COMMUNICATION

Profiling the nucleobase and structure selectivity of anticancer drugs and other DNA alkylating agents by RNA sequencing
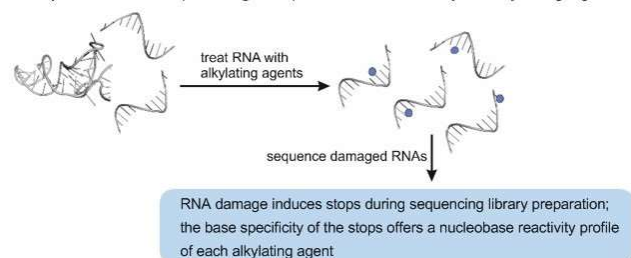
Basilius Sauter,[a] and Dennis Gillingham*[a]

Abstract: Drugs that covalently modify DNA are components of most chemotherapy regimens, often serving as first-line treatments. Classically, the reactivity and selectivity of DNA alkylating agents has been determined *in vitro* with short oligonucleotides. A statistically sound analysis of sequence preferences of alkylating agents is untenable with serial analysis methods because of the combinatorial explosion of sequence possibilities. Next-generation sequencing (NGS) is ideally suited for the broad characterization of sequence or structure selectivities because it analyzes many sequences at once. Herein, we use NGS to report on the chemoselectivity of alkylating agents on RNA and apply this technology to the previously uncharacterized alkylating agent trimethylsilyl diazomethane.

DNA alkylating agents are extremely toxic molecules, and yet they remain a cornerstone of chemotherapy.[1] In glioblastoma multiforme, for example, the alkylating agent temozolomide (**TMZ**)[2] is the only approved drug. One of the world's first anticancer agents was a sulfur mustard,[3] but many other DNA alkylating molecules[4] have been discovered either by sifting through natural products that bear electrophiles, or by focused medicinal chemistry efforts on electrophilic scaffolds with a known predilection for nucleic acid modification.[5] While alkylating DNA to cause apoptosis is their main mode of action, these alkylating agents are indiscriminately reactive, leading to many side effects. One of the off-target substrates is the more abundant RNA, but whether this contributes to toxicity is unknown. Damage to mRNA is mostly short-lived but can lead to longer-lasting effects if they cause errors during translation,[6] or if the damage causes a catalytic gain-of-function.[6e] With these findings and the discovery of RNA repair proteins,[7] the significance of RNA damage needs to be re-evaluated.[6a, 6b, 8] Since all of the alkylating drugs on the market were discovered before the advent of next generation sequencing (NGS), studies of their chemical reactivity on nucleic acids were necessarily serial, limiting analyses to a small set of synthetic oligodeoxynucleotides or to low-throughput biochemical assays such as ligation-mediated PCR.[9] Nucleobase selectivity was typically characterized by reactions with mononucleotides, which cannot report on sequence context or structure specificity. NGS would offer a more nuanced and statistically powerful profiling of alkylating agents, but this has gone largely untested. A couple of recent NGS approaches that look at sites of cisplatin reactivity[10] in both DNA and RNA are the lone exceptions. Other reports have looked at whether there might be natural or designed RNA aptamer sequences that can predispose an RNA to alkylation by organic electrophiles.[11] How natural post-transcriptional modifications influence sequencing datasets has recently been explored.[12] The techniques most similar in workflow to what we describe here are SHAPE sequencing and other types of high-throughput chemical RNA probing.[13] Whereas, however, chemical probing looks at characterizing RNA secondary structure or protein binding sites, we hope to offer a profile of the chemicals themselves: the types of bases they react with and in what ratios, as well as whether they prefer specific consensus sequences or specific types of structures.

