Chem Soc Rev

REVIEW ARTICLE

Properties and reactivity of nucleic acids relevant to epigenomics, transcriptomics, and therapeutics

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Dennis Gillingham,[a*]† Stefanie Geigle[a], and Anatole von Lilienfeld[b]

Developments in epigenomics, toxicology, and therapeutic nucleic acids all rely on a precise understanding of nucleic acid properties and chemical reactivity. In this review we discuss the properties and chemical reactivity of each nucleobase and attempt to provide some general principles for nucleic acid targeting or engineering. For adenine-thymine and guanine-cytosine base pairs, we review recent quantum chemical estimates of their Watson-Crick interaction energy, π-π stacking energies, as well as the nuclear quantum effects on tautomerism. Reactions that target nucleobases have been crucial in the development of new sequencing technologies and we believe further developments in nucleic acid chemistry will be required to deconstruct the enormously complex transcriptome.

Contents

Table of contents will be included after final revisions

1 Introduction

Nature uses chemical modification to regulate function in nucleic acids. For instance, dynamic methylations of CpG islands govern gene expression,[1] and recent evidence suggest that mRNA as well may be regulated by methylation.[2, 3] The toxicity of exogenous and endogenous electrophiles often derives from nucleic acid modification and studies to elucidate the repair pathways of DNA damage were awarded the 2015 Nobel Prize.[4] Nucleic acid engineering for research[5] or therapeutic purposes[6, 7] relies on chemical modification. Many natural modifications in DNA and RNA have been identified, but in most cases these are not yet understood.[8] For all of these functions DNA or RNA polymers are the substrates in chemical reactions. Here we will examine cases where DNA or RNA is modified in unusual ways and the impact of these changes on its function. In preparation for these discussions the early sections will outline some basic properties of the polymer which will help us rationalize its chemical reactivity. In the later sections we will describe how nucleic acid modification chemistry has been vital in the recent development of epigenomics and transcriptomics. The review will end with a brief discussion of where the field is likely headed over the next decade.

2 Some physical properties of the nucleic acid polymer

To understand where and how nucleic acids react it is important to know the key physical properties of the polymer.

2.1 Basicity of nucleobases

The first proton associations in nucleobases of adenosine monophosphate (AMP) and CMP occur between pH 3.5-4.2, other bases protonate closer to pH 2.[9] Although CMP is a tad more basic than AMP (pK_a of 4.2 versus 3.5),[10] the small difference means that in a larger DNA or RNA subtle effects such as sequence context can switch the order of basicity. In RNA pH values below 5 lead to decomposition (in the absence of other stabilizing effects) but, as we will see in section 2.3, protonation of RNA bases is still important to consider. The main message is that if one is looking at candidate sites for protonation on a particular nucleic acid, then adenine N[1] and cytosine N[3] are the likely targets.

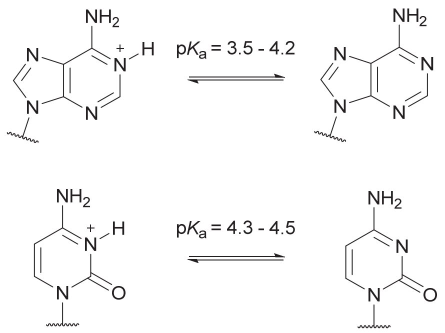


Fig. 1 First proton associations in nucleic acids occur at N^1 A and N^3 C.

2.2 Acidity of nucleobases

[a] Department of Chemistry, University of Basel, St. Johanns-Ring 19, Basel, CH-4056, Switzerland

[b] Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland

† To whom correspondence should be addressed: dennis.gillingham@unibas.ch

The first proton dissociations in DNA and RNA occur at guanine, thymidine, uracil, and inosine. The rank order of acidity seems to be inosine > guanine ≈ uracil > thymidine; although again the differences are small enough that sequence context and secondary structure could invert the acidity order. The second proton association and dissociation pK_as are likely irrelevant since these lie in unrealistic pH regimes.

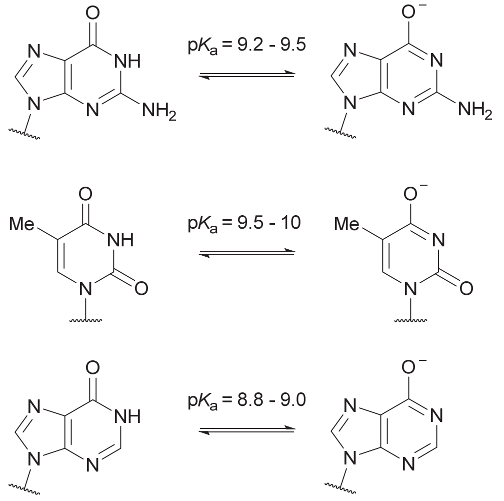


Fig. 2 First proton dissociations in nucleic acids occur at lactam-like N-H protons

2.3 pK_a perturbation depending on microenvironment

pK_as of amino acid side-chains in proteins can vary over many pH units depending on their environment.[11] Enzymes often take advantage of this property to control reactivity in their active sites. A striking example of pK_a shifting in proteins is the light-driven proton pump bacteriorhodopsin.[12] Conformational changes induced by photoisomerization of the retinal-Schiff base result in pK_a shifts (*i.e.* ΔpK_a) of between -1.8 to 8 units in the residues involved in proton transport. While pK_a modulation is well-established in proteins its importance in nucleic acids has only recently been appreciated.

HDV ribozyme. Watson-Crick base-pairing shifts pK_as of bases further away from neutrality;[10] hence in duplex DNA nucleobases are uncharged. This feature is important for the stability of the genome, where structural imperfections that result from proton gain or loss might lead to replication errors. In fact, it has been proposed that a preference for neutrality might have been one of the evolutionary selection criteria for the present set of genomic bases.[13] RNA is different. RNA exists as unstructured single-strands, or as complex folded motifs that contain both single- and double-stranded regions.[5] A variety of other secondary and tertiary structural elements are also possible. Folded structures in RNA can create protein-like pockets, increasing the probability of observing pK_a shifts. Furthermore, non-Watson-Crick base-pairing (wobble, Hoogsteen) can shift pK_as toward neutrality.[10] Unlike proteins, RNAs rely largely on metal cofactors[14, 15] or general acid-base chemistry to achieve catalysis.[16] The hepatitis delta virus (HBV) ribozyme,[17] for example, has a catalytically essential cytosine whose pK_a is shifted from the normal value of 4.1 in cytidine monophosphate to 6.4 in the active site of the HDV ribozyme.[18] The shifted pK_a places this cytidine in the same range as histidine residues, the most common general base in enzyme catalysis. Indeed the mechanistic picture that has emerged in the HDV ribozyme indicates an essential role for C75 (Fig. 3); Raman spectra of HDV crystals were used to obtain spectroscopic information on which vibrational modes were involved in the reaction. Protonation of C75 seems to play an essential role in the transformation.

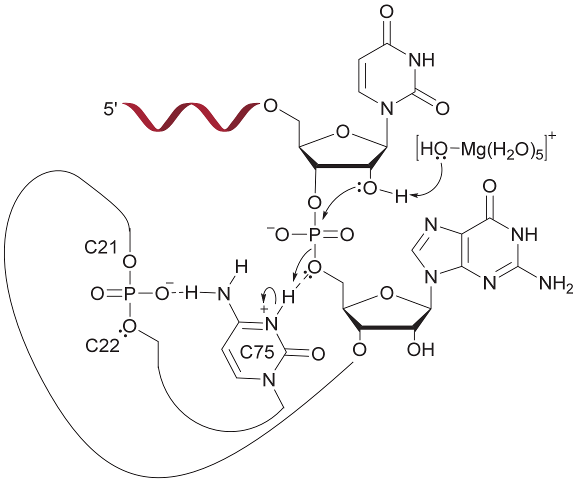


Fig. 3 Proposed transition state for HDV ribozyme involves a protonated cytosine

***i*-motif.** The *i*-motif fold contains intercalated non-Watson-Crick C-C$^+$ (C$^+$ = N^3 protonated cytosine) hemi-protonated base pairs. The human telomeric repeat sequence (CCCTAA)$_n$, when folded into an *i*-motif, contains six such base pairs and has an optimal stability *in vitro* of between pH 5.5 – 5.8.[19] The *i*-motif found in the promoter region of the oncogene Bcl-2 has a pK_a of 6.6.[20] When measured under molecular crowding conditions that mimic the intracellular environment cytosine basicity in *i*-motifs can be further increased to neutrality or above.[21] Although validated biological roles for the *i*-motif are rare and controversial,[22] recent evidence indicates that kinetic rather than thermodynamic factors may be the mechanism by which *i*-motifs exert their function.[19, 23, 24]

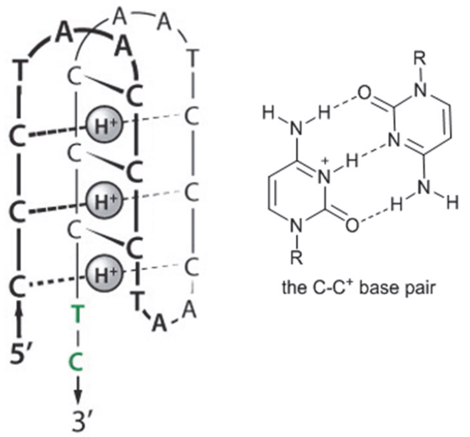


Fig. 4 The structure of the *i*-motif and the hemiprotonated C-C$^+$ base pair (adapted with permission from: Guillaume Mata; Nathan W. Luedtke; *J. Am. Chem. Soc.* **2015**, *137*, 699-707. DOI: 10.1021/ja508741u. Copyright © 2014 American Chemical Society).

