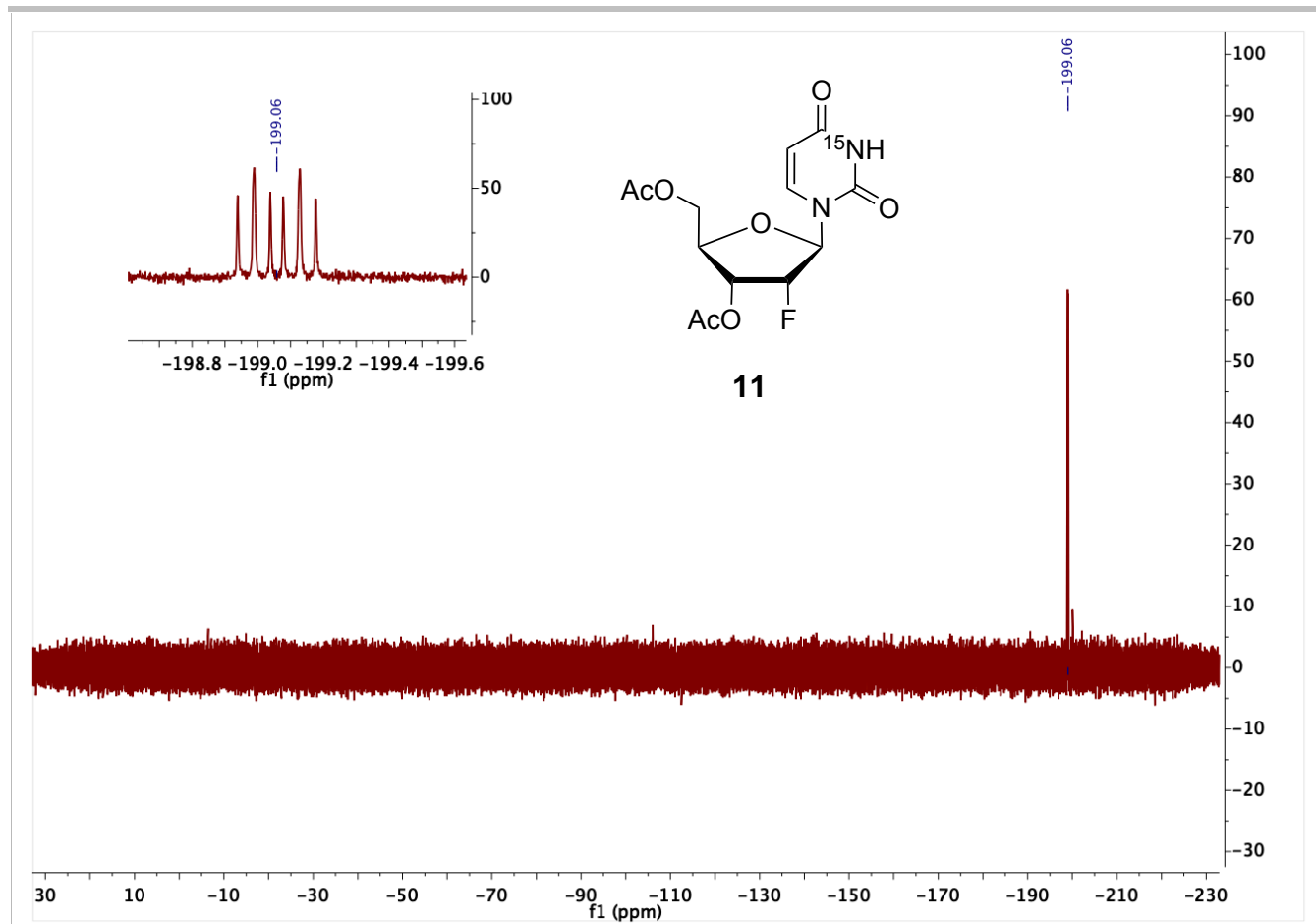
Figure SI-18: ^{19}F-NMR spectrum of compound **11**.

3',5'-bis-*O*-(*tert*-Butyl(dimethyl)silyl)-2'-deoxy-2'-(*R*)-fluoro-(*N*³-¹⁵N)-uridine (12)
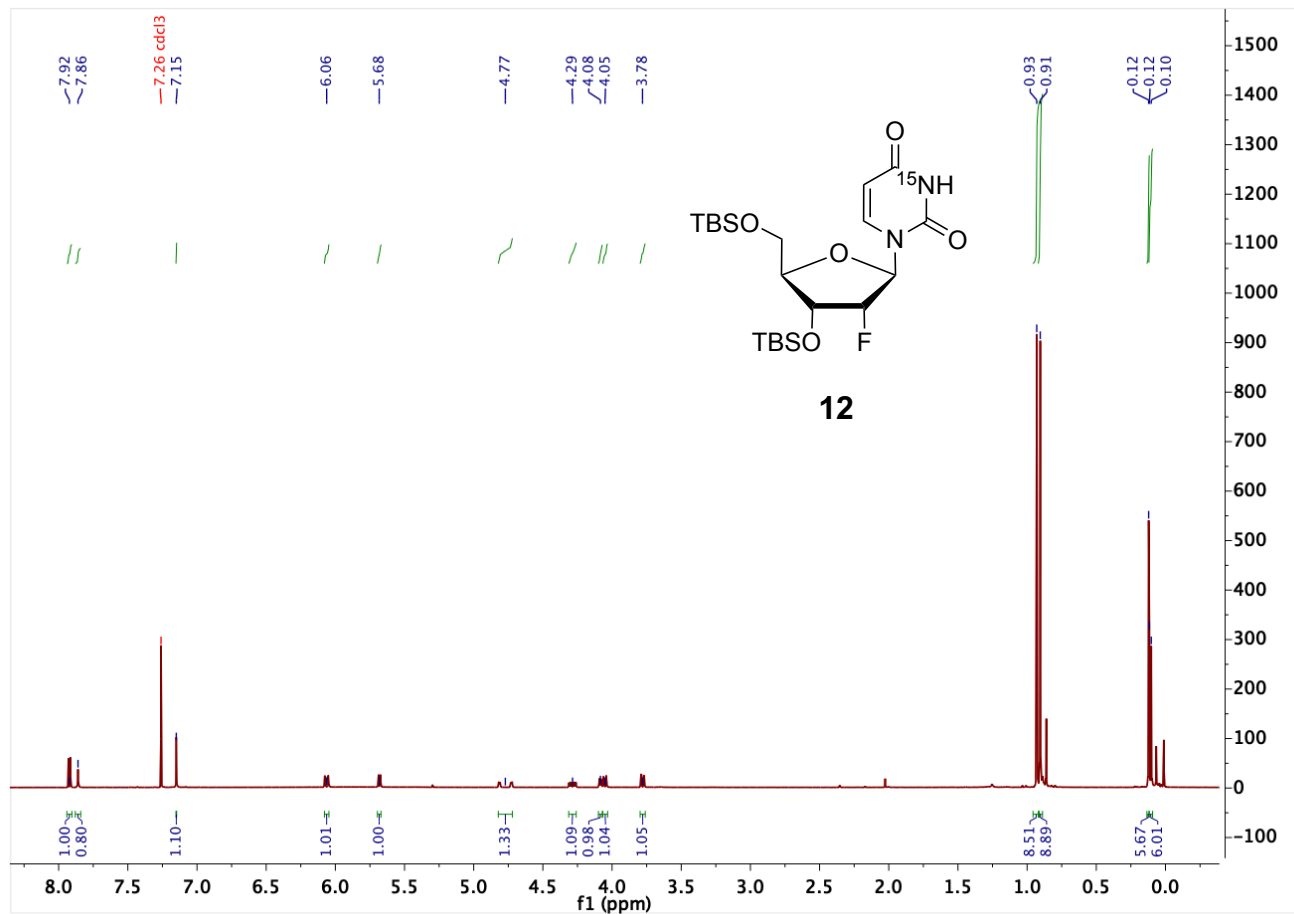
Figure SI-19: ¹H-NMR spectrum of compound **12**.
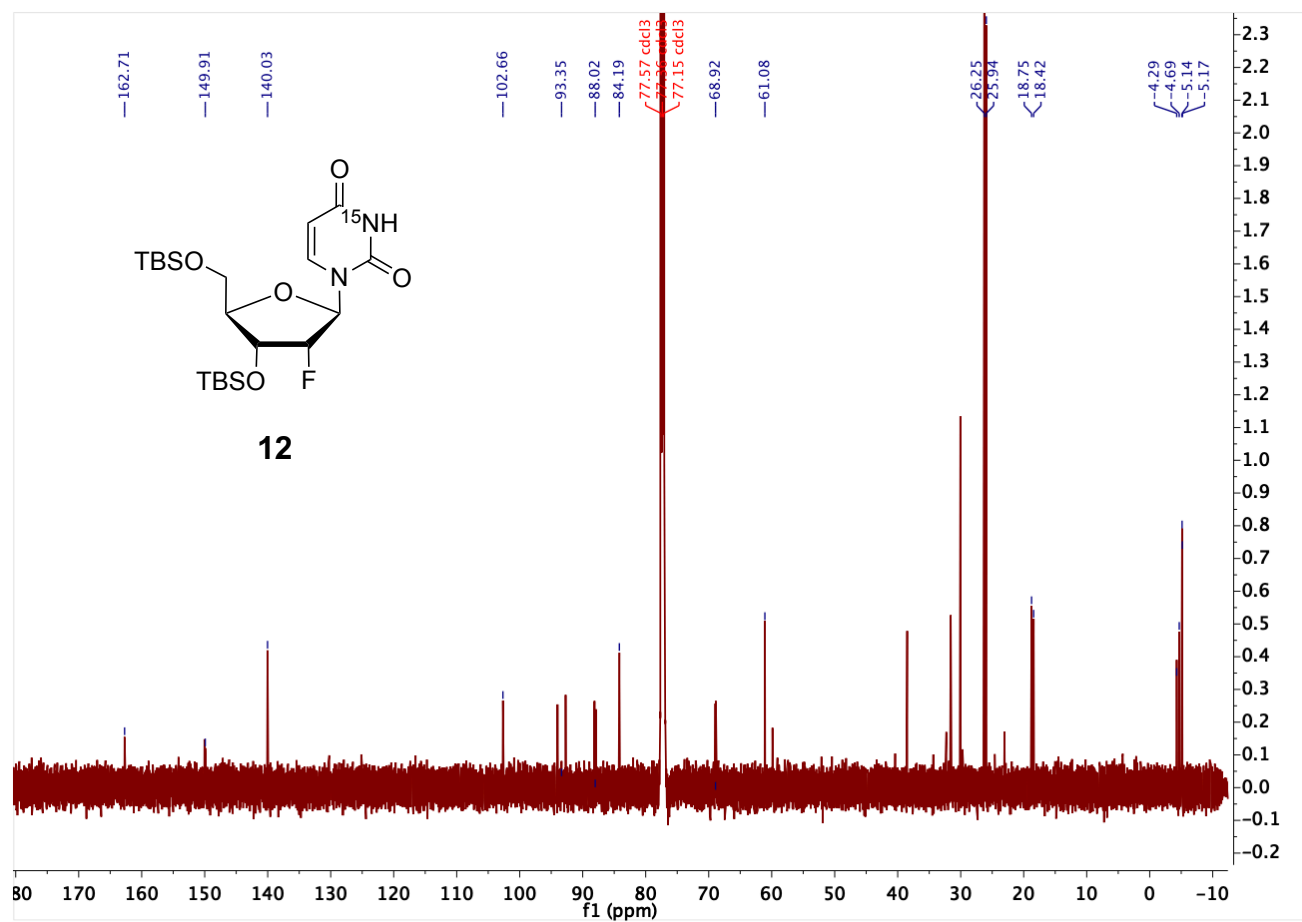
Figure SI-20: ^{13}C-NMR spectrum of compound **12**.
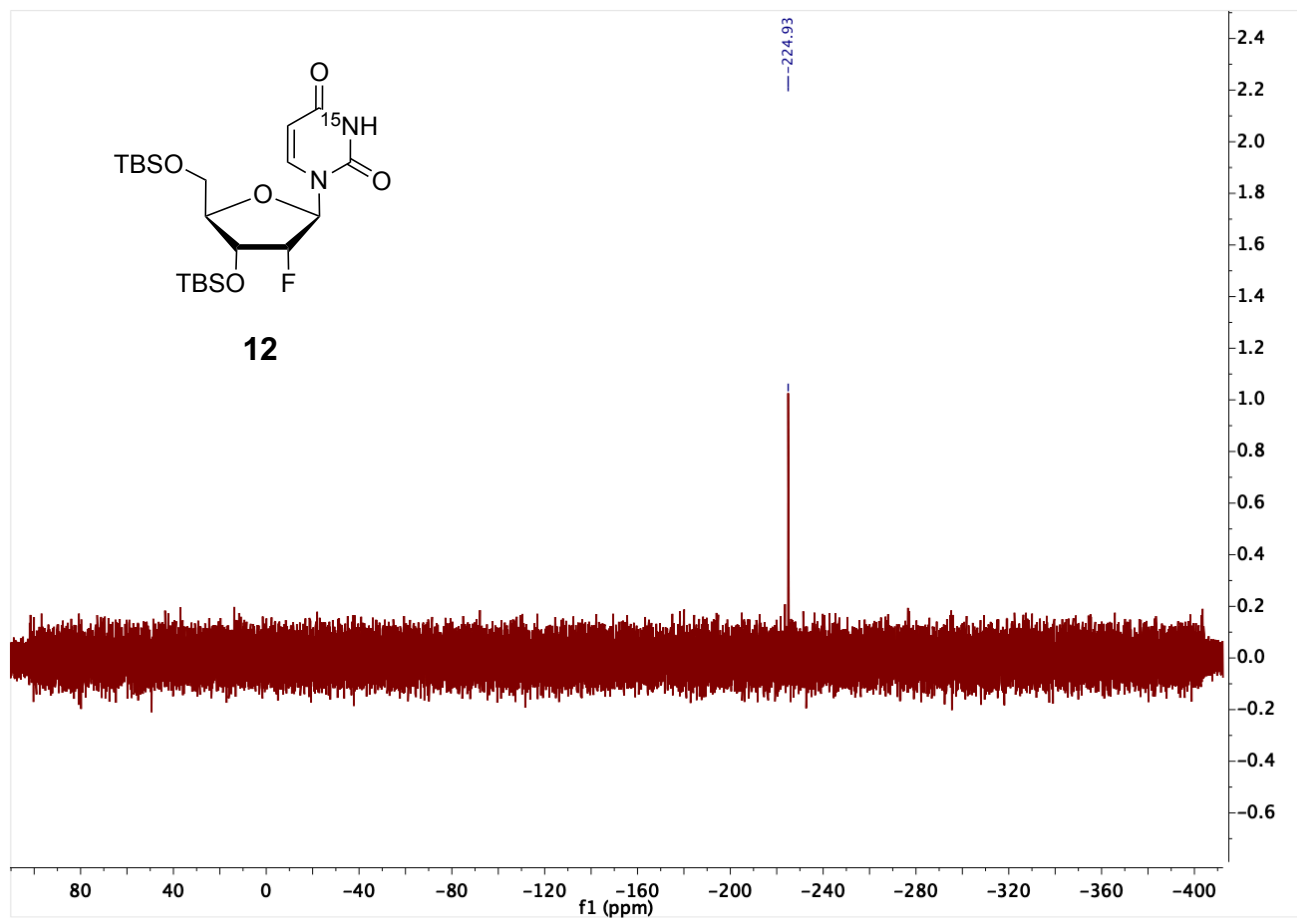
Figure SI-21: ^{15}N-NMR spectrum of compound **12**.
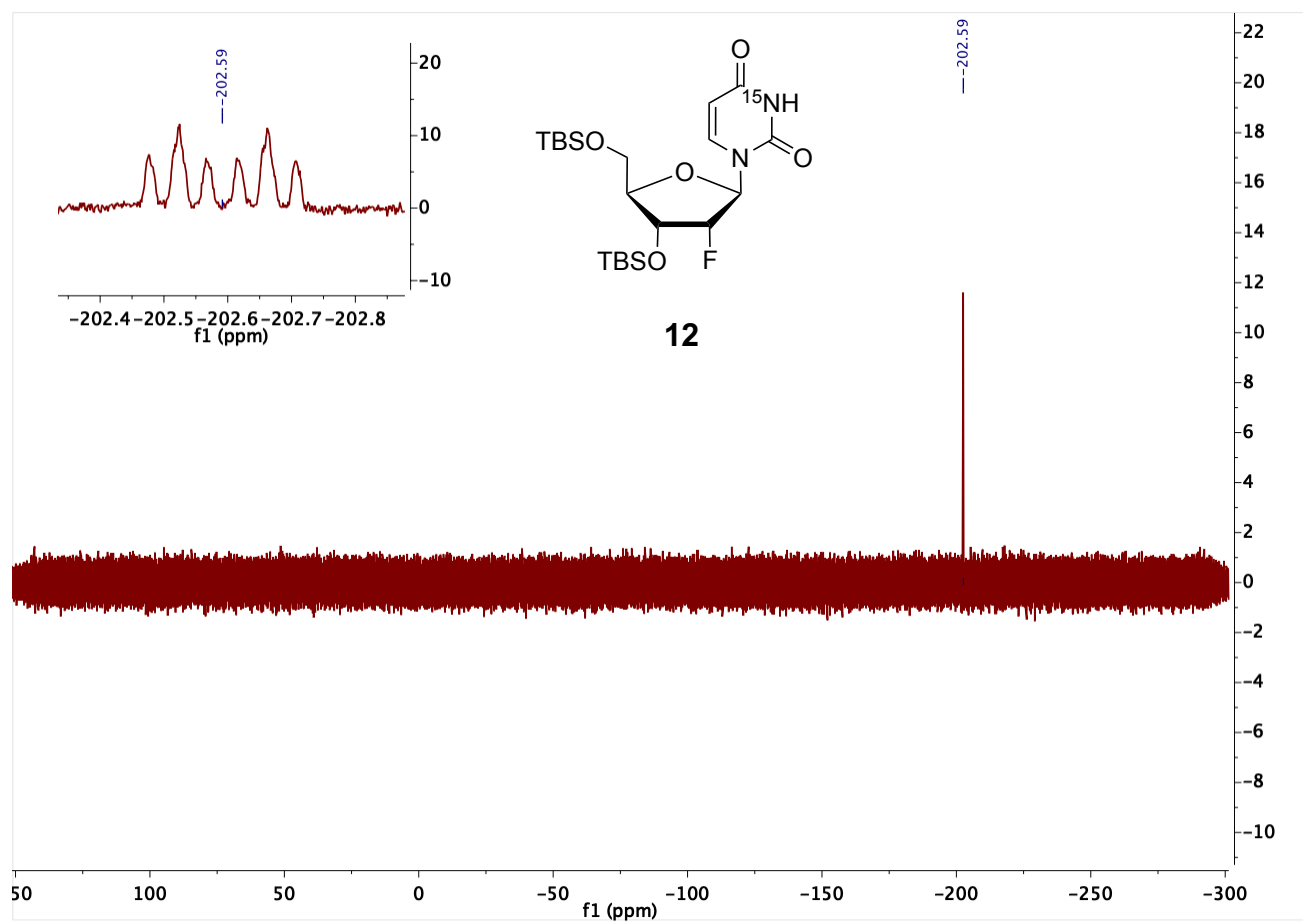
Figure SI-22: ^{19}F-NMR spectrum of compound **12**.