Figure 1. Profiling the selectivity of nucleic acid damaging agents by RNA sequencing

To develop the protocol, we choose to characterize two known drugs that operate by a similar mechanism and have a well-established chemical reactivity profile in DNA. **TMZ**[2b] and streptozotocin (**SZ**) (see Figure 1B for structures)[14] are both purported to generate methyldiazonium upon nucleophilic activation and this powerful electrophile[15] then preferentially targets G, modifying both its N7 and O^6 positions.[16] After establishing the protocol with these two molecules, we implemented the method to profile the reactivity of trimethylsilyldiazomethane (**TMS-DAM**) and dimethylsulfate (**DMS**). The reactivity of **TMS-DAM** with DNA and RNA is unstudied, but in principle should show similar reactivity as **TMZ** and **SZ** since it also generates a methyldiazonium in protic solvents.[17] **DMS**, on the other hand, is frequently used as a reagent in chemical RNA probing and in fact was one of the key reagents for the original Maxam-Gilbert sequencing.[18] Indeed, our sequencing data jibes with the observed chemical reactivity profile in Maxam-Gilbert sequencing. The advantage of NGS is

[a]   Basilius Sauter, Prof. Dr. Dennis Gillingham
        Department of Chemistry
        University of Basel
        St. Johanns-Ring 19
        4056 Basel
        E-mail: dennis.gillingham@unibas.ch

        Supporting information for this article is given via a link at the end of the document.

that new alkylating agents with unknown reactivity can be quickly profiled across a broad set of sequences in a single experiment. We started by developing the sequencing method and data analysis pipeline to identify sites of alkylation. Over the past few years classical RNA chemical profiling methods have been adapted for whole transcriptome NGS.[13] Data analysis approaches vary, but the basic concept in these techniques is the same: RNAs are chemically modified and then reverse transcribed (RT) to complementary DNAs (cDNAs). The chemically modified positions prematurely halt RT or cause mutations during read-through; modifications in the original RNA can then be inferred from premature stops or mutations in the sequencing reads. While these techniques were developed to study epigenetics and RNA structure, our goal was to adapt them for broad characterization of the chemo- and site-selectivity of nucleic acid alkylating agents.
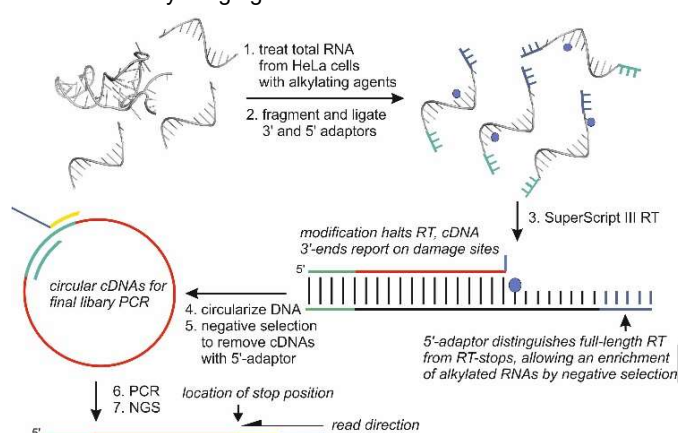

Figure 2. Sequencing library preparation protocol for profiling RNA damage selectivity.

Our protocol is similar to those used in previous reports in RNA structure based profiling,[13] but most closely resembles the Mod-Seq protocol reported by the McManus lab.[13c] The biggest changes we've made to the McManus protocol are that we use STAR[19] for genome alignment and that we use an alkaline fragmentation protocol (more on this below). In brief, we began by treating HeLa cells or HeLa whole transcriptome isolates with a variety of alkylating agents at high concentration (Figure 2, step 1). The resulting RNA pools were then fragmented, and adaptors were ligated at both ends (step 2). The 3' adaptor is a universal PCR primer, while the 5' adaptor is meant to serve as a stop-marker downstream in the bioinformatics analysis. An enrichment step with magnetic beads bearing a complement to the 5' adaptor insures that only successful ligation products proceed to the next step. RT then creates cDNAs of two types; RTs that stop prematurely will give cDNAs that lack a complement to the originally ligated 5' stop-marker, while those that copy normally will have this sequence at their 3' ends. The cDNAs are then circularized and the sequences that bear the stop-marker adaptor are removed by a magnetic bead negative selection (the beads contain an oligonucleotide complementary to the cDNA stop marker). The remaining cDNAs are then PCR amplified with primers that bear the Illumina sequences, purified with AMPure beads to minimize empty primer reads and then submitted for NGS sequencing.