Such unusual pK_a shifting as that seen in the cases outlined above in both DNA and RNA should serve as a note of caution that all instances of a specific base in a larger polymer are not necessary chemically uniform. Context and secondary structure can alter normal physical properties and open the door to unusual modes of reactivity.

2.4 Interaction Energies

Nucleobases in DNA typically interact with each other through (i) Watson-Crick (WC) interstrand hydrogen-bonding, and (ii) intrastrand π-π-stacking. While the former mode of binding was instrumental in establishing the double-helical structure of DNA,[25] the role of the latter has only recently been quantified.[26-28] Although the effect of the backbone on base-pair binding is rarely considered, recent calculations suggest that it too can have a significant influence on DNA and RNA conformations.[29]

Quantum chemistry estimates of the potential energy of interaction for WC binding amount to 15.4 and 28.8 kcal/mol for AT and GC, respectively.[30] These predictions were obtained using MP2 in the complete basis-set limit, augmented by a ΔCCSD(T) correction. More recent calculations from the same laboratory indicate a value of 28.5 kcal/mol for GC.[31] Dispersion corrected density functional theory estimates are in very good agreement, ranging from -14.2 to -15.7 kcal/mol and -27.7 to -30.5 kcal/mol for AT and GC, respectively.[32, 33] π-π stacking affords weaker, yet considerable bonding. Again, using MP2 in the complete basis-set limit, augmented by a ΔCCSD(T) correction, post-Hartree Fock method quality numbers have been obtained: They amount to -11.6 and -16.9 kcal/mol for stacked AT and GC base-pairs, respectively.[30] Corresponding dispersion corrected density functional theory estimates range from -9.5 to -11.2 kcal/mol and -14.6 to -17.8 kcal/mol for AT and GC, respectively.[32, 33] It is less obvious to assign high precision single numbers to larger complexes containing nucleic acids. Careful symmetry adapted perturbation theory based analysis, however, suggests that the dominant respective contribution of dispersion and electrostatic derived forces in DNA steps does alternate when the twisting angle is varied.[34] In light of all of this we consider accuracies corresponding to state-of-the art quantum chemistry a necessity in order to achieve truly predictive estimates of the decisive energetic features in DNA and RNA.

2.5 Electronic Properties

According to dispersion corrected DFT (BLYP_DCACP[35]), the norm of the respective dipole moments of A, T, G, and C amount to roughly 2.4, 4.4, 6.6, and 6.5 Debye.[33] Calculations at the Hartree-Fock level of theory seem to consistently overestimate nucleobase dipole-moments.[36, 37] A comprehensive study on the role of basis-set (ADZP, ATZP, AQZP), as well as electronic approximations within a single reference framework for Hartree-Fock and MP2, as well as for generalized gradient approximated DFT (BP86[38]), hybrid DFT (B3LYP[39]) was published in 2009 by Campos and Jorge.[40] Comparing these with other recent studies indicate a recurring theme: Hartree-Fock gives high estimates while other high-level approaches deliver internally consistent values (compare rows in Table 1).[41]

Table 1 A comparison of calculated dipole moments of canonical bases at different levels of theory

	HF	MP2	BP86	B3LYP	MP4	BLYP_ DCACP	CAM-B3LYP[42]
A	2.7	2.7	2.5	2.6	2.5	2.4	2.4
T	4.9	4.3	4.4	4.5	4.3	4.4	4.5
G	7.1	6.4	6.5	6.6	6.4	6.6	6.7
C	7.2	6.3	6.3	6.6	6.5	6.5	6.8
U	5.1	4.3	4.4	4.6	4.3		4.6

Although one needs to use caution when comparing gas phase calculations to solution phase measurements, the calculated values are nevertheless in agreement with experimental estimates.[43] These estimates have, for example, proven important in rationalizing the reactivity of bases with diazonium electrophiles.[44]

Polarizabilities calculated by Campos and Jorge[40], as well as by Alparone,[41] are in good agreement. As can be seen in Table 2, B3LYP or CAM-B3LYP appear to be in remarkable agreement with the highest level of theory used (MP4). Generalized gradient corrected DFT (BP86), by contrast, systematically overestimates the polarization, as one would expect.[37] We note that Hartree-Fock systematically underestimates the polarizability. Also this trend is not surprising: Typically, the band-gap is too wide when using Hartree-Fock.[26]

Table 2 Comparison of calculated polarizabilities of canonical bases at different levels of theory

	HF	MP2	BP86	B3LYP	MP4	CAM-B3LYP[42]
A	89	98	100	97	96	96
T	77	85	86	83	82	83
G	93	104	107	103	101	102
C	71	79	80	78	78	76
U	65	71	72	70	70	70

The standout in the calculation results is guanine; it has the highest dipole moment and the greatest polarizability. These results would suggest guanine as a common target in modification reactions. As we outline in the next sections these predictions are valid: guanine is the most vulnerable base in nucleic acids and is the primary target of most oxidants and electrophiles that damage nucleic acids.

3 Reactions of nucleobases

3.1 Mechanisms of nucleobase modification

Nucleic acids can serve as nucleophiles, electrophiles, or components in radical reactions – but they never do so easily. In early evolution, before compartmentalization and repair processes would have evolved to protect nucleic acids, a severe selection for chemically robust building materials would have been imposed. The Hadean and late-heavy bombardment period of the earth's history were harsh, unforgiving environments, and it is out of these circumstances that the earliest forms of life seem to have arose. Amongst other environmental insults, the earliest self-replicators likely had to suffer acid rain, cycles of hydration and desiccation, powerful UV radiation, and extremes of temperature. Genomes have to be durable. Nevertheless modification and decay does happen, and these undesirable processes have led to the evolution of DNA repair pathways – Sancar, Modrich, and Lindahl were awarded the 2015 Nobel Prize for their pioneering roles in elucidating some of those pathways. While radiation, oxidation, and hydrolytic damage constitute a large part of DNA damage, these topics have been covered extensively elsewhere. In this review we will focus primarily on nucleic acid alkylation since the role of alkylation of DNA and RNA in epigenetics is an unfolding story, and understanding chemical modification pathways will help in furthering this field. Therapeutic nucleic acids and artificial nucleic acids for biological research are typically heavily chemically modified. Particularly with long nucleic acids, such as mRNAs or long non-coding RNAs, a guidebook for chemical modification will be essential for future research.

Most assessments of nucleophilicity in nucleic acids have relied on qualitative comparisons of product distributions after reactions with different electrophiles. The dominant control element in chemical modification of double-stranded DNA is the shielding effect of the double helix itself.[44] Guanine is the most electron rich base and it is by far the primary target for electrophiles in duplex DNA: N7 is the most nucleophilic site and its position in the major groove makes it accessible to attack. N^2