3',5'-bis-*O*-[*tert*-Butyl(dimethyl)silyl]-2'-deoxy-2'-(*R*)-fluoro-(*N*³,*N*⁴-¹⁵N₂)-cytidine (13)
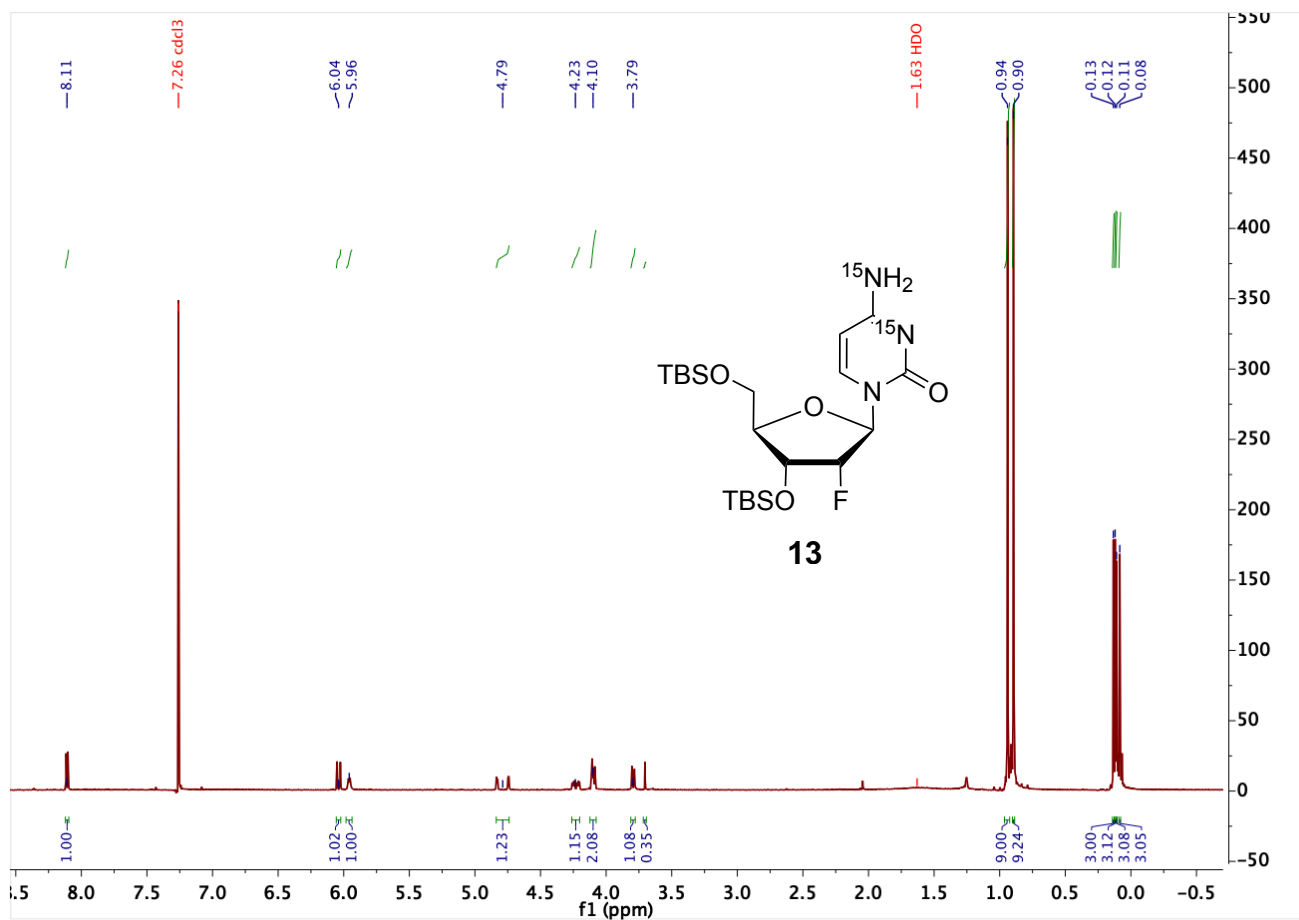
Figure SI-23: ¹H-NMR spectrum of compound **13**.

SUPPORTING INFORMATION
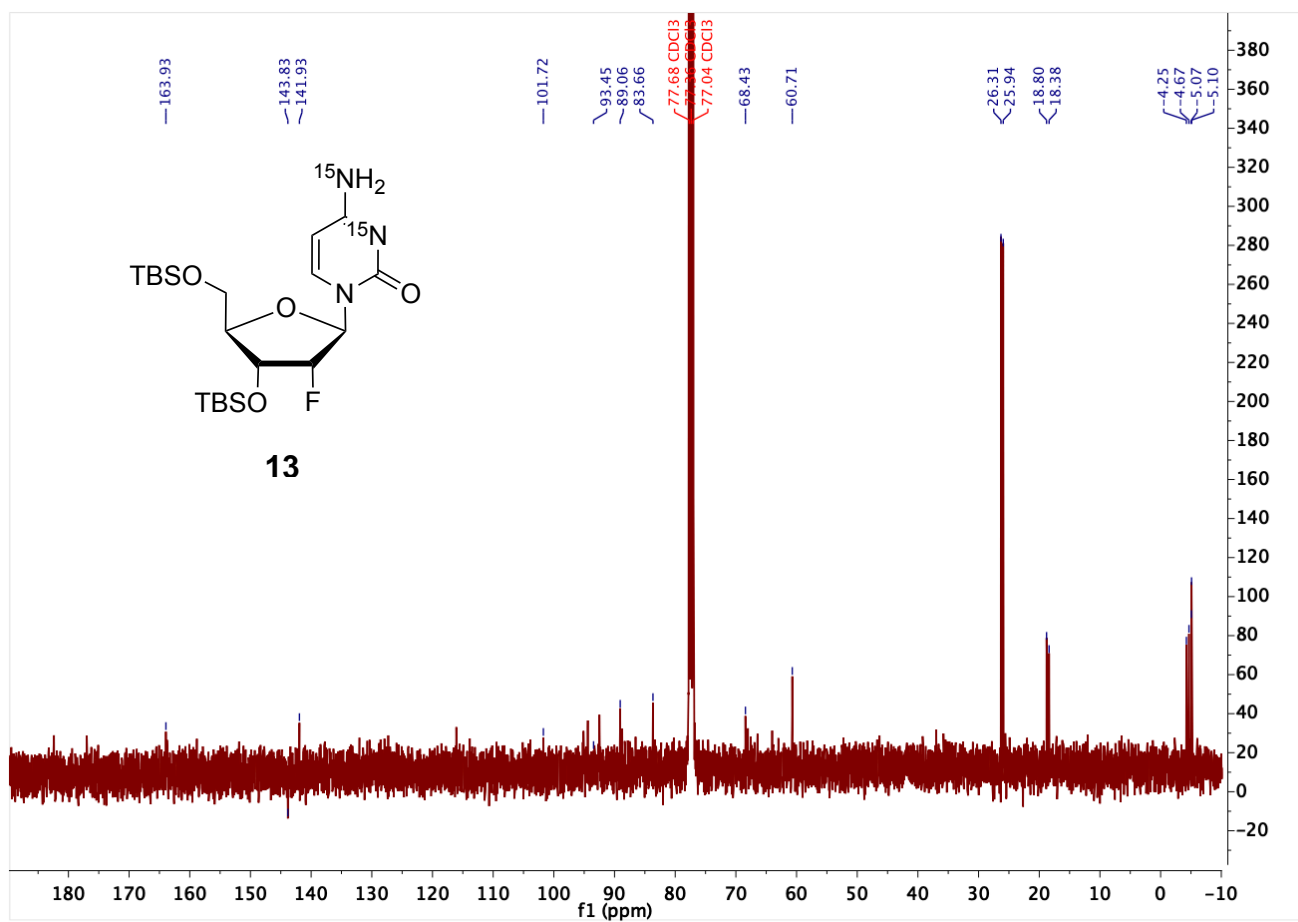
Figure SI-24: ^{13}C-NMR spectrum of compound **13**.
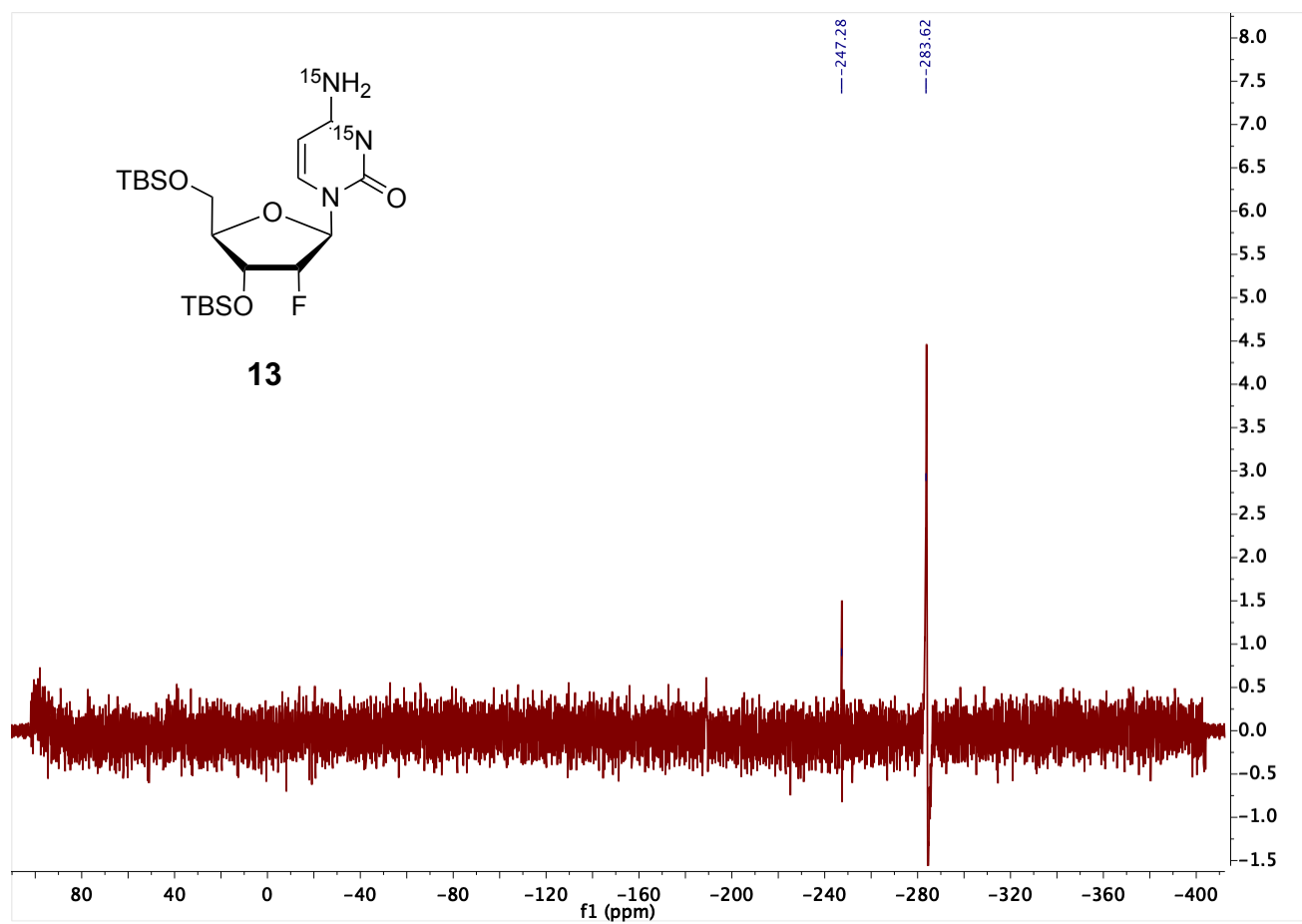
Figure SI-25: ^{15}N-NMR spectrum of compound **13**.
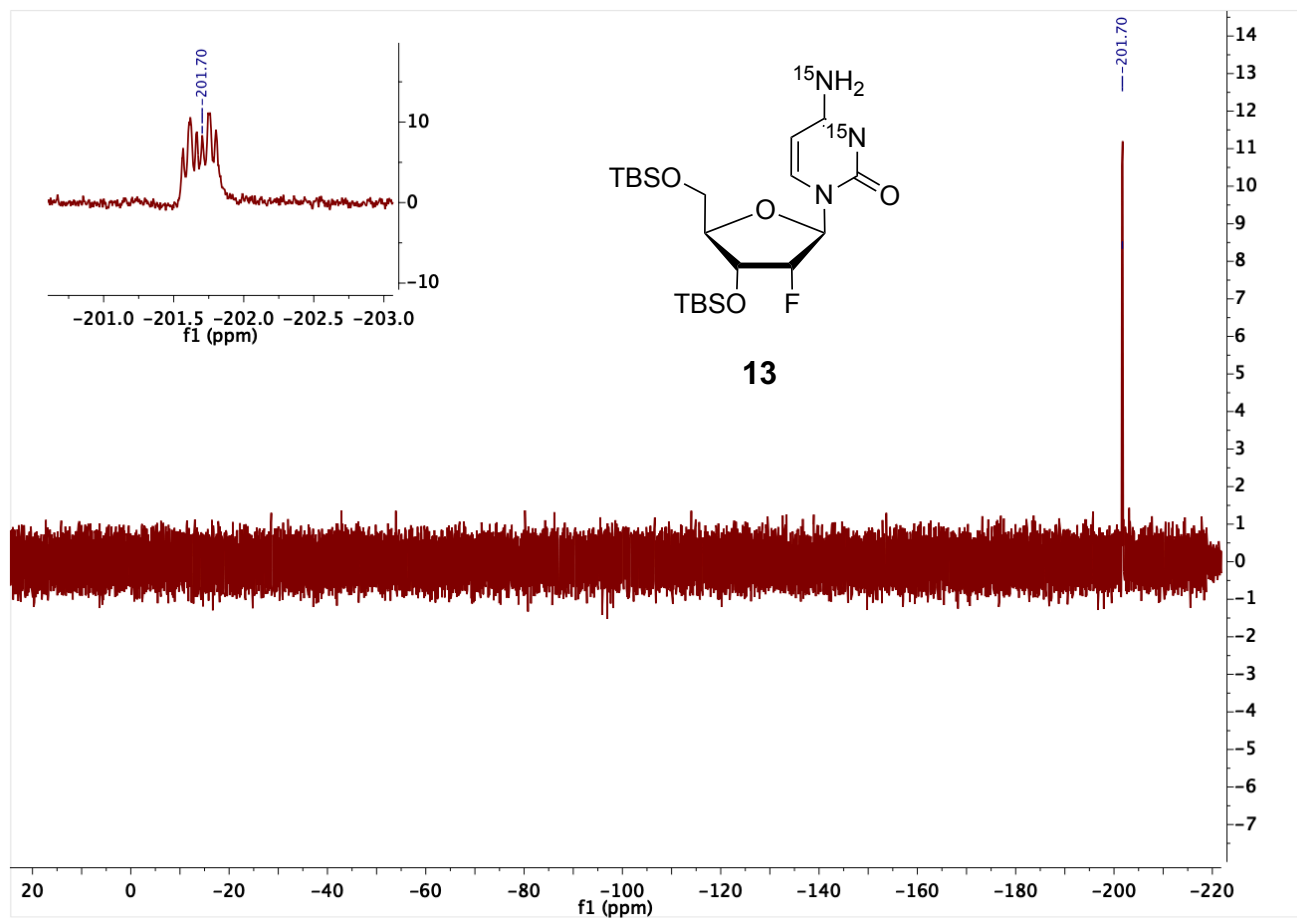
Figure SI-26: ^{19}F-NMR spectrum of compound **13**.