As already mentioned, the biggest difference between our method and SHAPE sequencing is that we use alkaline fragmentation. Although $ZnCl_2$ fragmentation was also an option,[13d] we opted for the alkaline method for chemical profiling since it reports on total modification (since N7G alkylations are depurinated under alkaline conditions and hence turn up as stops in the data), while the $ZnCl_2$ method (which does not efficiently depurinate) reports more on the modifications in the Watson-Crick face. Before doing any data filtering we first wanted to analyze whether the drug treated samples showed any consistent changes in stop positions across all stops. Indeed, in the treated samples we saw a substantial increase in the number of stops at guanine residues. We then reanalyzed the stops after applying a significance test filtering (Cochran-Mantel-Haenszel testing according to the method previously described).[13c] We first analyzed the stop-signal for the four alkylating agents shown in Figure 1B both *in vitro* and in live cells. For all but **DMS** this is the first time that these DNA alkylating agents have been characterized with NGS. As expected, they all showed a preference for guanine alkylation (signals are determined by comparison with matched DMSO controls). None of these purportedly unspecific alkylating agents show a sequence logo preference (see the ESI), consistent with their generating a potent but unselective alkylating agent. The alkylating profiles in live cells match closely those seen with transcriptome isolates except for **TMZ**. **TMZ** demonstrated extremely poor solubility in the culture medium, we believe this leads to a low effective treatment concentration and hence the stop-signals with **TMZ** are indistinguishable from the untreated controls. We also independently measured each base's propensity for RNA alkylation with **DMS** by reacting the monoribonucleotides with **DMS** and then examining the product ratios by UPLC-MS; this method gives conversions of 86, 12, 6, and 0 for G, A, C, and U respectively (see ESI figure S11 for data), exactly as the reactivity order predicted by sequencing based profiling *in vitro* (see **DMS** signals in panels B of Figure 3). These profiles are also consistent with earlier work on DNA modification[16, 20] with **DMS**, which suggests that the order of reactivity should be G>A>C>T in an approximately 78:20:2:0 ratio.

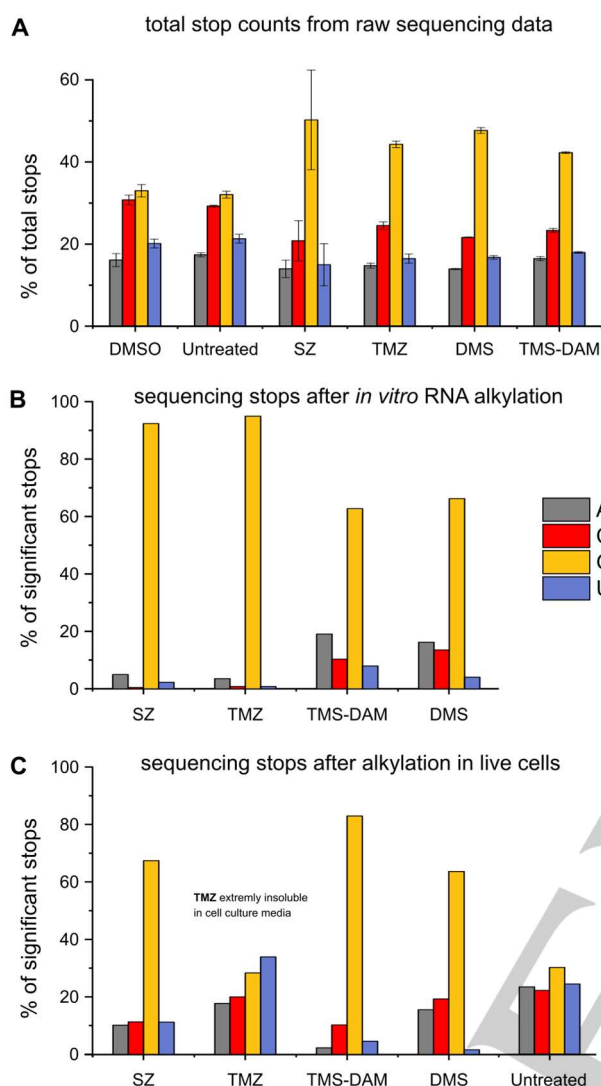**Figure 3.** (**A**) Crude sequencing data indicates a significant preference for guanine stops in comparison to untreated controls; (**B**) Applying the profiling sequencing workflow on the *in vitro* data delivers stops that show a strong preference for G; (**C**) Alkylation reactions in cell culture analysed by the chemical profiling method show consistent profiles except TMZ, which is highly insoluble in the culture medium (note: analyses in **B** and **C** excluded ribosomal RNA (rRNA) since they are the most populous and any selectivity inherent to rRNA would therefore dominate the data).