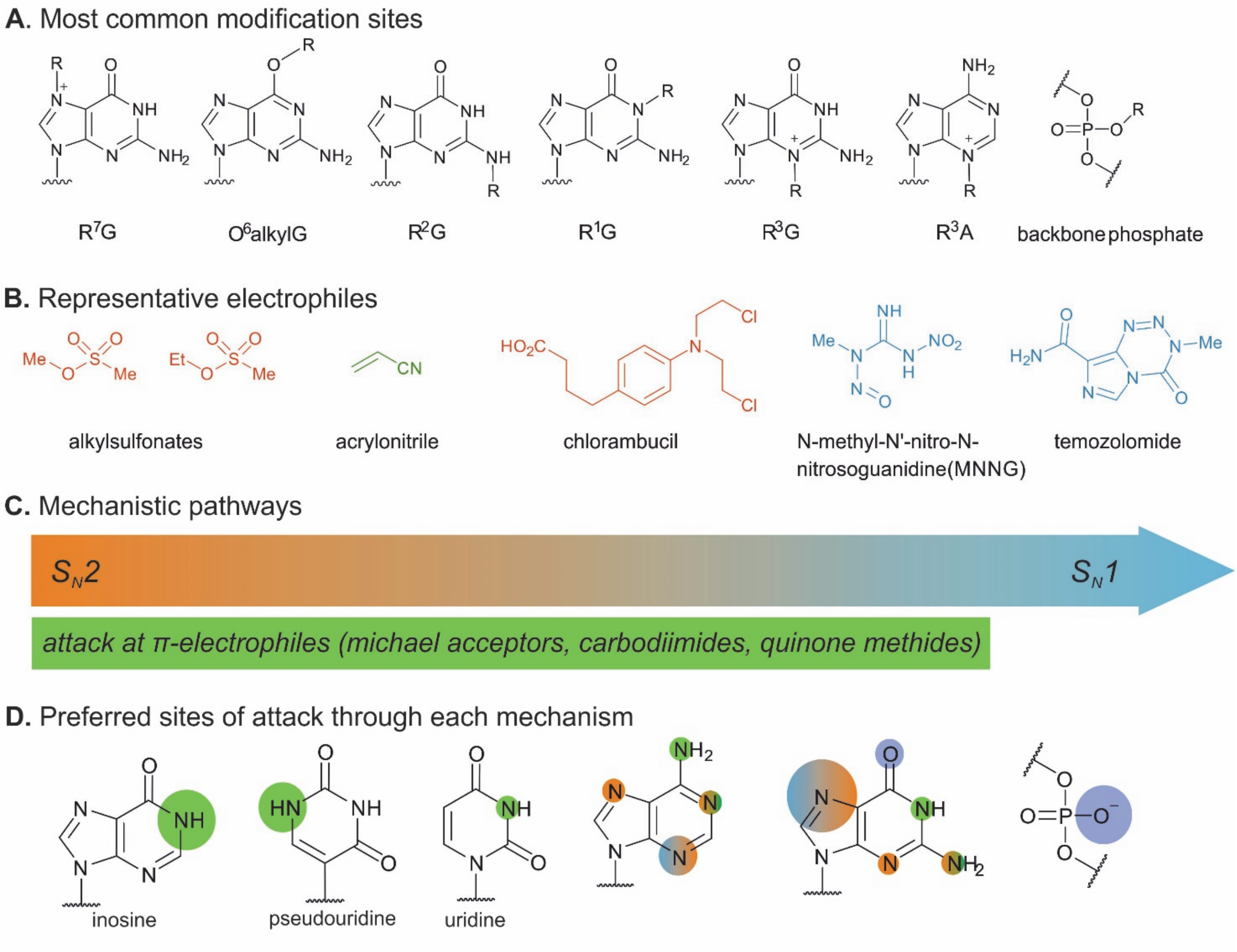

A. Most common modification sites
R^7G
O^6alkylG
R^2G
R^1G
R^3G
R^3A
backbone phosphate
B. Representative electrophiles
alkylsulfonates
acrylonitrile
chlorambucil
N-methyl-N'-nitro-N-nitrosoguanidine(MNNG)
temozolomide
C. Mechanistic pathways
S_N2
S_N1
attack at π-electrophiles (michael acceptors, carbodiimides, quinone methides)
D. Preferred sites of attack through each mechanism
inosine
pseudouridine
uridine

Fig. 5 The most nucleophilic sites in DNA and RNA (**A**), the electrophiles that react with them (**B**), and the mechanistic pathways leading to modification (**C & D**)

juts out in the minor groove and would also, in principle, be available for reaction, but the lack of free space and the spine of hydration in the minor groove limits the reactivity of N^{2}G in B-DNA. Single-stranded DNA and RNA offer a more pure assessment of nucleobase nucleophilicity. The alkylation products that are typically observed when DNA or RNA is treated with electrophiles are shown in Figure 5 panel A. Guanine dominates the alkylation spectra of nucleic acids (particularly in duplexes) while pyrimidines represent a small percentage of alkylated adducts. It would be ideal to have a quantitative assessment of the nucleophilicity of each site in DNA and RNA; Unfortunately the physical organic tools to study nucleophilicity quantitatively with systems that contain more than one nucleophile have only recently been developed,[45] and no one has revisited the nucleic acid question with these tools. Variations on the Swain-Scott treatment[46] have been employed for common DNA damaging electrophiles, but these have always used a proxy nucleophile such as pyridine for ease of analysis.[46] The studies on electrophilicity have shown that as one moves across the spectrum of S$_N$2 to S$_N$1 there are changes in site-selectivity (see panels C and D in Figure 5).[47] The Klopman-Salem[48, 49] distinction between charge-controlled and orbital controlled reactions has been invoked (and seems reasonable) to explain these trends. However, recent evidence suggests that Markus theory might be superior in understanding site-selectivity in reactions with multiple nucleophilic centres.[50] Nevertheless, based only on the empirical results the following trends have been observed:

- The most powerful nucleophilic base with σ-electrophiles is guanine
- S$_N$2-type electrophiles overwhelming favour N7G
- S$_N$1-type electrophiles (*e.g.* diazonium) deliver substantial amounts of N7G, O^{6}G, and backbone phosphate alkylation
- π-electrophiles typically operate under thermodynamic control and target N1 of inosine (e.g. acrylonitrile), the N1 of pseudouridine, N3 of uridine (with carbodiimide electrophiles), or the N^2 of guanine
- Bifunctional electrophiles such as malondialdehyde or acrylaldehyde target nucleobases that contain an aniline-type nitrogen (A, G, C), as these offer proximal nucleophiles that deliver irreversible cyclic adducts

3.2 Natural products that covalently modify nucleobases

Given the essential role of nucleic acids it is unsurprising that organisms have evolved strategies for targeting them. A number of distinct chemical mechanisms have emerged for targeting DNA and RNA, but we will only consider the alkylating natural products here. The reader is referred to the reviews[51-53] for a comprehensive discussion of the different mechanisms such as non-covalent binders,[54] radical cleavage,[55] reactive oxygen species generation.[56] In the following section we will give an overview of different alkylation strategies that have evolved to target nucleic acids. Of particular note is the fact that there are recurring structural motifs or mechanisms that distinguish DNA-targeting molecules. We will look at some natural products that alkylate DNA through six distinct mechanisms: activated cyclopropyl groups (Fig. 6), quinone methides (Fig. 7), three-membered heteroatom rings (Fig. 8), diazo compounds (Fig. 9), reactive iminium ions (Fig. 10), and photochemical cycloaddition (Fig. 11). With covalent modification the challenge is to hold the molecule's reactivity in check until it meets its intended target – Nature has evolved some ingenious solutions.

Duocarmycin A, shown in Fig. 6, is representative of a small family of natural products that contain the dienone cyclopropane motif.[53] This motif is relatively stable until the molecule binds the minor groove of A-T rich strands of DNA.[57] Binding in the minor groove requires some rotation about the amide bond, disrupting conjugation of the vinylogous amide and increasing the electrophilicity of the dienone cyclopropane.[58, 59] Its position next the N3 of adenine causes a proximity-induced alkylation of this heteroatom.[57] There are a number of examples of natural products that bear an enone adjacent to a cyclopropyl group. These behave chemically as a sort of one-carbon homologated enone electrophile. Cyclopropyl σ-bonds are unusually reactive (for σ-bonds) and, when in conjugation with π-systems, their reactivity is extended in a manner akin to the vinylogy concept.[60, 61] Ptaquiloside is another such cyclopropyl-enone bearing natural product (others include Illudin S[62] and myrocin C[63]). Ptaquiloside is the primary toxin found in fern bracken (*Pteridium aquilinum*) and can poison grazing animals.[64, 65] The molecule is also a carcinogen in mammals with a mode-of-action that involves alkylation of N3A and N7G.

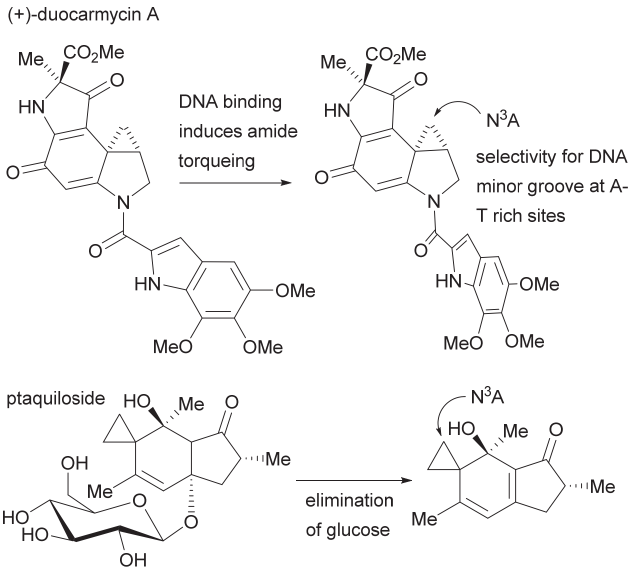


Fig. 6 Natural products that alkylate through a cyclopropyl enone electrophile: Top: Duocarmycin A is activated by the torqueing induced by DNA binding in the minor groove; Bottom: Ptaquiloside is activated by elimination of glucose.