3',5'-bis-*O*-(*tert*-Butyl(dimethyl)silyl)-5-iodo-2'-deoxy-2'-(*R*)-fluoro-(*N*³,*N*⁴-¹⁵N₂)-cytidine (14)
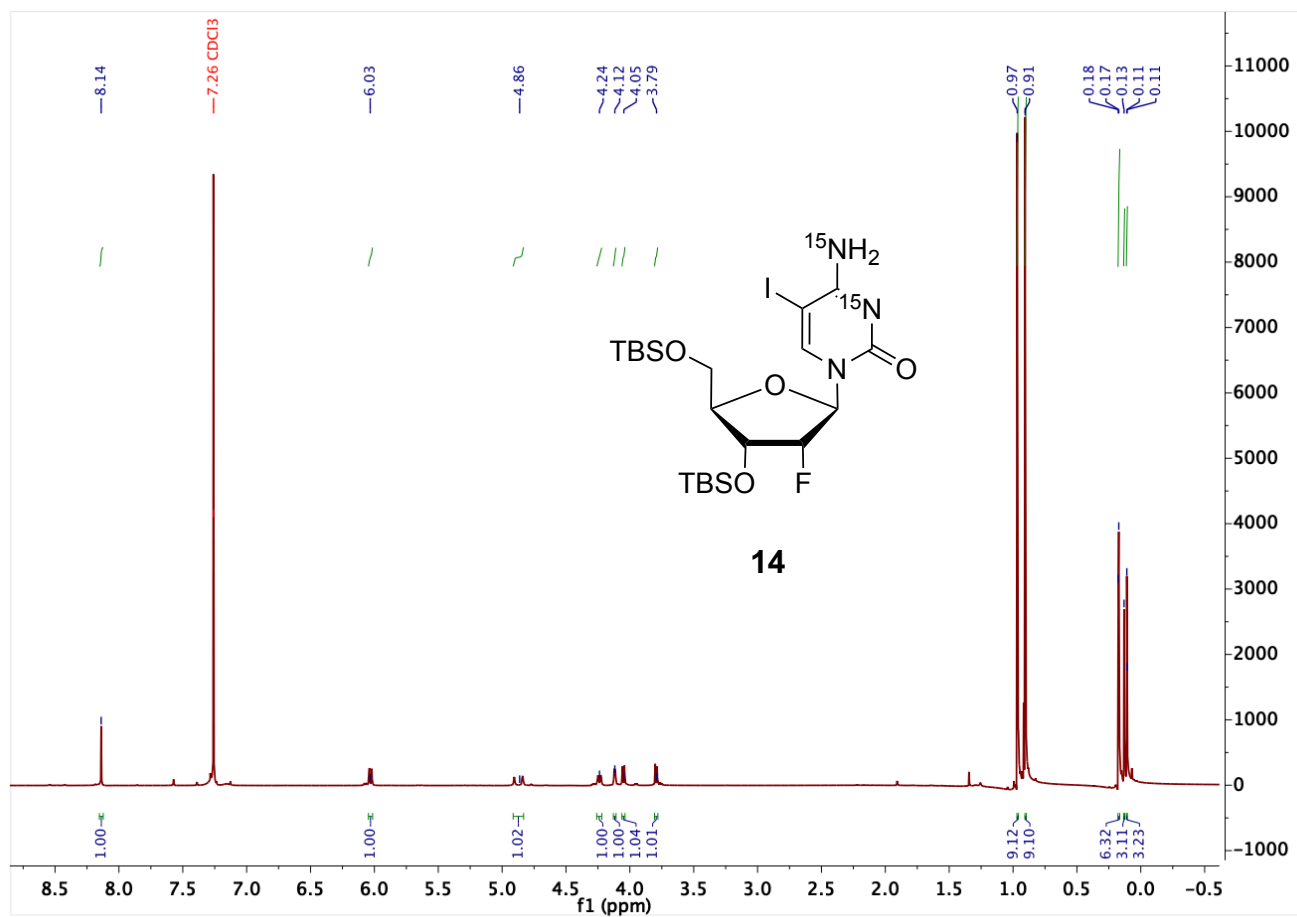
Figure SI-27: ¹H-NMR spectrum of compound **14**.
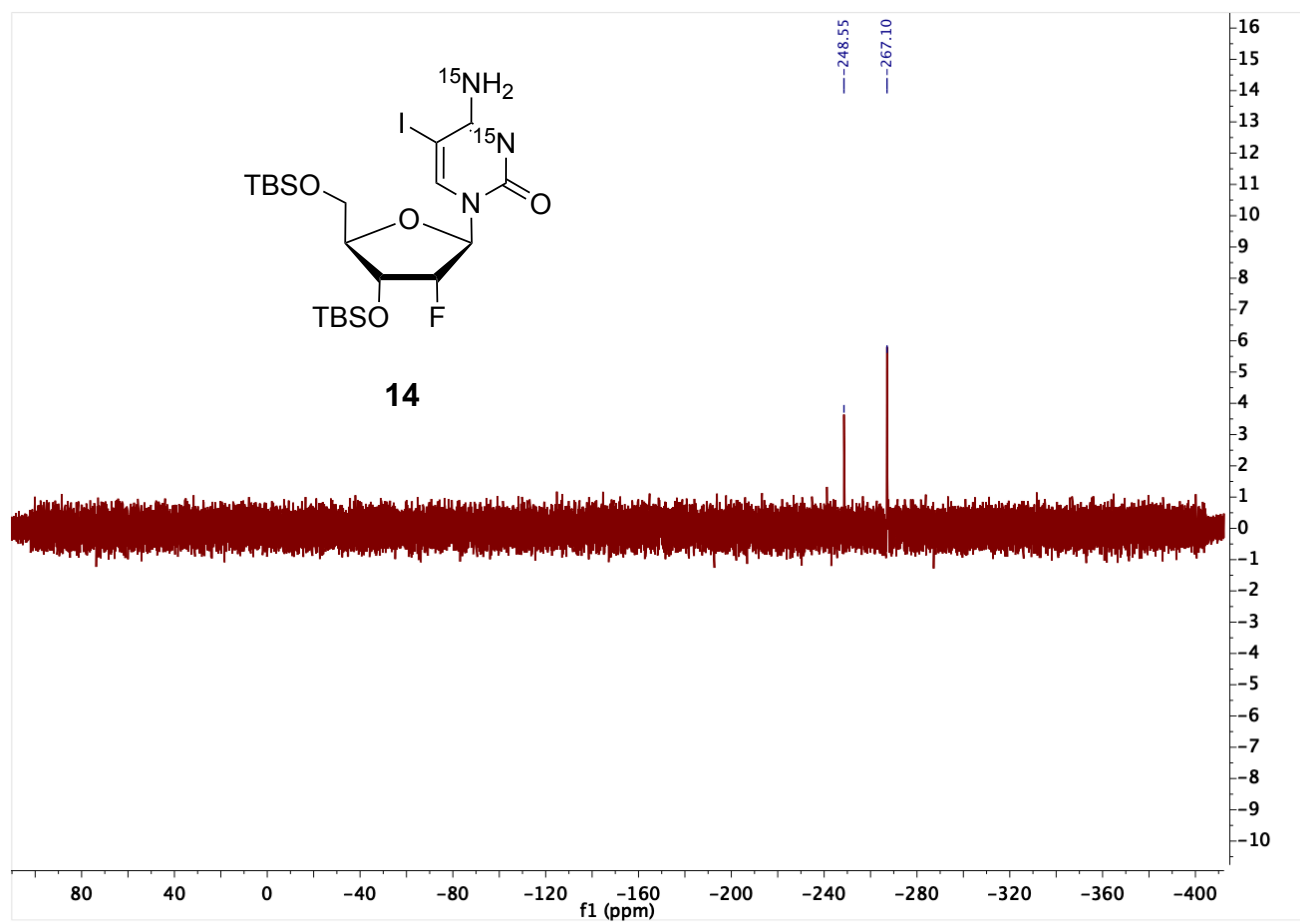
Figure SI-28: ^{15}N-NMR spectrum of compound **14**.
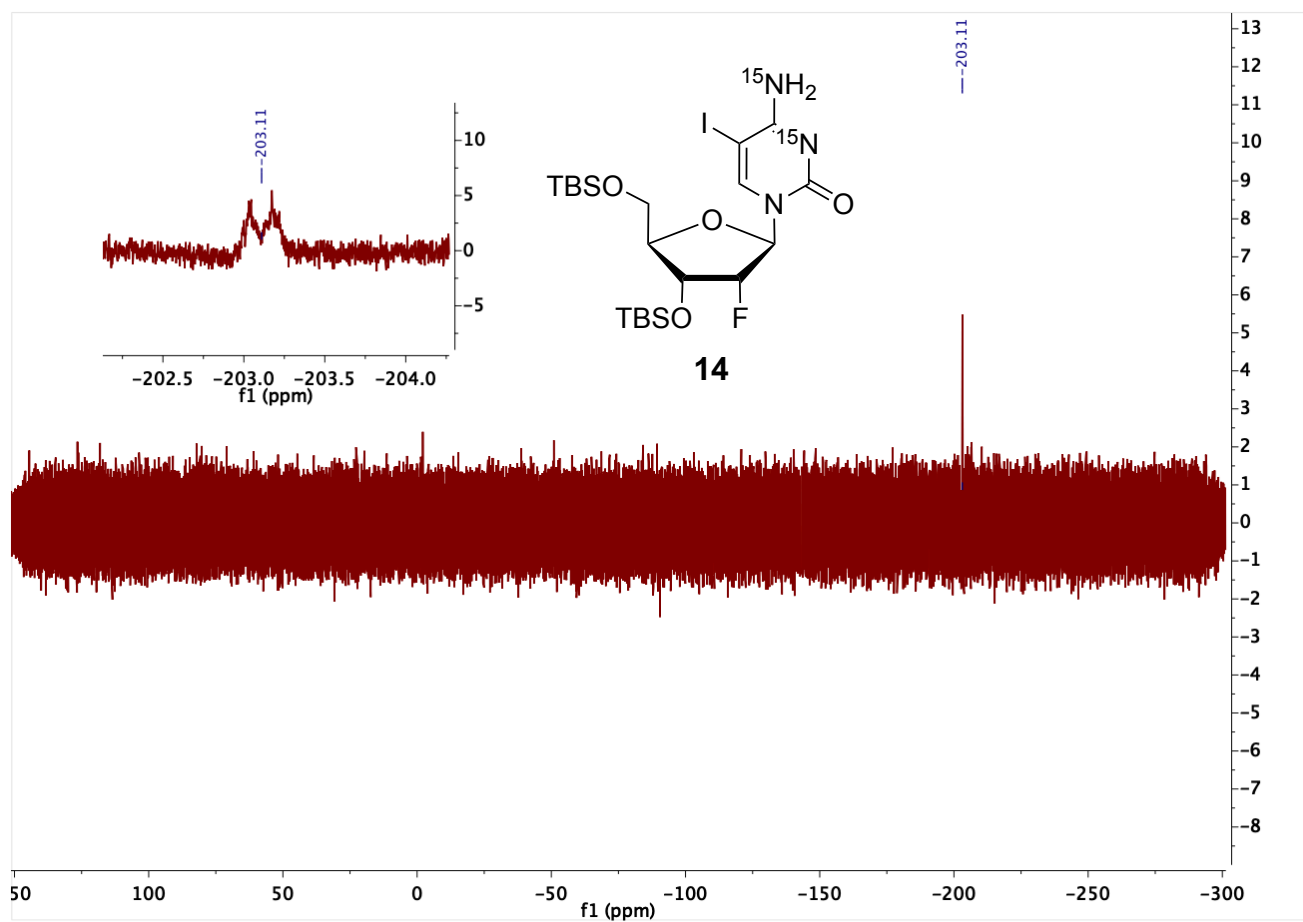
Figure SI-29: ^{19}F-NMR spectrum of compound **14**.