In structural probing tertiary structures are approximated by calculating a reactivity of each nucleobase position.[10d] In Figure 4 we focus on the 18S rRNA to examine in detail structural specificities of alkylation. In Figure 4A we see that **TMS-DAM** would be a poor structural probing reagent because it can target both double and single-stranded regions. On the other hand, if we look at the A and C stops for **DMS**, we see a strong preference for single stranded regions, a feature that makes **DMS** an excellent chemical probing reagent. The more broad reactivity of **TMS-DAM**, however, could be more advantageous for probing intermolecular interactions. Indeed if we look at the live cell results for both **TMS-DAM** and **DMS** we see that modification at double-

stranded regions is completely lost. Taken together, these analyses validate that NGS is an effective way to profile the structural selectivity of DNA and RNA damaging agents. As shown in Figure 4B, only guanine is alkylated in double-stranded regions in 18S rRNA in the live cell experiments. This unexpected observation could stem from several factors; the most interesting would be that this reactivity difference is reporting on tertiary or quaternary interactions of the 18S rRNA. Going forward we would like to increase sequencing depths so that we can search for putative sites of intermolecular interactions with statistical confidence, and then compare putative interactions with other recently described probing methods.[13f, 21]
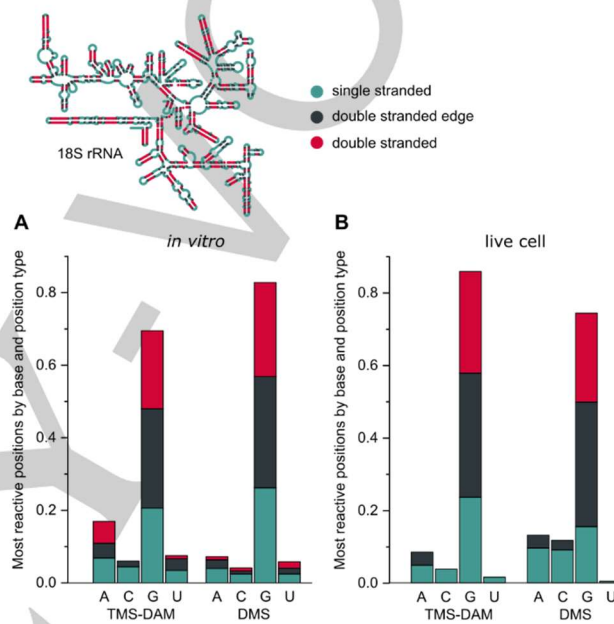

Figure 4. Selectivity of alkylation for different structural motifs in 18s rRNA in vitro (**A**) and in live cells (**B**).