Mitomycin C (see Fig. 7) is a potent cytotoxin in clinical use against a variety of cancers.[66] Its mode of action is DNA alkylation and cross-linking.[67] Reductive activation[68] of mitomycin delivers the hydroquinone, which leads to methanol elimination and an aziridine ring-opening to give an extended quinone-methide electrophile.[69] This activated form binds to the minor groove of DNA and the N^2 of guanine (a good minor groove nucleophile with a propensity for attacking π-electrophiles, see section 3.1), attacks the quinone-methide of activated mitomycin. After the initial attack, elimination of the carbamoyl group then generates an α,β-unsaturated iminium ion poised for a second attack by the N^2 of a different guanine. Both inter-[70, 71] and intramolecular[72] adducts have been observed, but the intermolecular lesion is likely the most toxic. Both the duocarmycins and mitomycins rely on the electrophilicity of a π-electrophilic system to drive their alkylation activity. As we will see later, this type of reactivity has also been exploited by chemists.

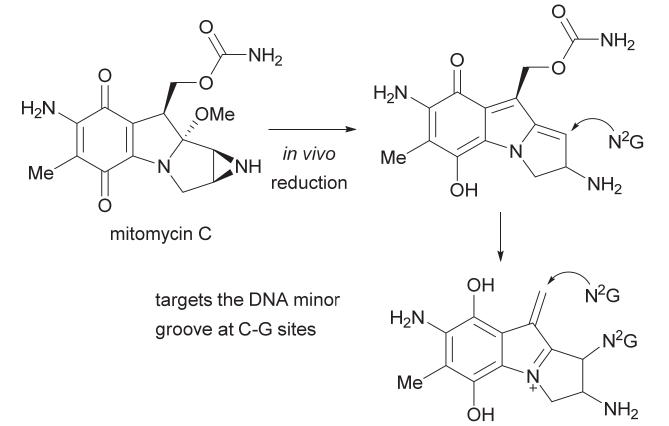


Fig. 7 Mitomycin C is a bifunctional electrophile. Biological reduction delivers a quinone-methide that alkylates N^2G in the minor groove; a subsequent elimination leads to a highly electrophilic α,β-unsaturated iminium ion, which is attacked by another N^2G on the opposing strand.

Three-membered ring heterocycles (epoxides, aziridines, and episulfoniums) are another common motif found in DNA-targeting natural products (see Fig. 8). The fasicularins,[73] for example, exploit a reactive aziridinium ion generated by intramolecular nucleophilic displacement of the thiocyanate group by the neighbouring nitrogen atom.[74] The resulting aziridinium has been shown to alkylate the N7 of guanine. The natural products azinomycin A and B are mechanistically related to fasicularin in the sense that they bear an aziridine-based warhead, but they also have a reactive epoxide in their structure.[75, 76] The combination of two electrophiles confers the azinomycins with the ability to cross-link DNA. They have a specificity for N7G-N7G cross-links with some N7G-N7A observed as well. The primary adduct at N7G is formed with the aziridine, followed by the secondary alkylation two bases downstream on the opposing strand. Perhaps the most toxic epoxide is that derived from aflatoxin B_1. Aflatoxin ingestion from fungal contamination of food supplies is a major cause of liver cancer worldwide, but is particularly acute in countries that have unregulated and poorly developed systems of food storage.[77] Once ingested, aflatoxin is epoxidized in the liver and then intercalates into DNA where the N7G serves as a nucleophile to open the epoxide.[78]

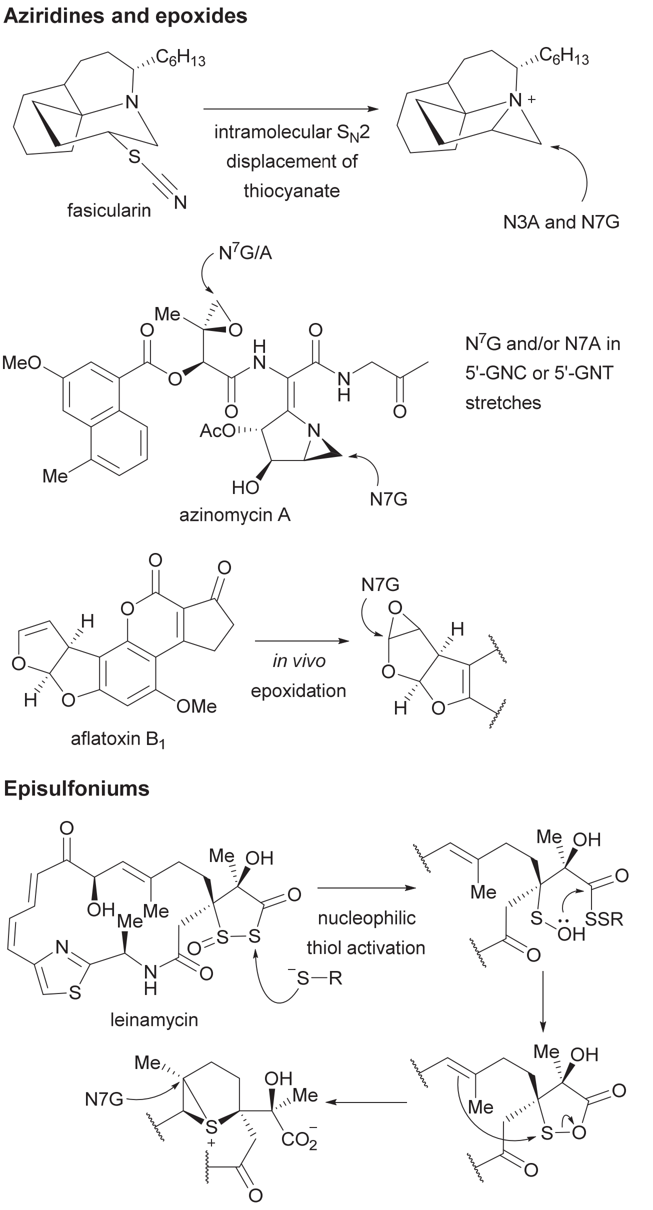


Fig. 8 Three-membered ring heterocycles are another common motif found in DNA alkylating natural products

Leinamycin[79] is a DNA alkylating natural product that operates by creating a transient episulfonium ion (see bottom of Fig. 8). The episulfonium forms from a nucleophilic attack on the unusual 1,2-dithiolan-3-one 1-oxide ring system, which seems unique to this natural product.[80] As shown in Fig. 8, the natural product is activated by nucleophilic attack of a thiol[81] (although water activation is also possible)[82], delivering a reactive dithioester and a sulfenic acid. The sulfenic acid then attacks the ester to form an oxathiolanone. In the next step the nearby olefin serves as a nucleophile to initiate episulfonium formation in a formal chelotropic process. The episulfonium is a potent and selective DNA alkylator with a preference for GG and GT tracts in duplex DNA.[81, 83] The extended π-system in leinamycin seems important in controlling its binding interaction with DNA.[84]

At first glance the natural products that contain diazo seem like unlikely structures. However, diazo compounds stabilized by conjugation into electron withdrawing motifs are quite stable and used extensively in organic synthesis.[85] Nature has produced twenty or so natural products bearing stabilized diazo functional groups.[86, 87] The most prominent of these is the kinamycin and lomaiviticin family of natural products. These, however, seem to operate by radical-induced strand-breaking[88] and hence will not be discussed here. The diazopeptides, exemplified in Fig. 9 by azaserine, are stable and long-lived enough to target intracellular structures. Diazo peptide natural products fell out of clinical study due to toxicity,[89-91] likely as a result of their activity on a number of essential processes. Although the primary mode-of-action of α-diazo carbonyl compounds is believed to be as glutamine antimetabolites, their mutagenic activity may suggest DNA as a secondary target.[92-94] In fact diazoacetate formation is believed to occur frequently in humans as a result of nitrosation of the N-terminus of peptides or glycine itself (found especially in persons with high red meat or cured meat diets).[95] DNA examined after exposure to diazoacetate contained O^6-carboxymethyl-G adducts;[96] hence it seems certain that α-diazo carbonyl compounds target DNA, but determining a precise mechanism and whether this has biological relevance will require more studies.