3',5'-bis-*O*-(*tert*-Butyl(dimethyl)silyl)-5-(methylcarboxy)-2'-deoxy-2'-(*R*)-fluoro-(*N*³,*N*⁴-¹⁵N₂)-cytidine (15)
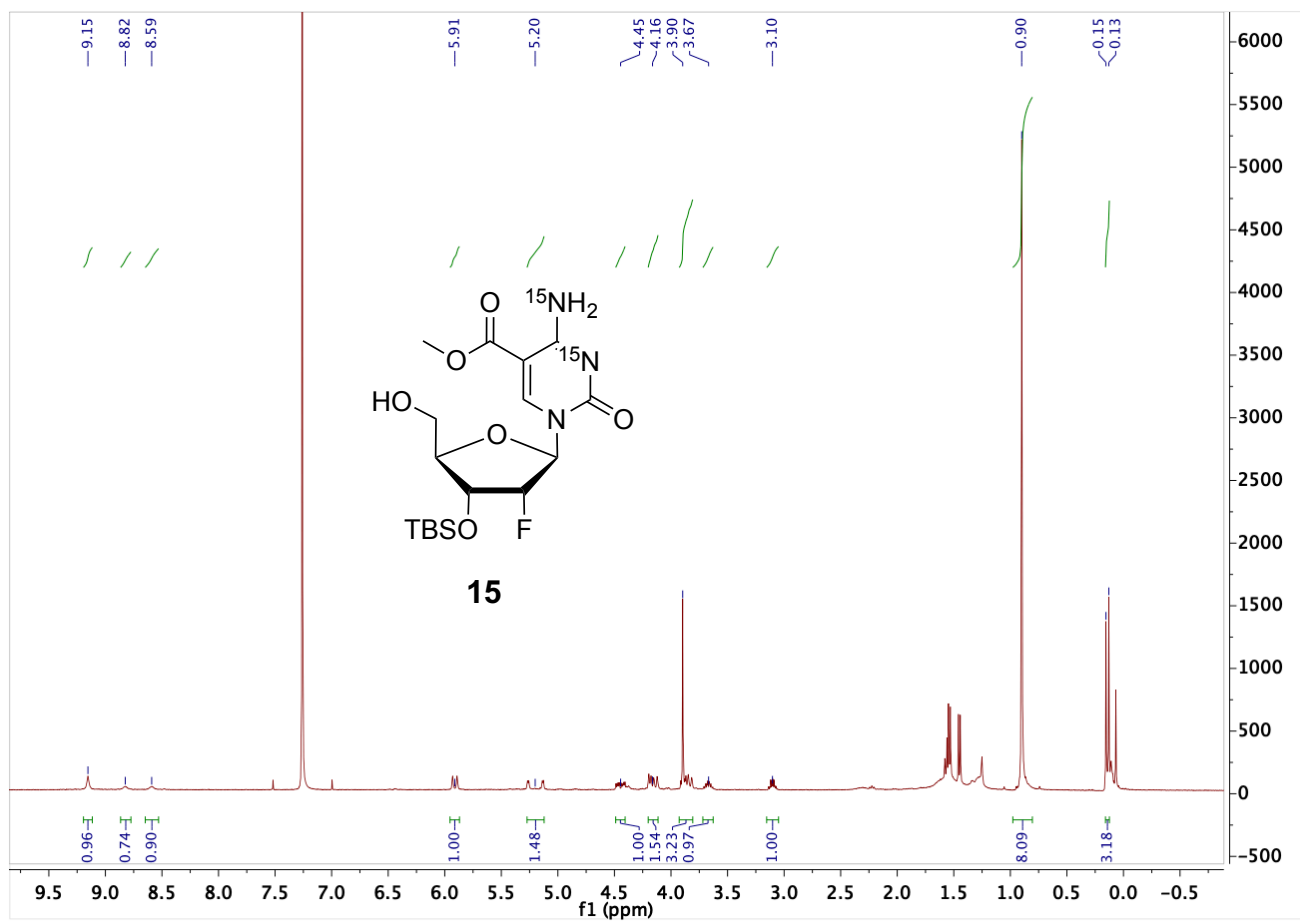
Figure SI-30: ¹H-NMR spectrum of compound **15**.
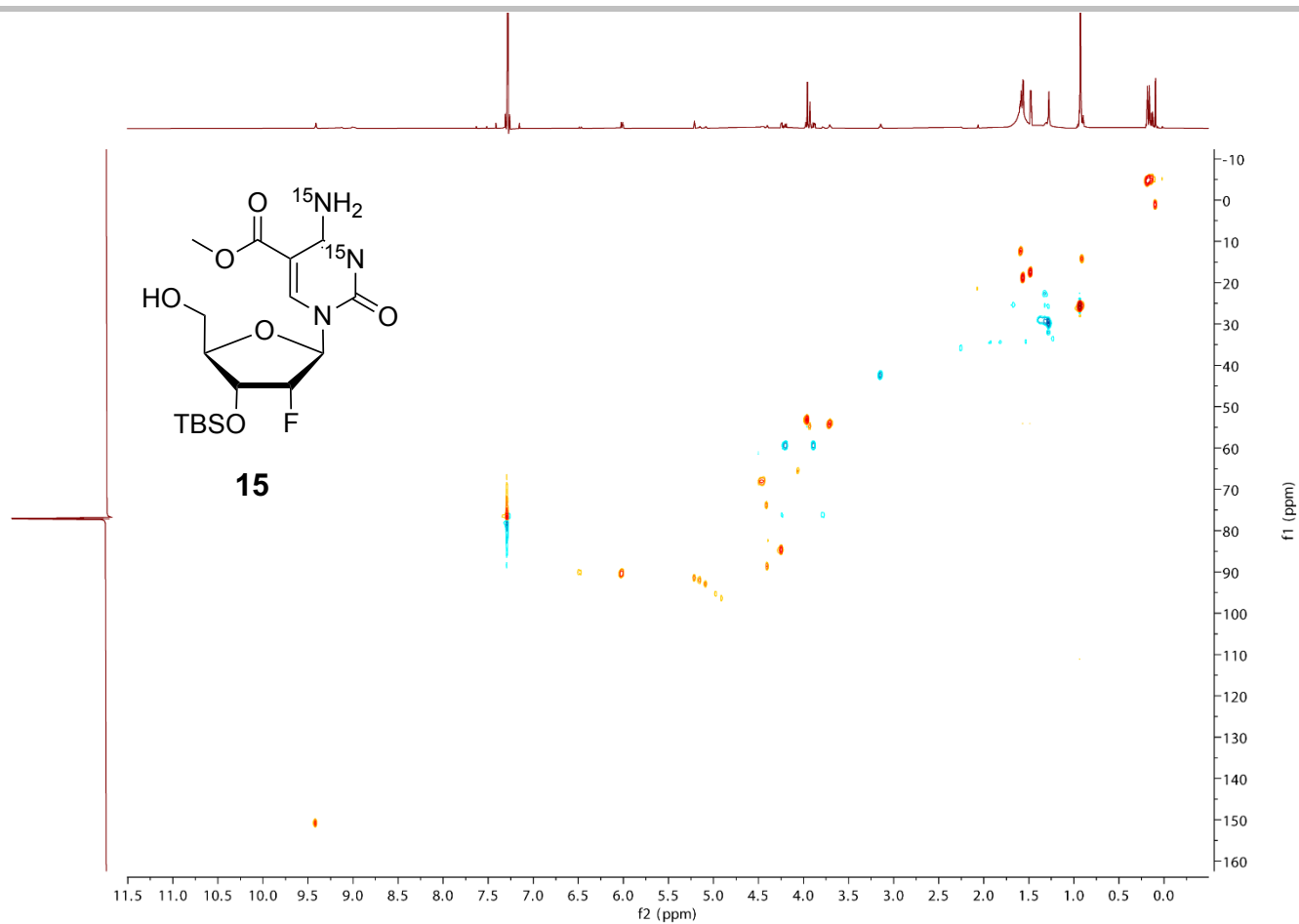
Figure SI-31: ^{1}H-^{13}C -HSQC-NMR spectrum of compound **15**.
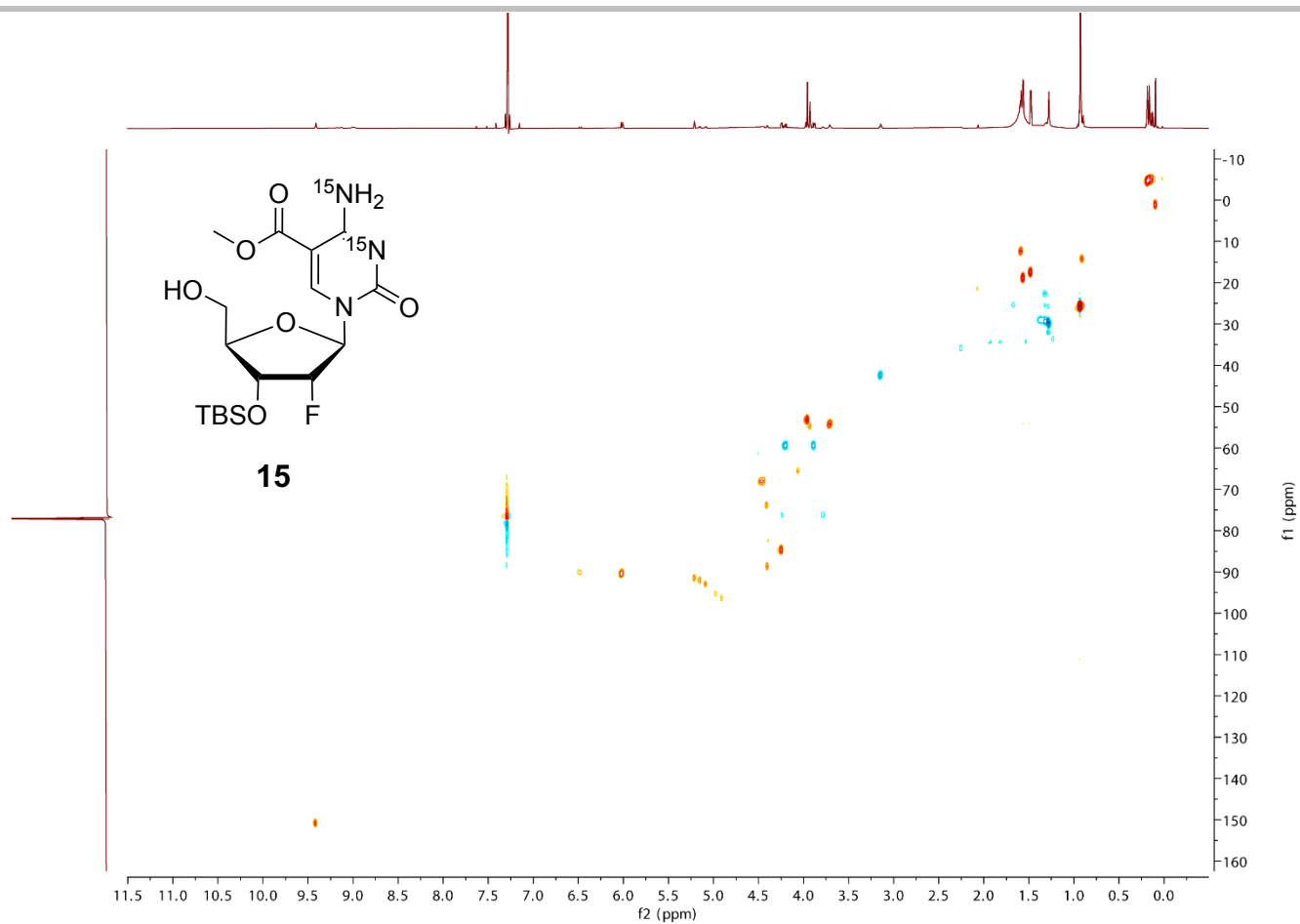
Figure SI-32: ^{1}H-^{13}C -HMBC-NMR spectrum of compound **15**.
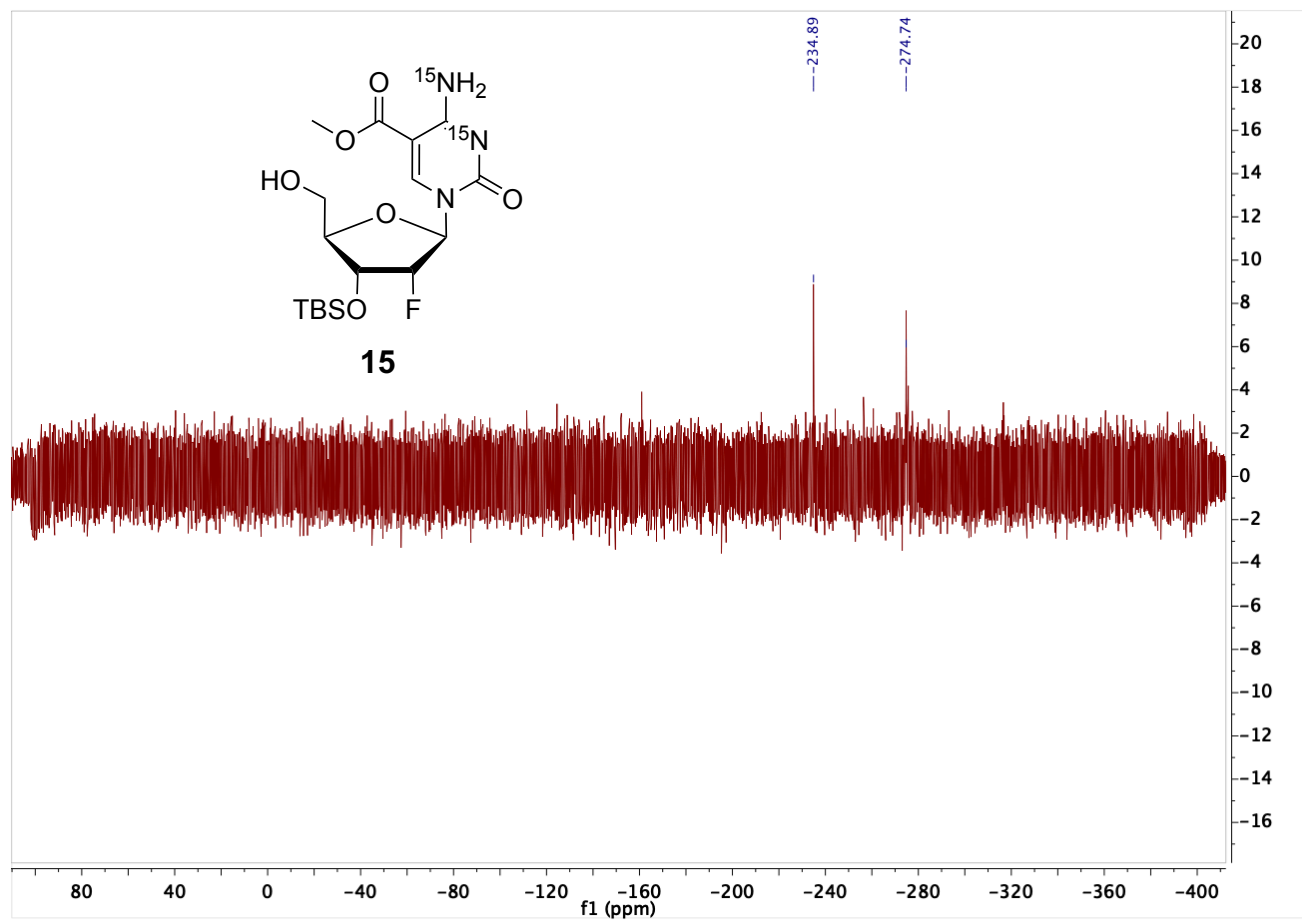
Figure SI-33: ^{15}N-NMR spectrum of compound **15**.