A disadvantage of the technique is that it is blind to modifications that cause neither RT stops nor frequent mutations. **TMS-DAM** offers an example of this shortcoming. Alkylation reactions with diazo electrophiles require a free proton to activate the diazo to a highly electrophilic diazonium.[22] Guanine and uracil have similar pK_as and they are the first ionisable protons in oligoribonucleotides,[23] which should give them similar reactivity with **TMS-DAM**. We tested the reactivity of all four RNA bases with **TMS-DAM** as mononucleosides or in HeLa total transcriptome RNA (Figure 5A). In both cases the major product was the m^3U alkylation product followed by m^6G and m^1G. The Kool lab has recently shown that m^3U is relatively well-tolerated during RT, making it difficult to detect in NGS datasets.[12] Our data is consistent with this conclusion since although there is substantial uridine modification according to chemical analysis, the NGS data show predominately guanine alkylation. With this in mind, a protocol for characterizing new alkylating agents would be to examine the *in vitro* specificity in mononucleotides and then use NGS to look for sequence or structural selectivity across many different sequences.
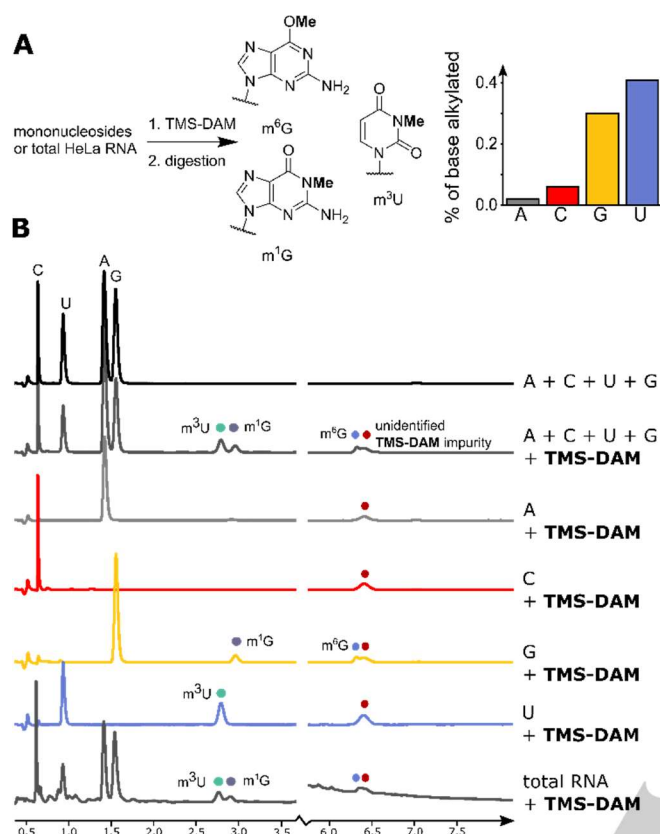
A

B

Figure 5. A. Alkylation reactions with TMS-DAM were carried out on both monoribonucleosides and total HeLa RNA. In the case of total RNA a digestion was carried out to generate monoribonucleosides for analysis; the bar graph shows the relative reactivity of each base; **B**. Crude HPLC traces show that A and C are not modified, while G gives two products (m^1G and m^6G) and U gives a single product (m^3U).

We've shown that RNA sequencing can be used to chemically profile alkylating anti-cancer agents or any other molecule that can covalently modify RNA and gives RT stops or mutations during cDNA synthesis. To develop the method, we first chose electrophiles with a well-established reactivity profile in DNA, and then profiled a substrate that has never been analysed in RNA alkylation (**TMS-DAM**) to show how the technique could be used in practice. To get a high enough signal-to-noise ratio, we needed to use high drug concentration. While this is an effective strategy for chemical profiling and RNA probing, it might obscure valuable additional information such as whether there are any sequence or transcript preferences. In the future we will include an enrichment step in reactions with modified electrophiles, allowing us to use reduce drug concentrations to levels that would not oversaturate any inherent drug selectivity and will help us identify whether alkylating drugs damage specific transcripts or consensus sequences. Already in its present form, however, the technique is an efficient way to profile the nucleobase and tertiary structural selectivity of DNA and RNA damaging agents.