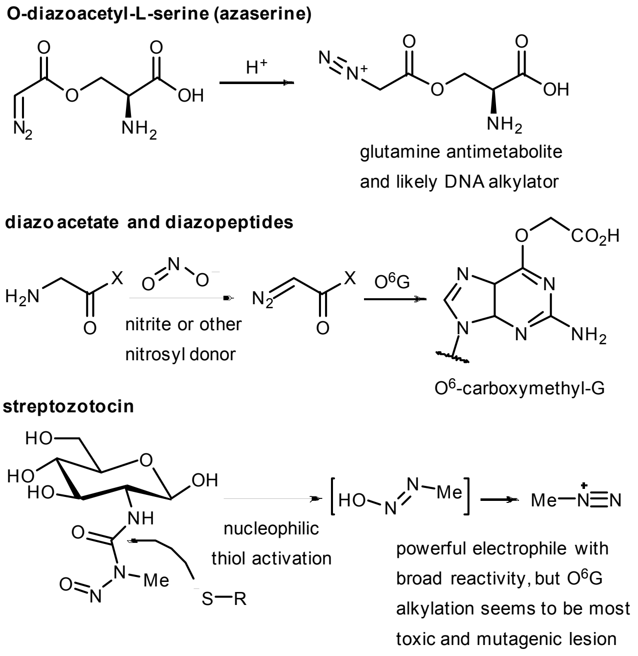

O-diazoacetyl-L-serine (azaserine)
O
O
O
OH
N2
NH2
H+
N
N+
O
O
O
OH
NH2
glutamine antimetabolite and likely DNA alkylator
diazo acetate and diazopeptides
H2N
X
O
O
N
O
nitrite or other nitrosyl donor
N2
X
O
O6G
O
CO2H
N
N
N
N
NH2
O6-carboxymethyl-G
streptozotocin
HO
HO
HO
O
OH
NH
O
N
Me
O
N
nucleophilic thiol activation
S
R
HO
N
N
Me
Me
N
N
powerful electrophile with broad reactivity, but O6G alkylation seems to be most toxic and mutagenic lesion

Fig. 9 Alkylation of DNA by diazo and diazonium groups

Unstabilized diazo compounds have half-lives on the order of a few seconds in water.[97] The natural product streptozotocin (see bottom of Fig. 9)[98] contains an N-nitroso-N-methylurea motif (MNNU). This functional group and the related guanidine analogue is found in a number of man-made cytotoxins and are known precursors to the extremely reactive methyldiazonium electrophile.[99] Alkyldiazonium species alkylate many of the heteroatoms in DNA, but the most mutagenic is the O^{6}G modification.[44] The glucosamine side-chain in streptozotocin is recognized by the glucose transport protein GLUT2,[100-102] which is highly expressed on the surface of pancreatic beta cells. The specific islet cell-targeting of steptozotocin has made it effective in treating pancreatic cancer and also as a chemical inducer of Type 1 diabetes in animal models.[102]

Another class of DNA alkylating natural products are the stable carbinolamines. These generate a reactive iminium ion that targets the N^{2}G in the minor groove upon binding with duplex DNA. In Fig. 10 the concept is shown with Ecteinascidin 743[103] but other molecules that seem to operate by this mechanism include the anthramycins[104] and the saframycins.[105] These are unique because they form a reversible N^{2}G adduct that is only stable in duplex DNA, hence replication forks and RNA are not affected.

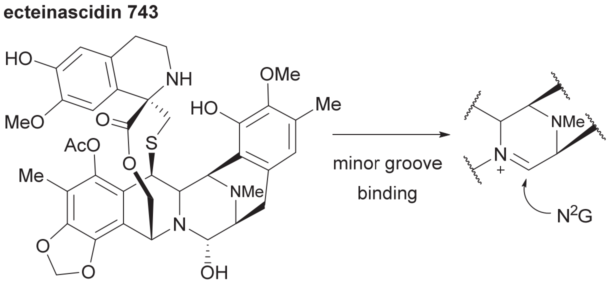


Fig. 10 Ecteinascidin 743 is representative of a group of natural products that generate a reactive iminium species that alkylates N^{2}G upon binding to the minor groove of duplex DNA

Finally, the furocoumarins operate by photochemical [2+2] cycloadditions on thymidine residues in DNA (see Fig. 11). Furocoumarins, such as the well-known psoralen, are found in a variety of plants and can intercalate into duplex DNA. Consecutive photocycloaddition at adjacent thymidine residues can lead to especially toxic DNA cross-links.[106, 107] Before these toxic effects were appreciated furocoumarins were used as tanning enhancers to suntan lotions[108] since the DNA adducts lead to melanogenesis. The toxicity of the furocoumarin derived DNA-adducts has been exploited in extracorporeal photopheretic therapy in cutaneous T cell lymphoma (CTCL) with excellent results.[109-111] In this treatment blood taken from the patients is treated with 8-methoxypsoralen and then treated with UV light before being reintroduced to the patient. A 55.7% overall response and 17.6% complete response has been achieved in CTCL. A similar treatment that employs only skin irradiation after oral or intravenous delivery of 8-methoxypsoralen has been used in the treatment of psoriasis and other skin disorders.[112, 113]

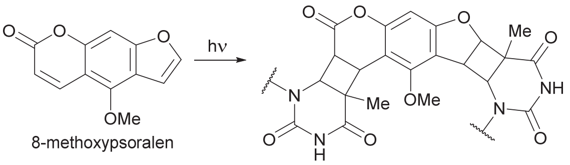


Fig. 11 The furanocoumarin ring system can lead to photocross-linking at A-T sites in duplex DNA

3.3 Artificial chemical agents acting on nucleobases

Nature has used only a few mechanisms to alkylate DNA. Researchers have found a number of molecules that also target DNA, and it is interesting to see that typically the same mechanisms have been employed.

The mustard family contains compounds bearing two chloroethyl side chains attached to a central nitrogen or sulfur atom (see top of Fig.12). While the nitrogen mustards (NM) are used as chemotherapeutics against various types of cancer, sulfur mustards (SM) were widely used as chemical weapons.[114-116] Despite the different mode-of-action of SMs (hydrolysis and conjugation to glutathione involving β-lyase and S-oxidation) many studies indicate that DNA alkylation is related to its toxicity.[117] The alkylation pathway of both mustards follow a similar mechanism through a three-membered heterocycle and preferentially modifying the N7 position of guanine in DNA, comparable to the natural products leinamycin and azinomycin discussed in the previous section. In a first step the aziridium or episulfonium is attacked by a guanine residue. The second chloroethyl side chain can then further react, causing inter- or intra-molecular cross-links (see bottom of Fig. 12).[118-121] These mono or cross-linked DNA modifications cause cellular damage, ultimately triggering apoptosis.[122] Chlormethine is the simplest member of the NM family and it remains a clinically important anti-cancer drug. It is a highly reactive DNA alkylation agent, but is also susceptible to hydrolysis. Varying the substituents on the nitrogen atom from aliphatic to aromatic allows tuning of reactivity and stability of the active electrophile.[123] Activation of chlormethine, for example, is easier compared to chlorambucil (where the nitrogen is N-arylated) due to the higher basicity of the aliphatic nitrogen.

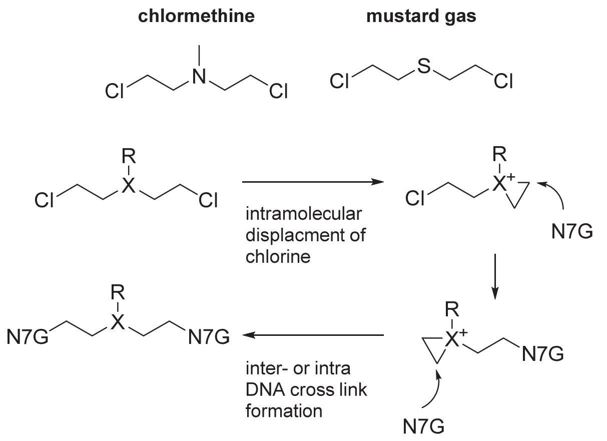

chlormethine
mustard gas
Cl
N
Cl
Cl
S
Cl
R
X
intramolecular displacment of chlorine
X+
N7G
inter- or intra DNA cross link formation
N7G

Fig. 12 Two mustards: chlormethine and mustard gas, representative of synthetic three-membered ring heterocyclic DNA alkylating agents, preferentially alkylate N7G

Temozolomide (TMZ), a synthetic antitumor prodrug, acts as a methyl group transfer agent. Spontaneous breakdown of TMZ to monomethyl triazene (MTIC) in the presence of water and further decomposition yields a highly reactive methyldiazonium electrophile (see Fig. 13).[124, 125] This methylating agent preferentially modifies N7 of guanine in G rich regions, but the O^6 position of guanine and N3 in adenine are also alkylated. However, the cytotoxic and mutagenic effects of this anticancer prodrug are mainly caused by O^6 damage. Polymerases often misincorporate T opposite O^6-alkylated guanines during DNA replication and this can lead to mutagenesis or apoptosis. TMZ is often administered orally to fight brain cancer and *glioblastoma multiform* (a tumour located in the central nervous system). The effects of TMZ can be reduced or completely suppressed by direct repair of this alkylation damage with methylguanine-DNA methyltransferase (MGMT) or low DNA mismatch repair (MMR) activity.[126-128] Besides TMZ there are numerous DNA-targeting molecules of the nitrosourea family (which all follow a similar alkylation mechanism) in the pharmacopeia. These compounds also decompose under physiological conditions to form a diazonium ion which reacts with DNA.[114, 129]