5-(Carboxy)-2'-deoxy-2'-(R)-fluoro-(N^3,N^4-$^{15}N_2$)-cytidine (3)
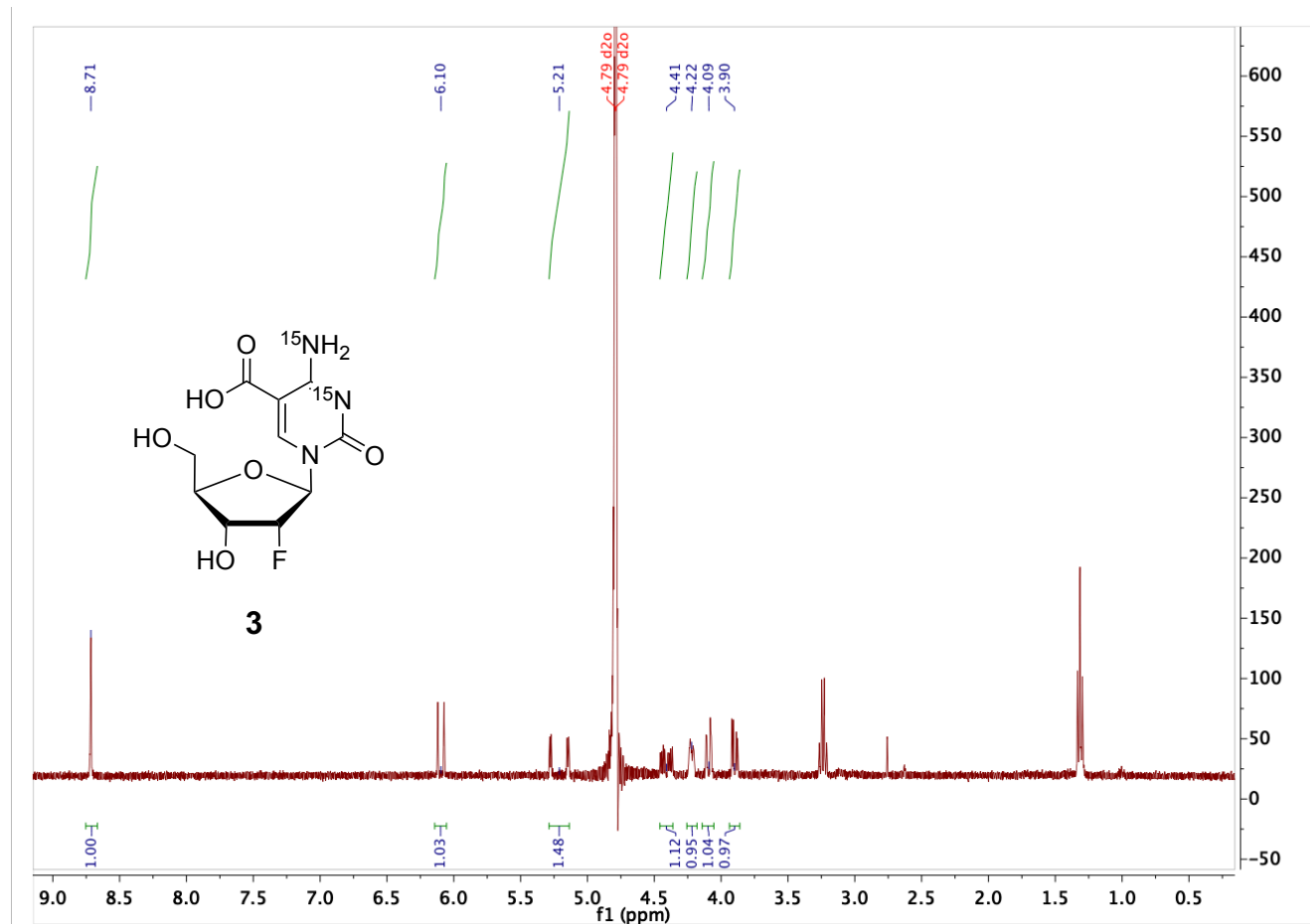
Figure SI-34: ^{1}H-NMR spectrum of compound **3**.
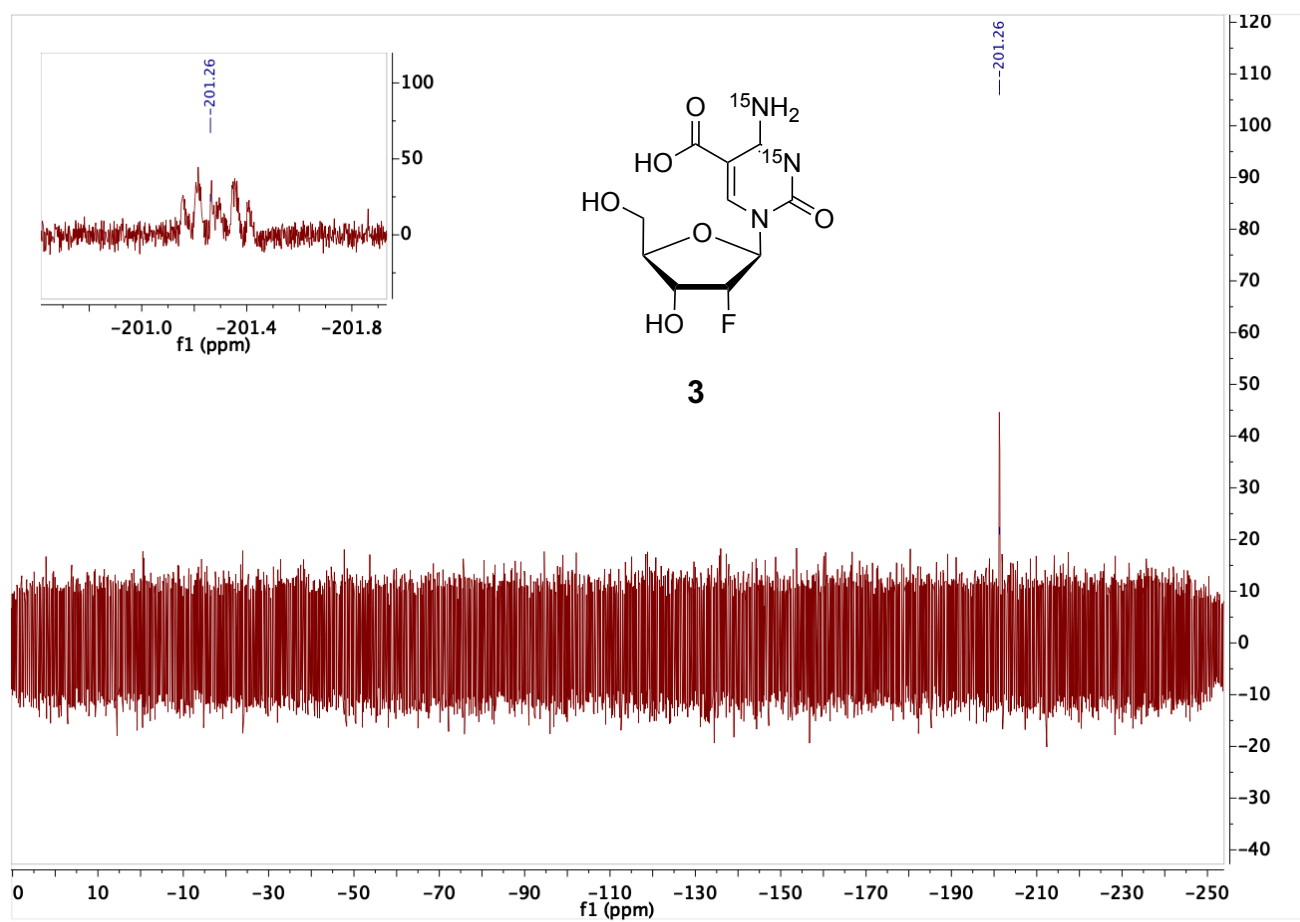
Figure SI-35: ^{19}F-NMR spectrum of compound **3**.
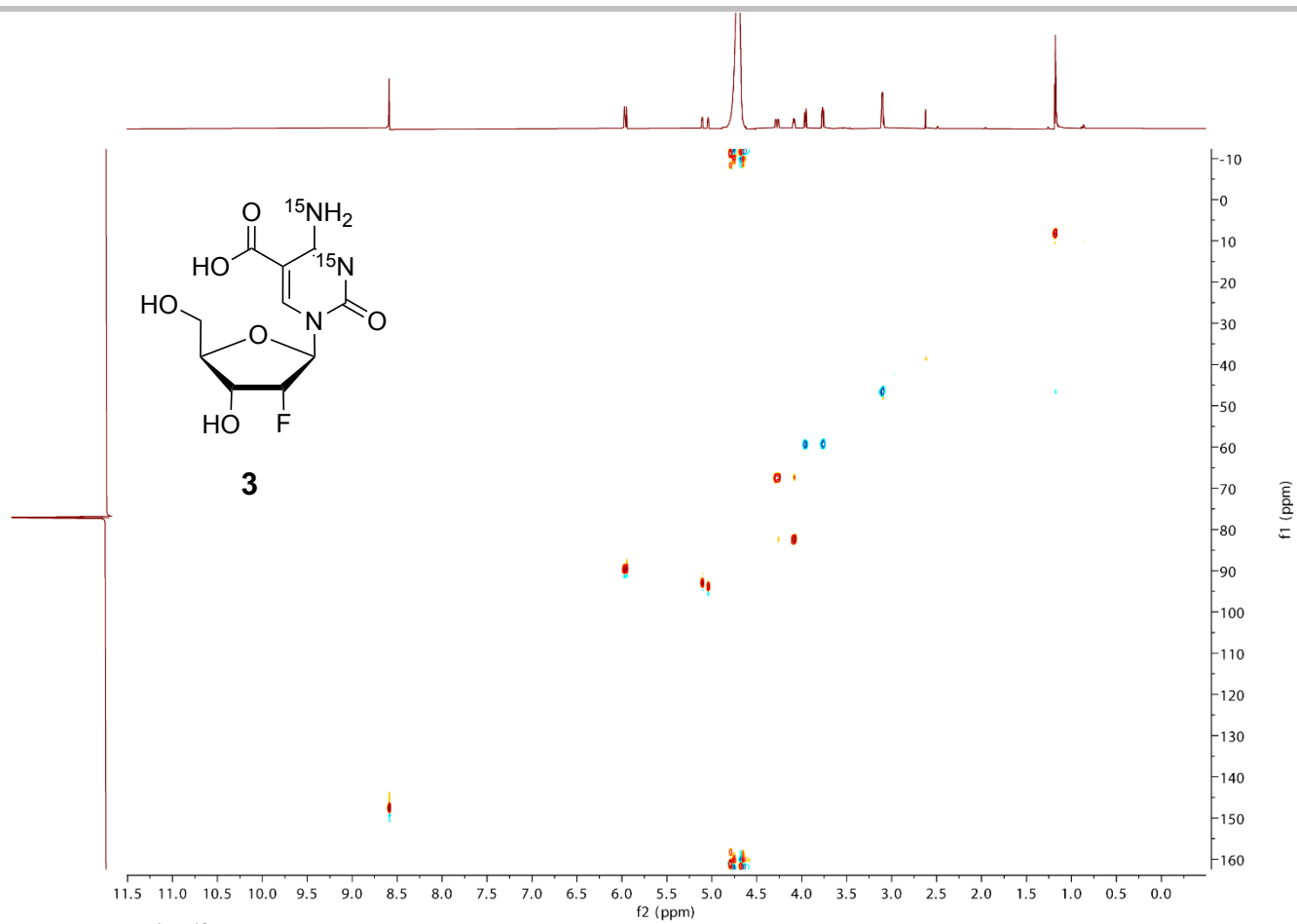
Figure SI-36: ^{1}H-^{13}C -HSQC-NMR spectrum of compound **3**.
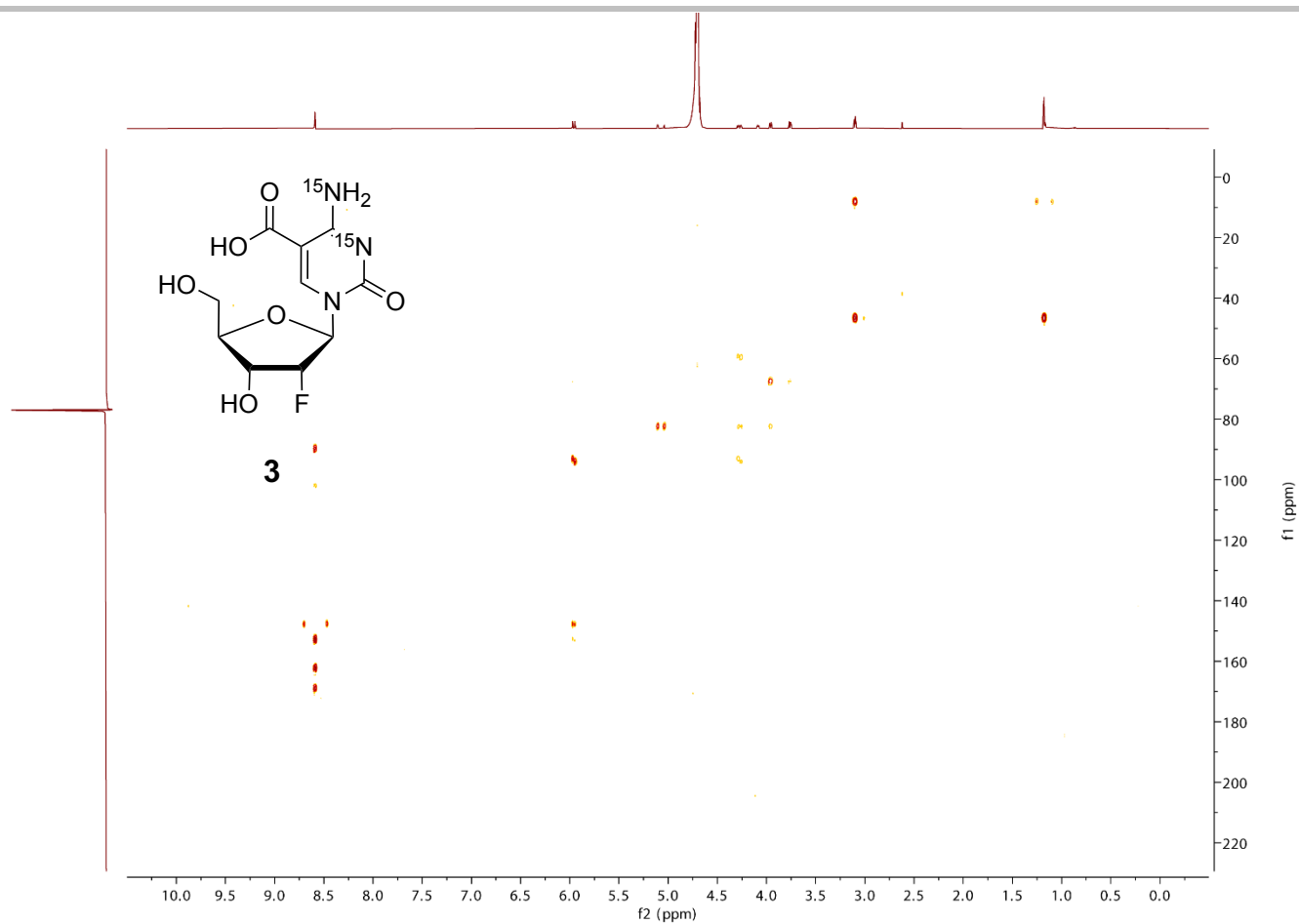
Figure SI-37: ^{1}H-^{13}C -HMBC-NMR spectrum of compound **3**.