Acknowledgements

[1]   A. P. Francisco, M. d. J. Perry, R. Moreira, E. Mendes, in *Anticancer Therapeutics*, John Wiley & Sons, Ltd, **2008**, pp. 133-158.
[2]   (a) Z. Jihong, F. G. S. Malcolm, D. B. Tracey, *Curr. Mol. Pharm.* **2012**, *5*, 102-114; (b) E. Newlands, M. Stevens, S. Wedge, R. Wheelhouse, C. Brock, *Cancer Treat. Rev.* **1997**, *23*, 35-61.
[3]   F. E. Adair, H. J. Bagg, *Ann. Surg.* **1931**, *93*, 190-199.
[4]   A. Haddow, G. A. R. Kon, W. C. J. Ross, *Nature* **1948**, *162*, 824.
[5]   (a) W. C. Tse, D. L. Boger, *Chem. Biol.* **2004**, *11*, 1607-1617; (b) S. E. Wolkenberg, D. L. Boger, *Chem. Rev.* **2002**, *102*, 2477-2496.
[6]   (a) C. L. Simms, H. S. Zaher, *Cell. Mol. Life Sci.* **2016**, *73*, 3639-3653; (b) B. H. Hudson, H. S. Zaher, *RNA* **2015**, *21*, 1648-1659; (c) J. R. Brickner, J. M. Soll, P. M. Lombardi, C. B. Vågbø, M. C. Mudge, C. Oyeniran, R. Rabe, J. Jackson, M. E. Sullender, E. Blazosky, A. K. Byrum, Y. Zhao, M. A. Corbett, J. Gécz, M. Field, A. Vindigni, G. Slupphaug, C. Wolberger, N. Mosammaparast, *Nature* **2017**, *551*, 389; (d) H. S. Pettersen, T. Visnes, C. B. Vågbø, E. K. Svaasand, B. Doseth, G. Slupphaug, B. Kavli, H. E. Krokan, *Nucleic Acids Res.* **2011**, *39*, 8430-8444; (e) J.-X. Wang, J. Gao, S.-L. Ding, K. Wang, J.-Q. Jiao, Y. Wang, T. Sun, L.-Y. Zhou, B. Long, X.-J. Zhang, Q. Li, J.-P. Liu, C. Feng, J. Liu, Y. Gong, Z. Zhou, P.-F. Li, *Mol. Cell* **2015**, *59*, 50-61.
[7]   (a) G. Zheng, John A. Dahl, Y. Niu, P. Fedorcsak, C.-M. Huang, Charles J. Li, Cathrine B. Vågbø, Y. Shi, W.-L. Wang, S.-H. Song, Z. Lu, Ralph P. G. Bosmans, Q. Dai, Y.-J. Hao, X. Yang, W.-M. Zhao, W.-M. Tong, X.-J. Wang, F. Bogdan, K. Furu, Y. Fu, G. Jia, X. Zhao, J. Liu, Hans E. Krokan, A. Klungland, Y.-G. Yang, C. He, *Mol. Cell* **2013**, *49*, 18-29; (b) M. Westbye, E. Feyzi, P. Aas, C. Vågbø, V. Talstad, B. Kavli, L. Hagen, O. Sundheim, M. Akbari, N.-B. Liabakk, G. Slupphaug, M. Otterlei, H. Krokan, *J. Biol. Chem.* **2008**, *283*, 25046-25056; (c) F. Drabløs, E. Feyzi, P. A. Aas, C. B. Vaagbø, B. Kavli, M. S. Bratlie, J. Peña-Diaz, M. Otterlei, G. Slupphaug, H. E. Krokan, *DNA Repair* **2004**, *3*, 1389-1407; (d) P. Aas, M. Otterlei, P. Falnes, C. Vågbø, F. Skorpen, M. Akbari, O. Sundheim, M. Bjørås, G. Slupphaug, E. Seeberg, H. Krokan, *Nature* **2003**, *421*, 859-863.
[8]   C. L. Simms, B. H. Hudson, J. W. Mosior, A. S. Rangwala, H. S. Zaher, *Cell Reports* **2014**, *9*, 1256-1264.
[9]   G. P. Pfeifer, R. Drouin, G. P. Holmquist, *Mutat. Res.* **1993**, *288*, 39-46.
[10]   (a) X. Shu, X. Xiong, J. Song, C. He, C. Yi, *Angew. Chem. Int. Ed.* **2016**, *55*, 14246-14249; (b) J. Hu, J. D. Lieb, A. Sancar, S. Adar, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11507-11512; (c) K. Plakos, V. J. DeRose, *Chem. Commun.* **2017**, *53*, 12746-12749.
[11]   (a) R. I. McDonald, J. P. Guilinger, S. Mukherji, E. A. Curtis, W. I. Lee, D. R. Liu, *Nat. Chem. Biol.* **2014**, *10*, 1049-1054; (b) A. K. Sharma, J. J. Plant, A. E. Rangel, K. N. Meek, A. J. Anamisis, J. Hollien, J. M. Heemstra, *ACS Chem. Biol.* **2014**, *9*, 1680-1684.
[12]   A. M. Kietrys, W. A. Velema, E. T. Kool, *J. Am. Chem. Soc.* **2017**, *139*, 17074-17081.
[13]   (a) D. Loughrey, K. E. Watters, A. H. Settle, J. B. Lucks, *Nucleic Acids Res.* **2014**, *42*, 000; (b) J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, A. P. Arkin, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 11063-11068; (c) J. Talkish, G. May, Y. Lin, J. L. Woolford, C. J. McManus, *RNA* **2014**, *20*, 713-720; (d) Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, S. M. Assmann, *Nature* **2014**, *505*, 696-700; (e) C. K. Kwok, Y. Ding, Y. Tang, S. M. Assmann, P. C. Bevilacqua, *Nat. Commun.* **2013**, *4*, 2971; (f) R. C. Spitale, R. A.