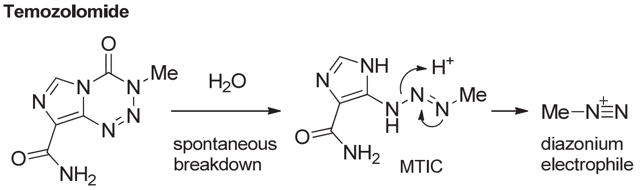


Fig.13 Alkylation of DNA by diazonium groups derived from artificial agents

3.4 Chemical approaches toward sequence selective nucleic acid alkylation

Reactions controlled by Watson-Crick base pairing. One approach for sequence-specific modifications of single-stranded nucleic acids (NAs) is to use a short NA or NA mimic that is covalently linked to a reactive functional group (see Fig. 14). The NA moiety bearing the reactive group is designed to complement a target NA sequence and therefore guides the reactive molecule to a location determined by base-pairing. The functional group may react spontaneously once annealed to the DNA (facilitated by the increased effective molarity induced by duplex formation, as in the Michael addition shown in the top left of Fig. 14)[130] or in other cases its reactivity can be triggered by an external stimulus such as light (see top right of Fig. 14).[131] The Rokita group has pioneered the use of quinone methide electrophiles to target NAs (see bottom right of Fig. 14).[132, 133] A unique feature of the quinone methide system is the reversibility of the alkylation reaction. Reversibility is advantageous because off-target alkylations are not dead-ends; Only the combined effects of base-pairing and covalent bond formation lead to stable alkylation adducts. Although selective, all of the approaches outlined are likely only useful for applications in NA detection since any native function of the NA will be masked by the duplex introduced through binding of the guiding sequence. The functionality transfer technique, introduced by Sasaki,[134] deftly side-steps this problem by making the addition reaction also trigger a secondary reaction that releases the guiding sequence. In practice a 1,4-addition/elimination reaction to transfer a malonyldienone-type electrophile (see bottom left of Fig. 14) has been the most successful.[134] The guiding sequence is connected to the electrophile by a carbon-sulfur bond that is cleaved in the process of transfer. Its precision, without the interference caused by a covalently-linked guiding strand, makes this strategy unique in chemical DNA and RNA alkylation.

A recent addition to the 'guided' reactive functional group concept is the use of groove-binders such as acridine or Hoechst with an electrophilic 2-amino-6-vinylpurine to target abasic sites. The groove-binders sample different binding sites on the DNA and alkylate only abasic sites when bound to A-T rich sites in duplex DNA.[135]

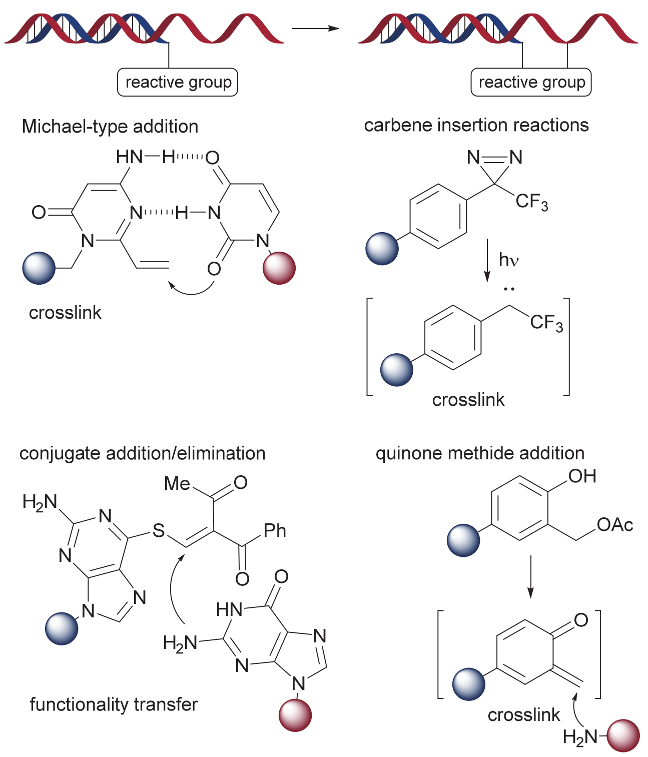


Fig. 14 Attaching a reactive group to an oligonucleotide enables sequence selective reactions through a number of mechanisms. These typically lead to cross-linking and hence are best as detection systems. The functionaly transfer process (bottom left), however, transfers a dienone motif to the target nucleic acid.

Reactions controlled by a self-alkylating sequence. New sequencing technologies have spurred great advances in transcriptomics. Most techniques, however, can only globally assess transcript levels. Localization, timing, and dynamics of specific transcripts remain difficult to determine. The most common technique involves integrating multiple copies of an RNA stem-loop sequence into the target transcript that binds MS2 bacteriophage coat protein with high-affinity.[136-138] The MS2 has proven powerful for large transcripts but smaller RNAs such as siRNAs and miRNAs cannot be visualized with this technique because it would require too large a tag. Minimal systems that might fit the loop region of a hairpin would be a valuable innovation. Evolved self-alkylating ribozymes might be an alternative to the MS2 system, although so far the ribozymes discovered would suffer the same limitations as the MS2 system since they are large and require multiple copies for detection. Nevertheless further development is warranted since ribozymes would engender key advantages over current tagging systems: in particular, a strong covalent tether and avoiding the need for additional expressed proteins. Since the seminal report on self-alkylating ribozymes[139] little progress had been made until recently. The Heemstra group recently described the application of systematic evolution of ligands through exponential enrichment (SELEX) towards identifying self-alkylating ribozymes.[140] They incubated an RNA library with a fluorescein iodoacetamide electrophile and performed pull-down with an anti-fluorescein antibody. After twelve rounds of affinity selection they identified self-alkylating ribozymes with second-order rate constants of 100 ± 8 $M^{-1}s^{-1}$, and these were selective for the evolved aptamer sequence. Although they did not explicitly determine the structure of the alkylaton, the consensus sequences of the most active clones suggest primarily guanine alkylation and since fluorescein iodoacetamide is an S_N2-type electrophile (see section 3.1) it is likely that N7G is the target (see Fig. 15 for putative structure). This would also be consistent with the mode of modification of the first self-alkylating ribozyme.[139]

The Liu group has described a novel approach toward identifying self-alkylating ribozymes that depends on naturally encoded RNA sequences.[141] They created RNA pools derived from genomically encoded sequences of organisms from all three domains of life and treated these with eight common electrophiles that were unreactive with typical RNAs (including epoxide, thioester, α,β-unsaturated amide, α-halo acetamides, fluorophosphonate, and ester). Each electrophile was functionalized with a biotin motif for pull-down and high-throughput sequencing. Using this technique they identified three sequences of interest but two were eliminated because they were either the result of gene fusions generated during library preparation, or were too dependent on sequence-context to be practical. A 63 nucleotide fragment from the thermophilic archaea *Aeropyrum pernix* represented 4.4% of the sequencing reads and was selected for further analysis. They determined that this sequence selectively reacts with the epoxide electrophile (see bottom of Fig. 15) and that a 42-nt truncated sequence was equally competent in self-alkylation. LC/MS and NMR analysis identified the N7 of guanine 9 in the sequence as the site of alkylation. By re-randomizing the RNA pool, alkylating, and then performing high-throughput deep-sequencing they could generate a sequence logo and characterize the critical elements for effective alkylation. While this RNA may be helpful in providing a specific handle for labelling transcripts, an interesting open question is whether self-alkylation might be the natural function of this particular RNA sequence.

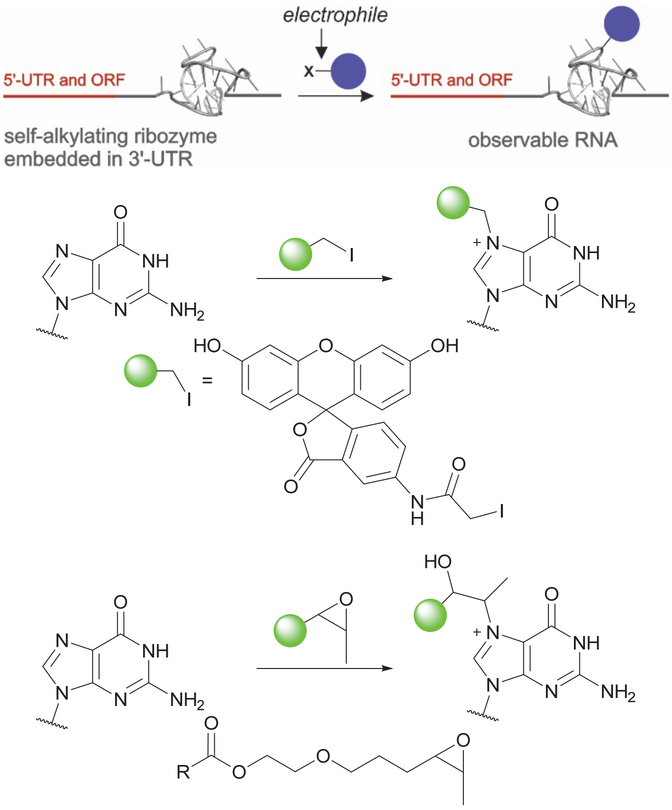


Fig. 15 Top: Self-alkylating ribozymes are capable of activating an electrophile to attack one of its one RNA bases; an mRNA is shown in this specific case. Bottom: Aptamer sequences have been identified for two completely different reaction types: iodoacetamide alkylation and epoxide opening.