3'-5'-O-bis(*tert*-Butyldimethylsilyl)-2'-desoxy-2'-(R)-fluoro-cytidine (16)
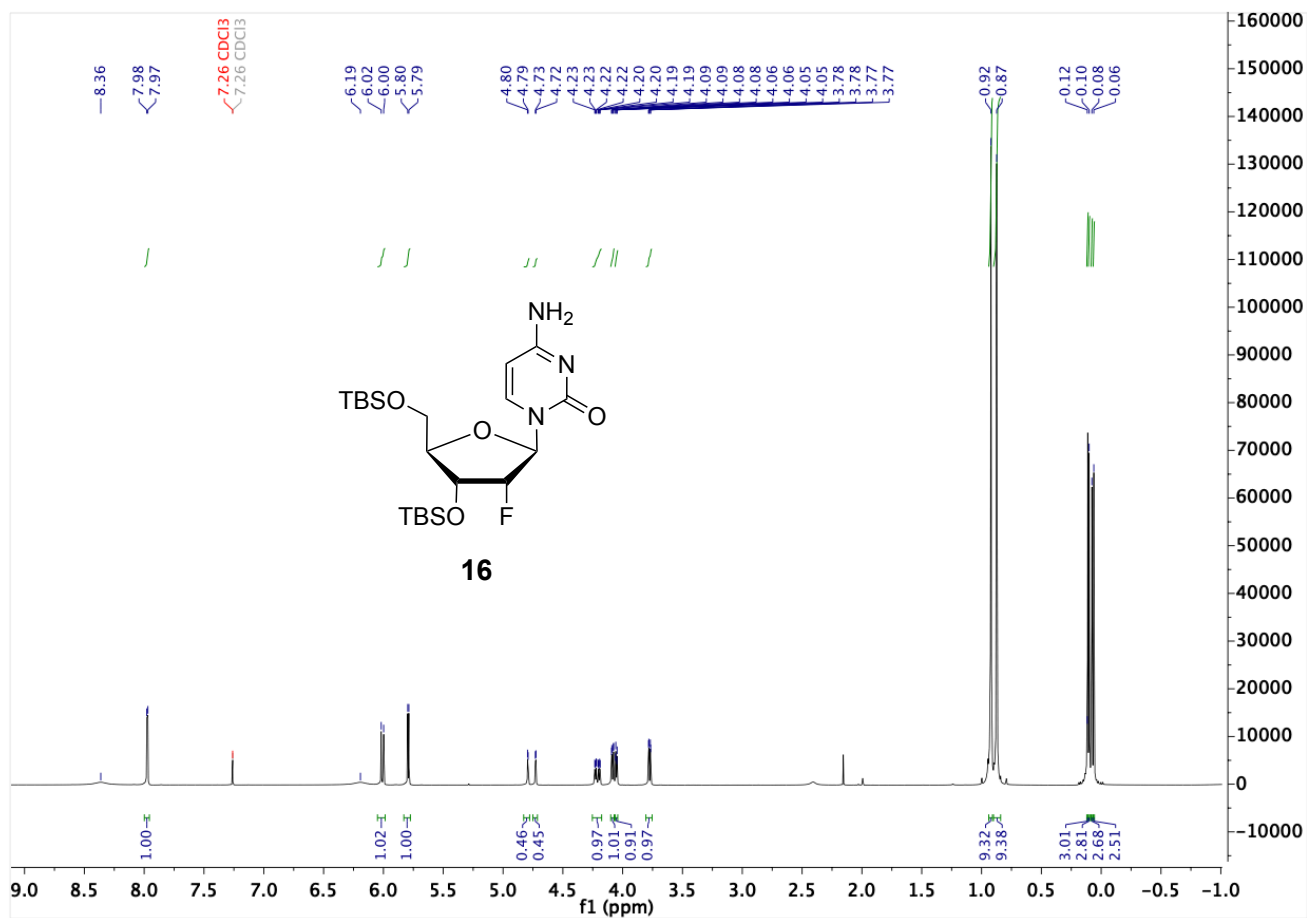
Figure SI-38: ^{1}H-NMR spectrum of compound **16**.
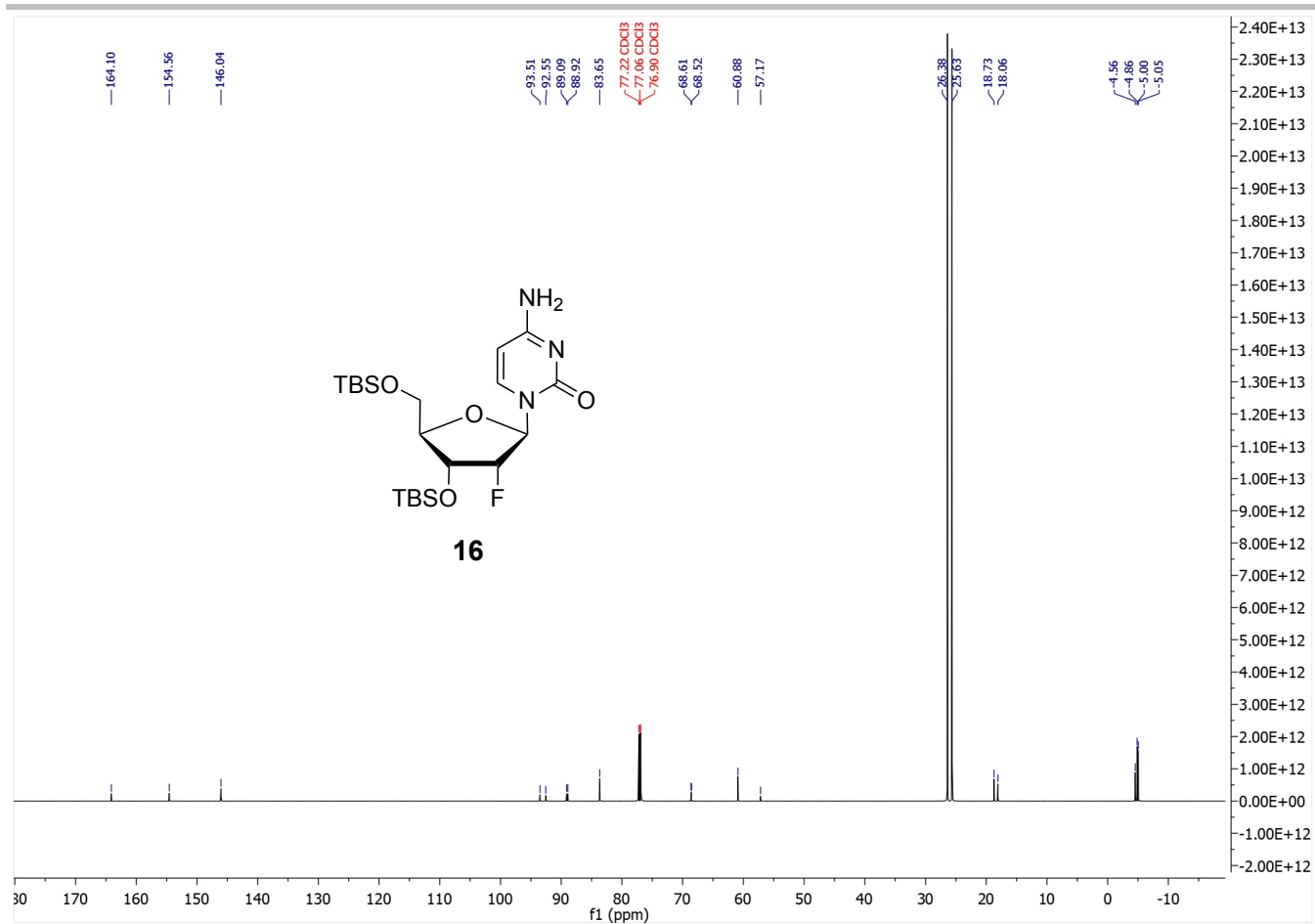
Figure SI-39: ^{13}C-NMR spectrum of compound **16**.
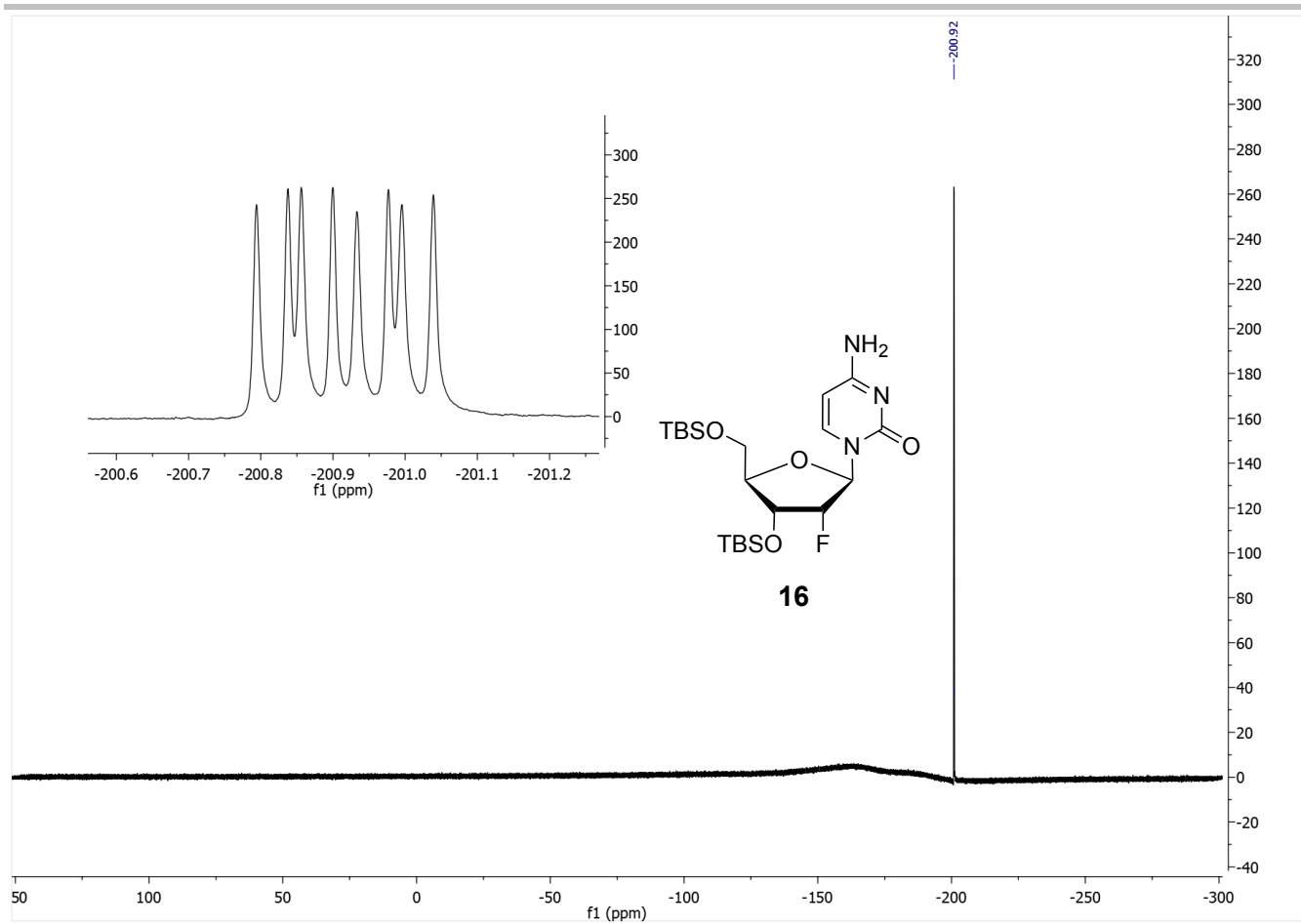
Figure SI-40: ^{19}F-NMR spectrum of compound **16**.

3'-5'-O-bis(*tert*-Butyldimethylsilyl)-5-iodo-2'-desoxy-2'-(R)-fluoro-cytidine (17)
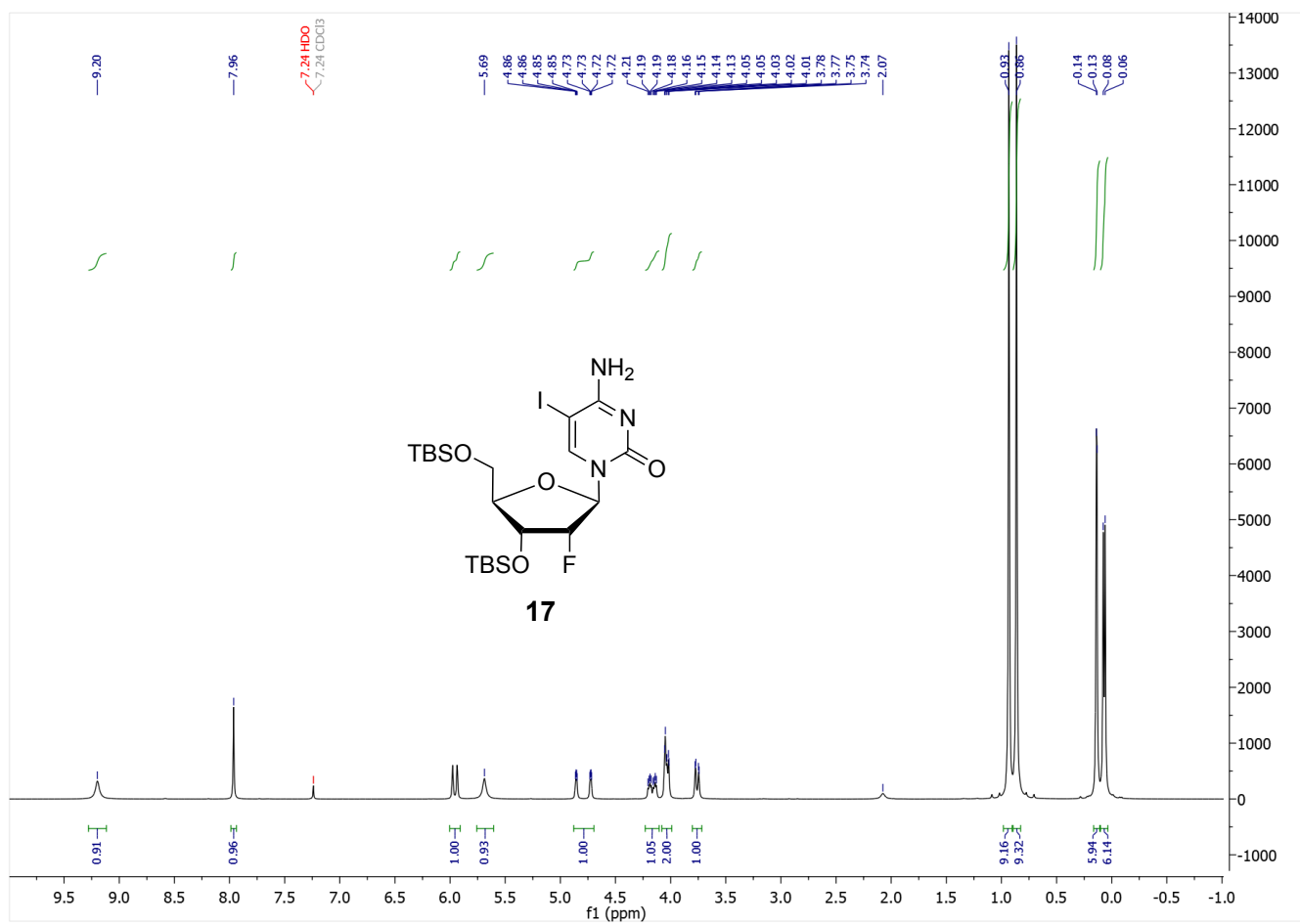
Figure SI-41: ^{1}H-NMR spectrum of compound **17**.
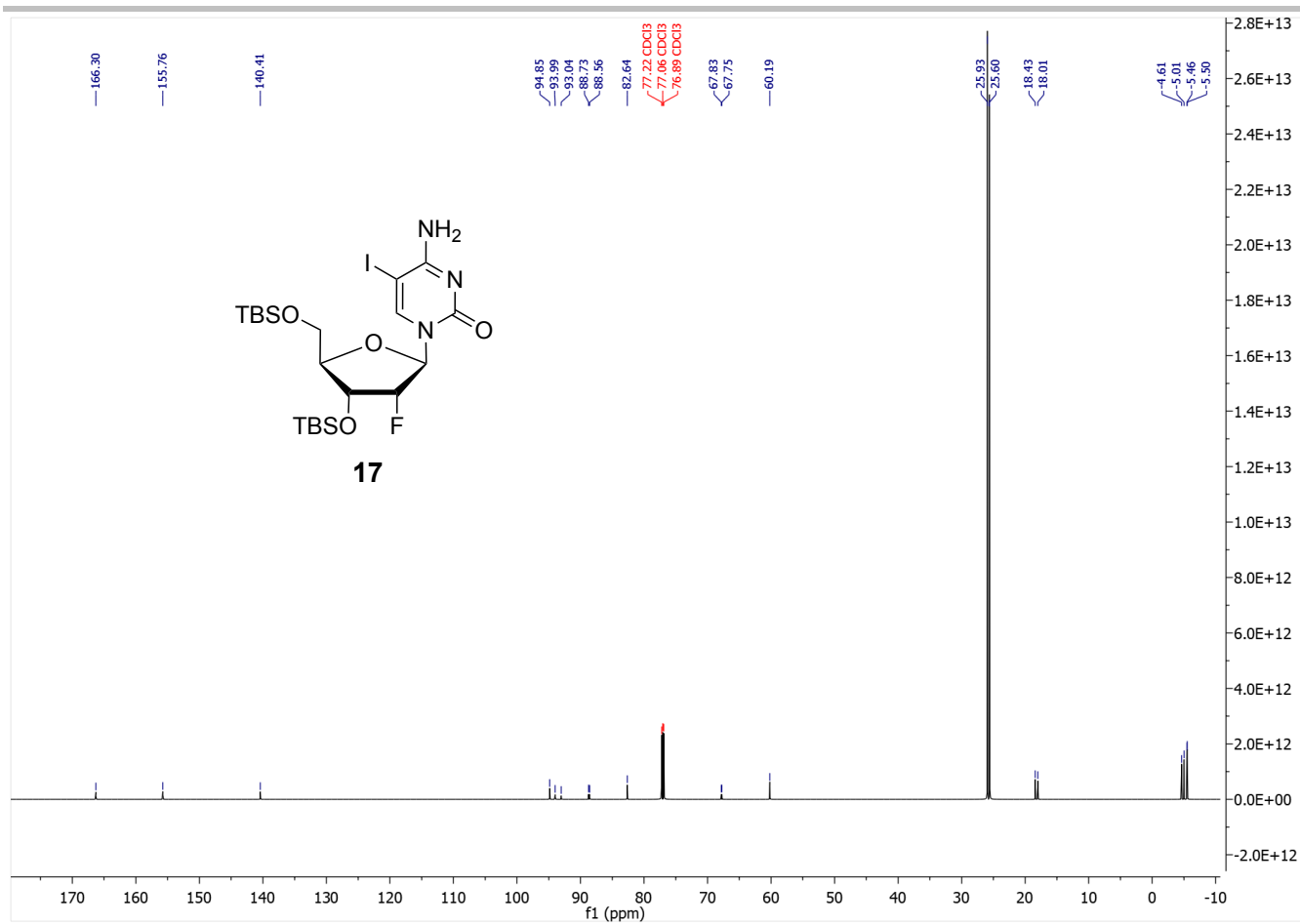
Figure SI-42: ^{13}C-NMR spectrum of compound **17**.

Figure SI-43: ^{19}F-NMR spectrum of compound **17**.