Flynn, Q. C. Zhang, P. Crisalli, B. Lee, J.-W. Jung, H. Y. Kuchelmeister, P. J. Batista, E. A. Torre, E. T. Kool, H. Y. Chang, *Nature* **2015**, *519*, 486-490; (g) R. C. Spitale, P. Crisalli, R. A. Flynn, E. A. Torre, E. T. Kool, H. Y. Chang, *Nat. Chem. Biol.* **2013**, *9*, 18-20; (h) L. J. Kielpinski, J. Vinther, *Nucleic Acids Res.* **2014**, *42*, e70.

[14] A. D. Bolzán, M. S. Bianchi, *Mutat. Res.-Rev. Mutat.* **2002**, *512*, 121-134.

[15] D. Gillingham, S. Geigle, O. A. von Lilienfeld, *Chem. Soc. Rev.* **2016**, *45*, 2637-2655.

[16] D. T. Beranek, *Mutat. Res.* **1990**, *231*, 11-30.

[17] E. Kühnel, D. D. P. Laffan, G. C. Lloyd-Jones, T. Martínez del Campo, I. R. Shepperson, J. L. Slaughter, *Angew. Chem. Int. Ed.* **2007**, *46*, 7075-7078.

[18] A. M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 560-564.

[19] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, *Bioinformatics* **2013**, *29*, 15-21.

[20] D. T. Beranek, C. C. Weis, D. H. Swenson, *Carcinogenesis* **1980**, *1*, 595-606.

[21] K. E. Watters, T. R. Abbott, J. B. Lucks, *Nucleic Acids Res.* **2016**, *44*, e12.

[22] N. Fei, B. Sauter, D. Gillingham, *Chem. Commun.* **2016**, *52*, 7501-7504.

[23] J. C. González-Olvera, J. Martínez-Reyes, E. González-Jasso, R. C. Pless, *Biophys. Chem.* **2015**, *206*, 58-65.

Entry for the Table of Contents (Please choose one layout)

Layout 1:

COMMUNICATION

Text for Table of Contents	((Insert TOC Graphic here))	*Author(s), Corresponding Author(s)** ***Page No. – Page No.*** **Title**

Layout 2:

COMMUNICATION
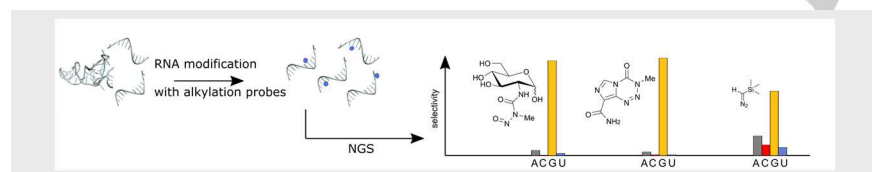
Basilius Sauter, Dennis Gillingham*

Page No. – Page No.

Title

Drugs that covalently modify DNA are components of most chemotherapy regimens, often serving as first-line treatments. Classically, the reactivity and selectivity of DNA alkylating agents has been determined *in vitro* with short oligonucleotides. Here we use next generation sequencing to report on the chemoselectivity of alkylating agents on RNA and apply this technology to the previously uncharacterized alkylating agent trimethylsilyl diazomethane.