3.5 Recent advances in chemical mapping of nucleic acid structure with next-generation sequencing technologies

Chemical modification of nucleic acids with reactive reagents (*i.e.* chemical[142-145] or hydroxyl radical footprinting[146]) and nuclease footprinting[147, 148] are classical methods of determining structure. The reagents in current use however (dimethyl sulfate (DMS), lead(II) cleavage, kethoxal) all have biases in their reactivity and/or cannot be used *in vivo*.[5] For example dimethyl sulfate only alkylates N1-A and N3-C. Selective 2’-hydroxyl acylation analysed by primer extension (SHAPE) makes use of the hydroxyl group that is universally present in RNAs to determine aspects of their structure.[149] Structural information can be gleaned from the fact that 2’-hydroxyls in unpaired or flexible regions are far more reactive than those in duplex regions. 2’-hydroxyl acylation with N-methylisatoic anhydride is effective for *in vitro* SHAPE analysis, but its short half-life in water and cross-reactivity with other nucleophiles has limited its applicability in complex environments. Recently two new reagents have been developed that seem to show no bias in the alkylation of specific 2’-hydroxyls and are stable enough to be used in cell culture.[150] Combined with the information from DMS profiling and next generation sequencing technologies these reagents have recently been employed for the *in vivo* structural profiling of RNA. Although a number of variants[150-154] of the *in vivo* SHAPE-Seq protocol have been developed the basic workflow is the same: add reagent to cells that should alkylate the RNA, isolate total or partial RNA, reverse transcribe to cDNAs, which are truncated wherever a 2’-acylation has occurred. The next step is adaptor ligation, followed by next generation sequencing on the cDNA libraries. Through this approach transcriptome wide information on RNA structure can be gleaned. By comparing *in vitro* and *in vivo* SHAPE-Seq data one can identify instances of altered folding or sites hidden by binding to other partners.[155, 156]

3.6 Catalytic mechanisms for modification of DNA and RNA

One of us recently reviewed catalytic methods for DNA and RNA modification[157] and so in this section we will focus on developments since the appearance of that review. The most effective strategies for site-specific labelling co-opt enzymatic processes. Three recent advances in this area are shown in Fig. 16. The Rentmeister group has developed S-adenosylmethionine analogues that can be transferred to the 5’-cap structure found in eukaryotic mRNAs through the action of trimethylguanosine synthases (GlaTgs).[158-160] They have most recently applied this technology to transfer a vinylbenzyl motif as a ligation handle, which is important since this functional group could be applied with high-rate contant click reactions such as the tetrazole photoclick and the tetrazine inverse electron demand Diels-Alder reaction.[161-163]

tRNA has a large number of post-transcriptional modifications.[8] Most tRNA-modifying enzymes have very specific recognition structures or sequences. Depending on their tolerance toward unnatural substrates, these enzymes could be co-opted to introduce reporter motifs in an RNA of interest. Two groups have recently reported the adaptation of tRNA-modifying enzymes for introducing bioorthogonal reporter molecules into different transcripts.[164, 165] Li et. al. took advantage of a tRNA^{Ile2}-agmatidine synthetase from *Archaeoglobus fulgidus*, which normally transfers an agmatidine moiety to the C34 of the tRNA (see Panel B in Fig. 16), to introduce, for example, a cyclooctyne motif that could later be engaged in a selective reporting reaction in lysates or cells.

In a related approach the Devaraj lab has taken advantage of a bacterial guanine transglycosylation (TGT) enzyme which normally exchanges guanine for the guanine analogue PreQ1 in the wobble position of the tRNA anticodon loop (see panel C in Fig. 16). While the TGT enzyme normally operates on whole tRNA, these researchers have found that a 17 nucleotide stem-loop recognition sequence is enough to make the RNA a substrate for TGT. The ability to use a short 17 nucleotide recognition sequence and the tolerance of TGT to large molecules in the PreQ1 analogues is a great advantage of this method.

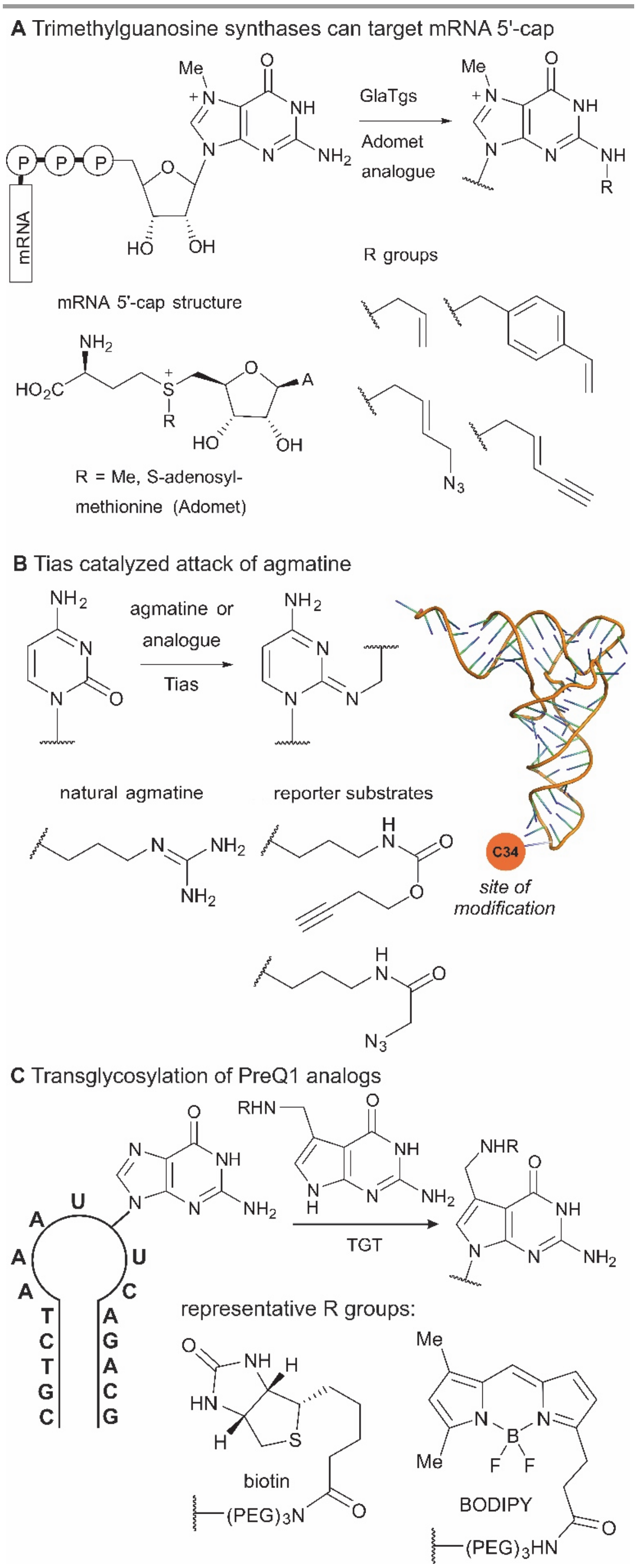


Fig. 16 Co-opting natural enzymes for transcript labelling

4.0 Epigenomics applications of chemical nucleobase modification

In the previous section we examined Nature's approaches toward chemical modification of DNA and how scientists have built molecules that emulate these mechanistic strategies. In the following section we will look at how some chemistry that is unprecedented in Nature has been used to reveal some subtle aspects of nucleic acid biology. Of particular current importance is the availability of chemistry that facilitates the accurate sequencing of non-canonical bases in whole genomes and transcriptomes at single nucleotide resolution. Parsing the 'epigenome' (*i.e.* the regulation of genes occurring above the level of primary genome sequence)[166, 167] has proven to be an enormous challenge,[168] but some simple and selective chemical reactions that distinguish epigenomic bases from both each other and the canonical bases have been essential tools for the task. Excellent reviews on modified RNA have appeared recently,[8, 169] this section will focus on the chemistry that has enabled recent developments in epigenomics and transcriptomics.