3'-5'-*O*-bis(*tert*-Butyldimethylsilyl)-5-methoxycarbonyl-2'-desoxy-2'-(R)-fluoro-cytidine (18)

Figure SI-44: ^{1}H-NMR spectrum of compound **18**.
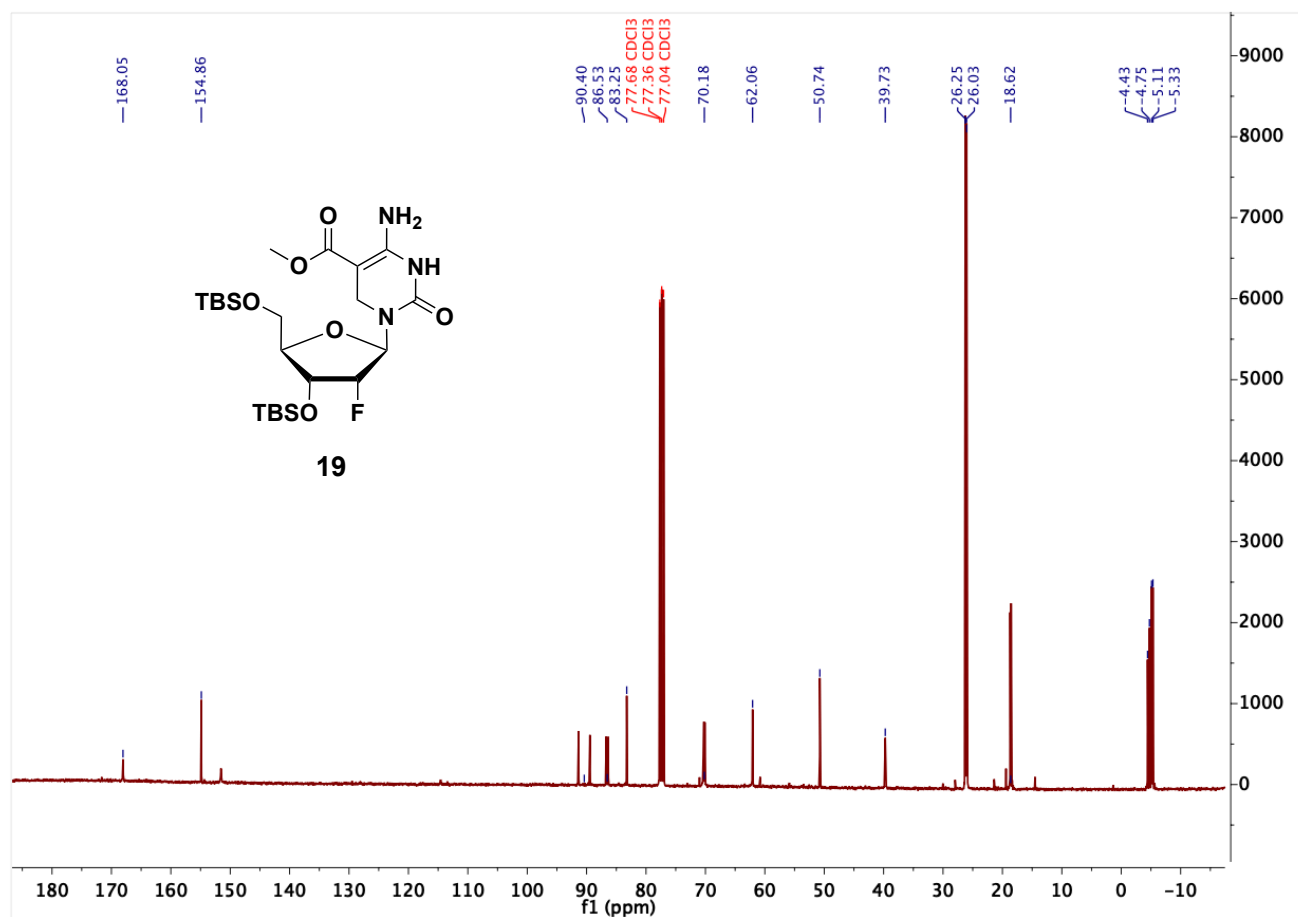
SUPPORTING INFORMATION
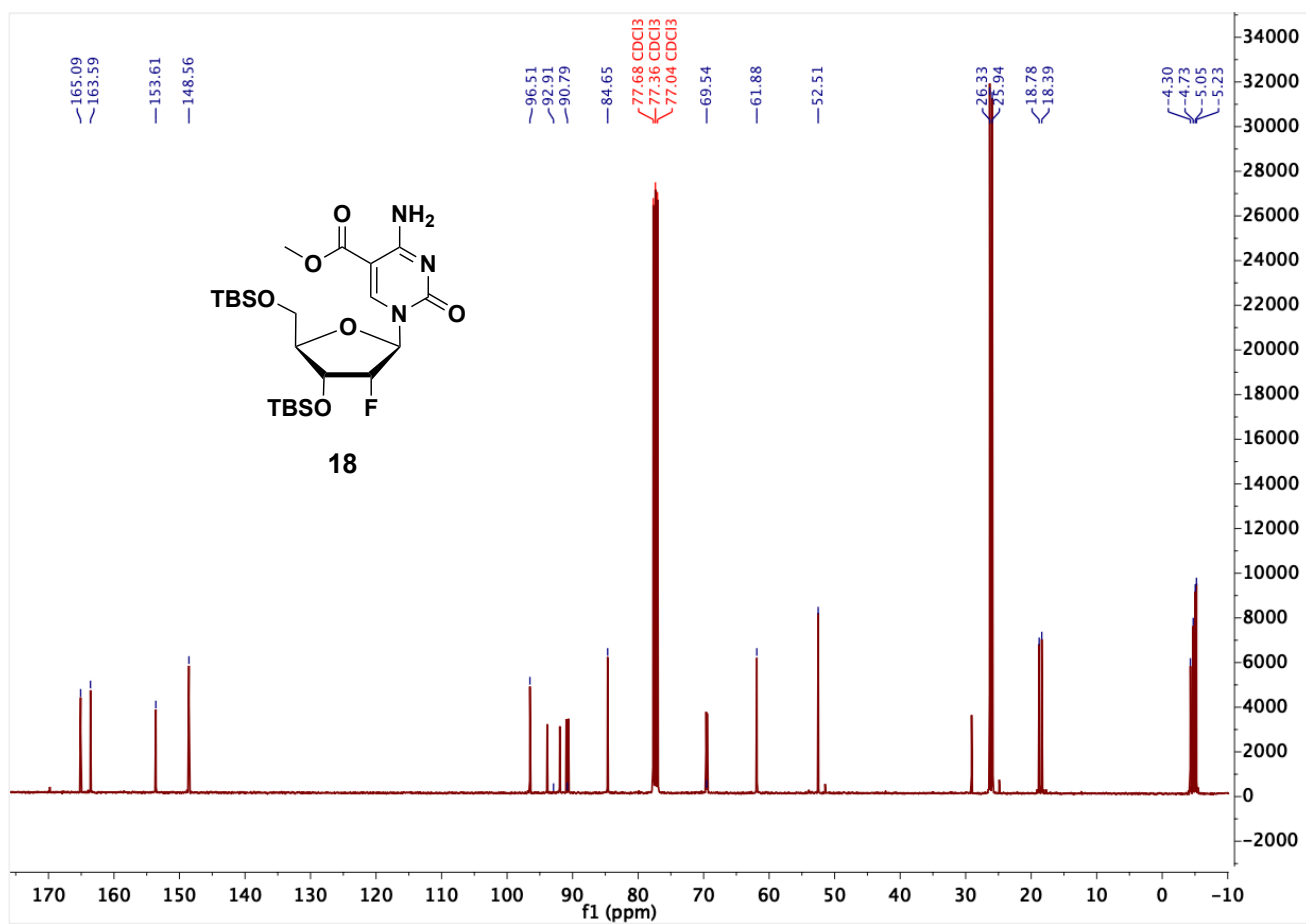
Figure SI-45: ^{13}C-NMR spectrum of compound **18**.
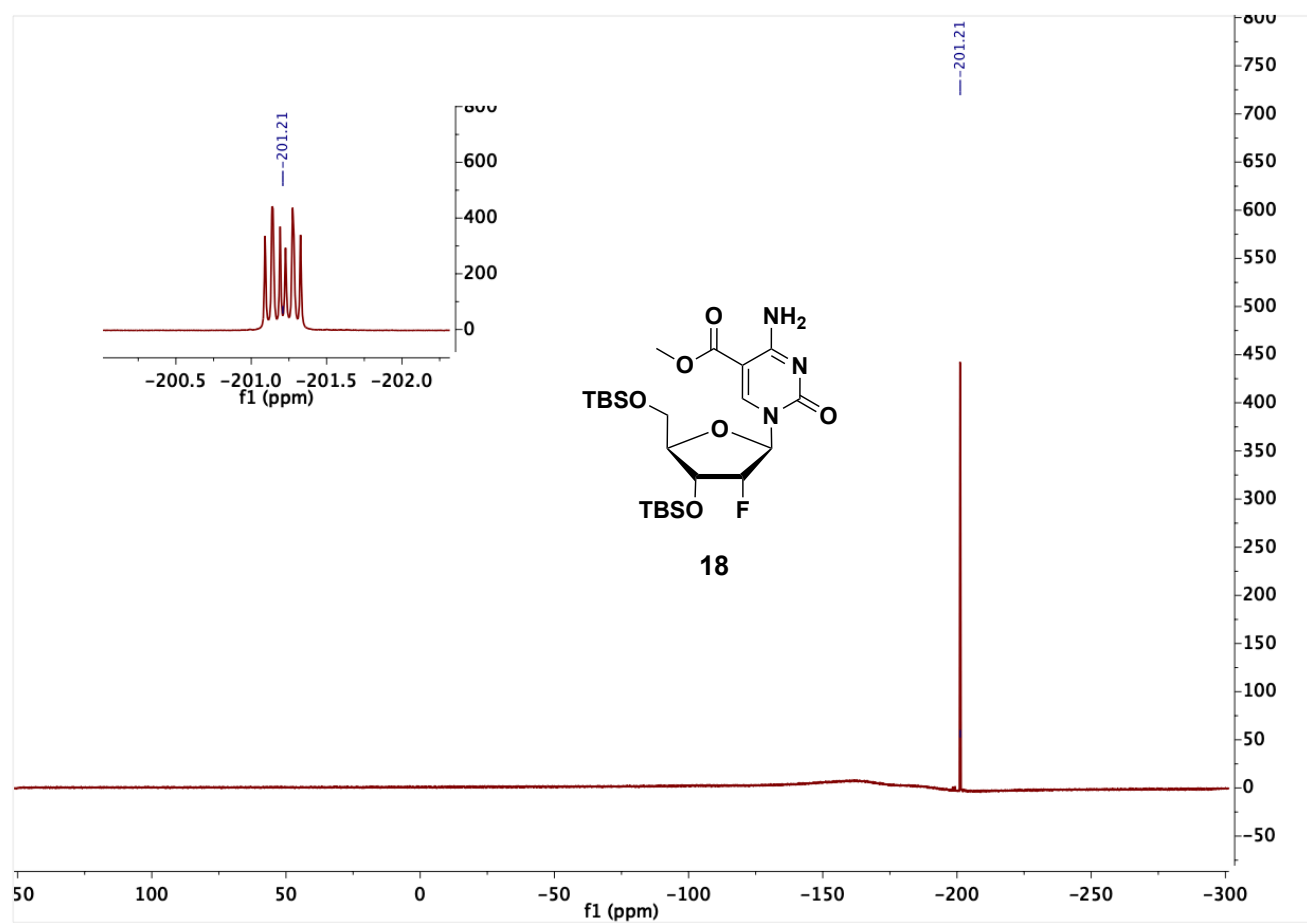
Figure SI-46: ^{19}F-NMR spectrum of compound **18**.

3'-5'-O-bis(*tert*-Butyldimethylsilyl)-5-methoxycarbonyl-2'-desoxy-2'-(R)-fluoro-5,6-dihydrocytidine (19)
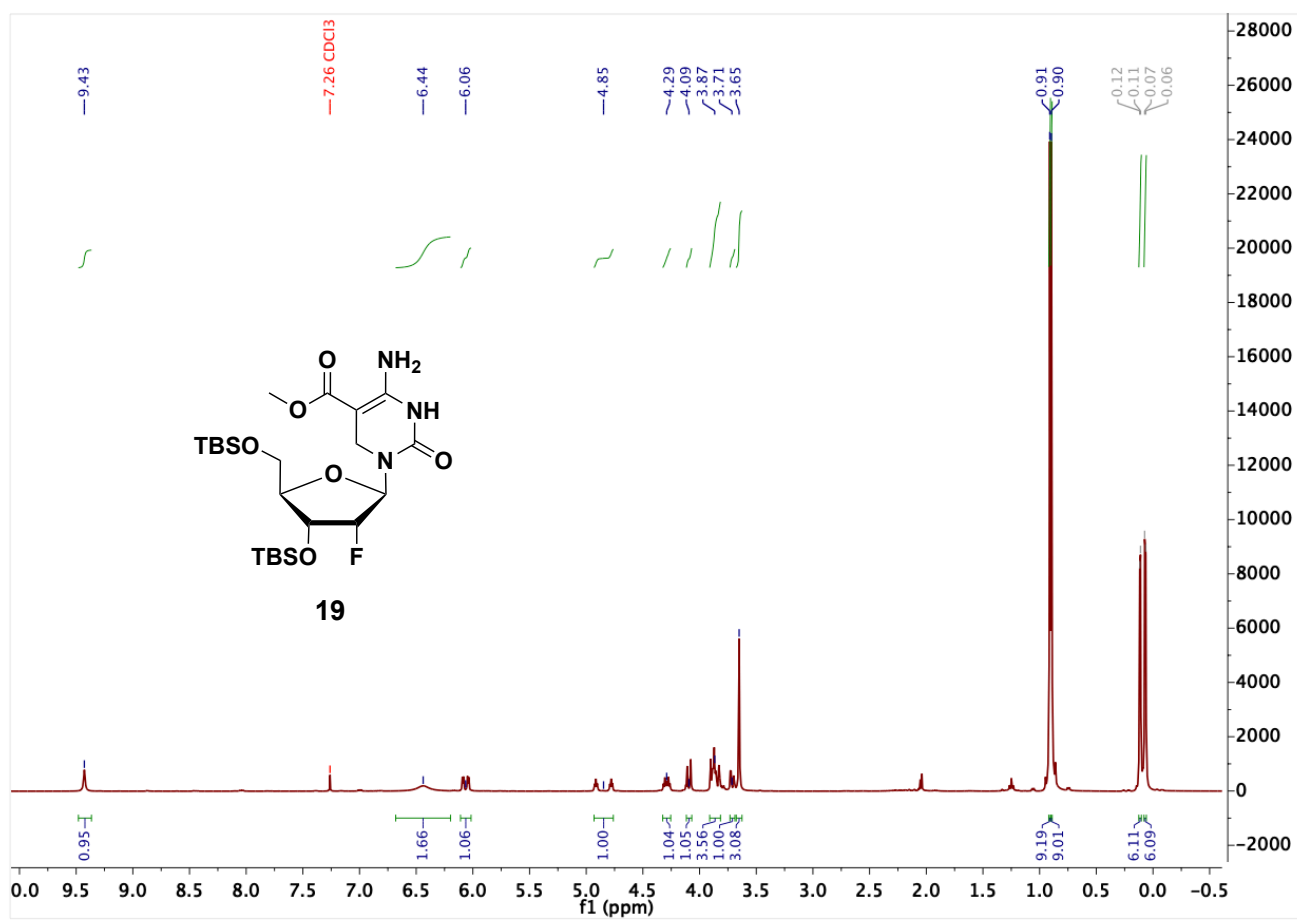
Figure SI-47: ^{1}H-NMR spectrum of compound **19**.

Figure SI-48: ^{13}C-NMR spectrum of compound **19**.
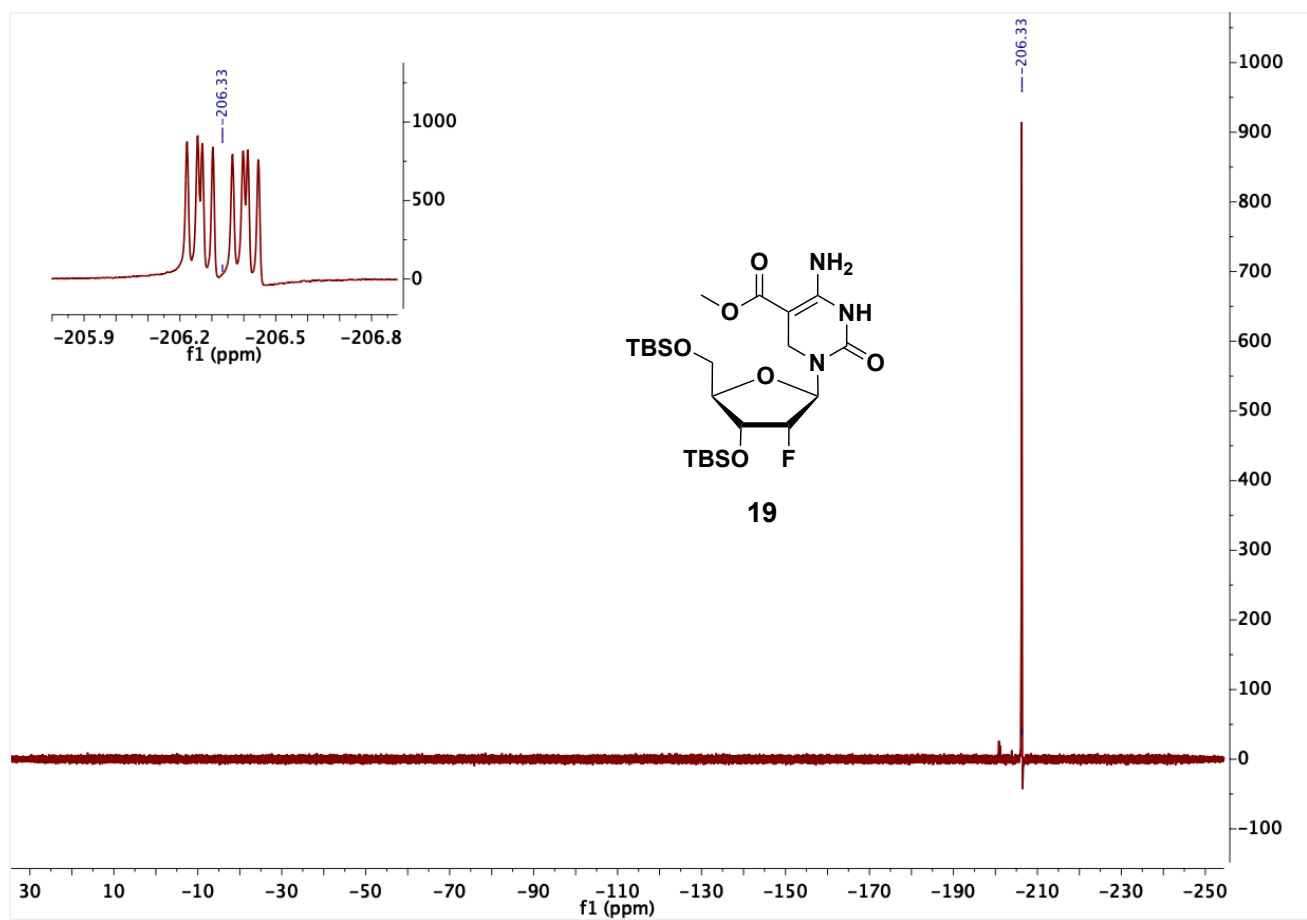
Figure SI-49: ^{19}F-NMR spectrum of compound **19**.

3'-5'-O-bis(*tert*-Butyldimethylsilyl)-2'-desoxy-2'-(R)-fluoro-5,6-dihydrocytidine (20)
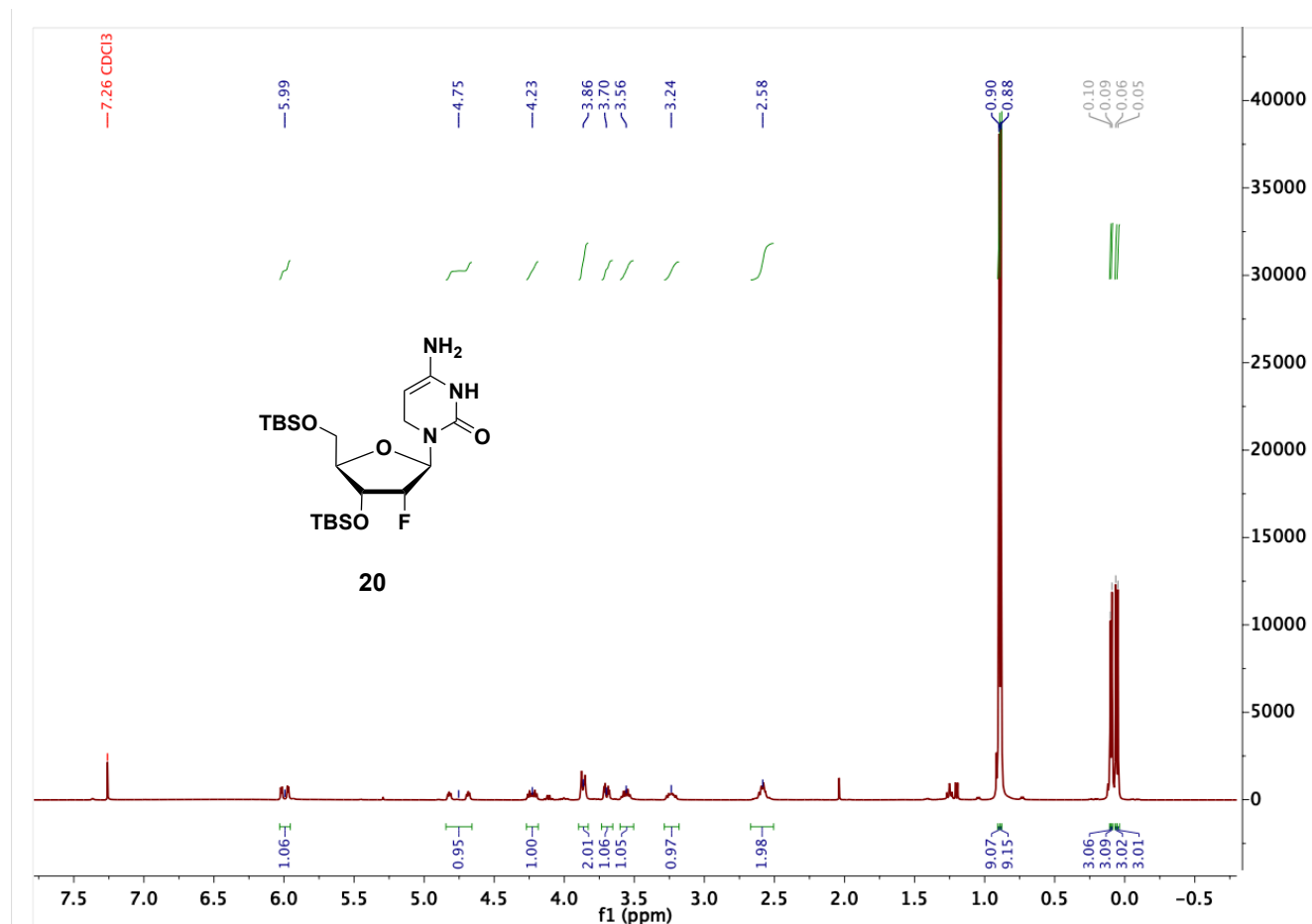
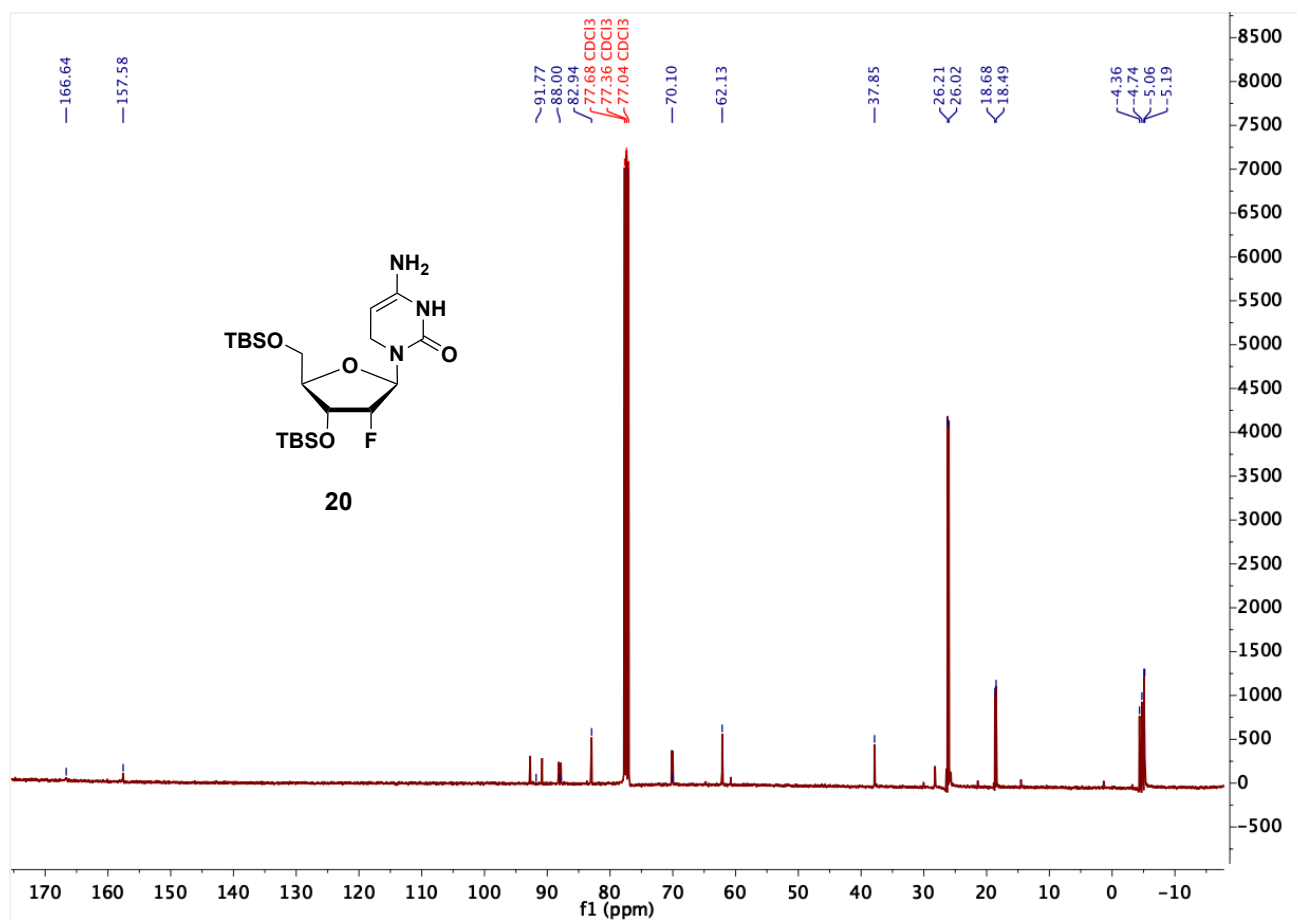**Figure SI-50**: ^{1}H-NMR spectrum of compound **20**.

Figure SI-51: ^{13}C-NMR spectrum of compound **20**.
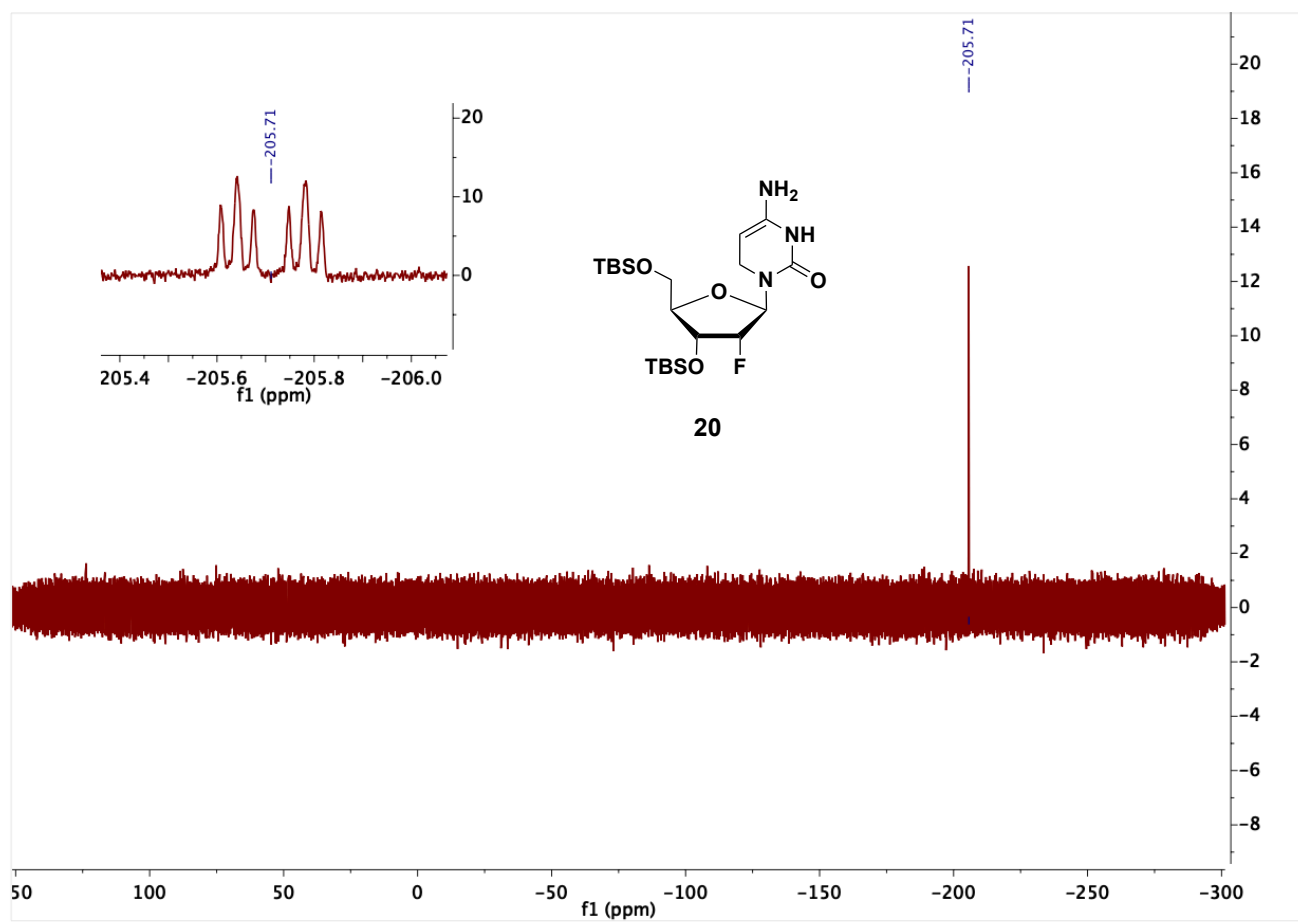
SUPPORTING INFORMATION

Figure SI-52: ^{19}F-NMR spectrum of compound **20**.

References

[1]   F. Yuan, Y. Bi, J.-Y. Zhang, Y.-L. Zhou, X.-X. Zhang, C.-X. Song, *RSC Adv.* **2019**, *9*, 29010-29014.
[2]   W. J. Lowary PT, *J Mol. Biol.* **1998**, *276*, 19-42.
[3]   K. Iwan, R. Rahimoff, A. Kirchner, F. Spada, A. S. Schröder, O. Kosmatchev, S. Ferizaj, J. Steinbacher, E. Parsa, M. Müller, T. Carell, *Nat. Chem. Biol.* **2018**, *14*, 72-78.
[4]   F. R. Traube, S. Schiffers, K. Iwan, S. Kellner, F. Spada, M. Müller, T. Carell, *Nat. Protoc.* **2019**, *14*, 283-312.
[5]   A. S. Schröder, E. Parsa, K. Iwan, F. R. Traube, M. Wallner, S. Serdjukow, T. Carell, *Chem. Commun.* **2016**, *52*, 14361-14364.
[6]   A. S. Schröder, O. Kotljarova, E. Parsa, K. Iwan, N. Raddaoui, T. Carell, *Org. Lett.* **2016**, *18*, 4368-4371.