4.1 Bisulfite sequencing for m^{5}C

Early work in nucleic acid chemistry focused on understanding reactivity in the nucleobases. These efforts were motivated by (among other things) the need for new sequencing methods,[170] chemical determination of nucleic acid structure (e.g. SHAPE analysis, hydroxyl radical footprinting), and the hope that through characterizing nucleic acid reactivity we could understand which chemicals might be dangerous for genomes. In 1970 a series of papers appeared[171-173] which independently described that uracil and cytosine bases were susceptible to nucleophilic attack by the bisulfite anion. The cytosine addition product (see Fig. 17) was unstable and hydrolysed over several hours with a maximum rate at pH 5.8. Thus a common intermediate from reaction with uracil and cytosine is obtained, which, when the pH is raised to 10, both eliminate the bisulfite to generate uracil. Overall uracil remains unchanged but cytosine is converted to uracil. The reaction only occurs at an appreciable rate in single stranded nucleic acids,[174] which inspired employing the bisulfite reaction for determining structure in folded nucleic acids such as tRNAs.[175]

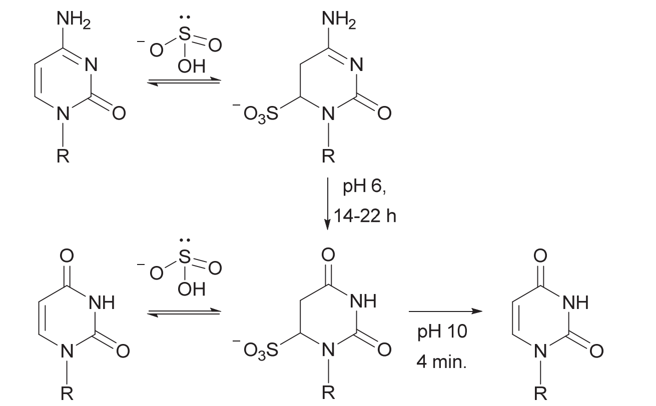


Fig. 17 Bisulfite adds to the 6 position of pyrimidines. In the case of cytosine this leads to deamination

5-methylcytosine (m^5C) occurs widely in DNA[176] and RNA.[177] It is an important regulator of gene expression in eukaryotes and a guardian against methylation-sensitive endonucleases in prokaryotes. The existence of methylated forms of cytosine had been known for a long time[178] and some functions had been speculated,[179] but it was the development of modern DNA sequencing technologies that allowed us to truly appreciate the many biological roles of m^5C. Although Lindahl had determined that m^5C was about four-fold more susceptible to hydrolysis than C,[180] a method to sequence C versus m^5C would require far greater chemoselectivity. The bisulfite reaction was the perfect solution. It was known that m^5C could react with bisulfite,[172] but it turns out the rate is slow enough that a sequencing strategy could be developed (see Fig. 18).[181, 182] First, isolated DNA is sequenced by normal methods. Normal sequencing cannot distinguish C from m^5C. Next, the same DNA is treated with bisulfite to convert Cs to Us, while leaving the m^5Cs intact (see Fig. 18). PCR amplification and sequencing gives a second data set where all Cs are now read as Ts, but m^5Cs are still read as Cs. By comparison with the first sequencing dataset the m^5C sites can be determined with single-nucleotide accuracy. After optimization of conditions,[183] bisulfite sequencing has gained widespread adoption in the biology community and has enabled the sequencing of whole m^5C epigenomes.[184, 185] With methylase assisted bisulfite sequencing, even dynamic changes in DNA methylation states can be monitored.[186]

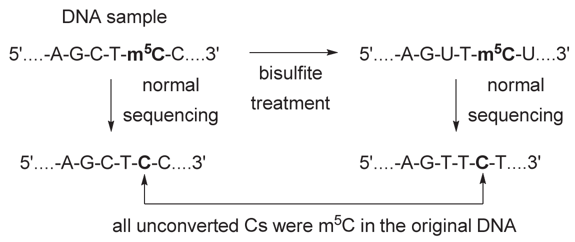


Fig. 18 Concept and workflow for bisulphite sequencing

4.2 Oxidative and reductive bisulfite sequencing for hm^5C and f^5C

In 2009 additional modifications to m^5C were found to exist in mammalian DNA: 5-hydroxymethyl (hm^5C),[187, 188] 5-formyl (f^5C)[189, 190] and 5-carboxylcytosine (ca^5c).[189] hm^5C and f^5C were detected in many cell types (major organs and embryonic stem cells), whereas ca^5C was found in just a few, and at significantly lower concentrations.[189, 191] These modifications were initially viewed as intermediates on the mechanistic pathway for restoration of C from m^5C. The ten-eleven translocation (TET) enzyme, for example, is suggested to catalyse the oxidation of m^5C to its derivatives (hm^5C, f^5C, ca^5c), with subsequenct excision by thymine DNA glycosylase.[192] Further data suggested that each oxidation state of the carbon may exhibit preferential binding to proteins compared to m^5C and each may function as epigenetic transcriptional regulators in their own right.[193]

Accurate detection of these modifications in genomic DNA, on the level of single-base resolution, is needed to fully understand their detailed functions. Different approaches have been developed to map the oxidized forms of m^5C in genomic DNA: antibody-affinity or chemical tagging using hydroxylamines.[194, 195] These methods will not be further discussed in detail due to their low resolution at a genomic level and nonlinearity in their quantitation compared to control assays.[194]

In the previous section it was shown that general bisulfite sequencing can be used for m^5C detection at a genomic level. In contrast, treatment of hm^5C with sodium bisulfite yields a highly stable cytosine-5-methylsulfonate (CMS), which is sequenced as C[196] (see Fig. 19). Differentiation between m^5C and hm^5C is impossible using this method. Therefore, two slightly adapted methodologies were developed to map hm^5C and f^5C: oxidative and reductive bisulfite sequencing (oxBS-Seq[197], redBS-Seq[198]). After oxidation of hm^5C to f^5C, bisulfite treatment leads to deamination and formation of U, which is sequenced as T (see right panel of Fig. 19). In contrast, the redBS-Seq pathway yields a T to C transformation for f^5C after sequencing (see left panel in Fig. 19). In summary the combination of oxidative and reductive BS-Seq allows the quantification of methylcytosines (hm^5C, f^5C) at single-base resolution in genomic DNA. However, the major limitations of this method are the need for multistep bisulfite reactions and deep sequencing in order to achieve enough reads. The combination of chemical modification and enzymatic digestion seems so far to be the best solution to overcome these limitations. For a detailed overview on alternative sequencing methods and recent developments the reader is referred to a recent review by Wu and Zhang.[199]

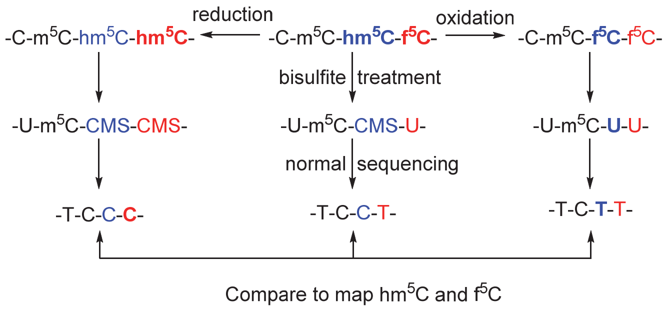


Fig. 19 Concept and workflow for oxidative and reductive bisulphite sequencing

Another sequencing method for hm^{5}C detection is Tet-assisted bisulfite sequencing (TAB-seq), published by He and coworkers.[200] The principle is similar to oxBS-Seq where initial oxididation is followed by bisulfite treatment. In contrast to chemical oxidation the recombinant Tet1 protein is used to transform m^{5}C to ca^{5}C (see Fig. 20). Prior to oxidation hm^{5}C has to be protected by glycosylation with β-glucosyltransferase to form β-glucosyl-5-hydroxymethylcytosine (gm^{5}C). Final bisulfite treatment of the resulting DNA converts all formed ca^{5}C to ca^{5}U whereas gm^{5}C is untouched. After sequencing ca^{5}U is read as T and gm^{5}C as C.

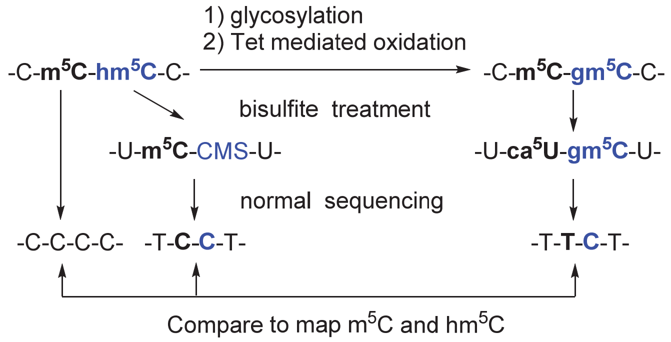


Fig. 20 Concept and workflow for TAB sequencing

4.3 Sequencing of m^{6}A and m^{1}A

Methylated adenine (m^{6}A) has been detected in bacterial, archaeal and eukaryotic genomes[201, 202] as well as in eukaryotic mRNA.[203, 204] These natural compounds add an extra layer of information to DNA/RNA to control gene expression[205] and mRNA metabolism.[206] There are a few examples in the literature for m^{6}A sequencing but most of them are based on specific enzymatic reactions[207] or antibody recognition.[208, 209] The only chemical approach for m^{6}A sequencing was reported by Nakatani in 2010.[210] Selective interstrand cross-link formation (ICL) of a specific DNA duplex bearing a formyl group discriminated between A and m^{6}A based on sensitive imine formation (see Fig. 21). Obvious limitations of this sequencing method are the need for a specific DNA strand with known sequence and synthesis of its complement bearing the required formyl group. Hence, for epigenomic sequencing this method seems to be inappropriate and chemical mapping of m^{6}A remains a challenge. For the foreseeable future it is likely that m^{6}A sequencing will rely on immunoprecipitation.

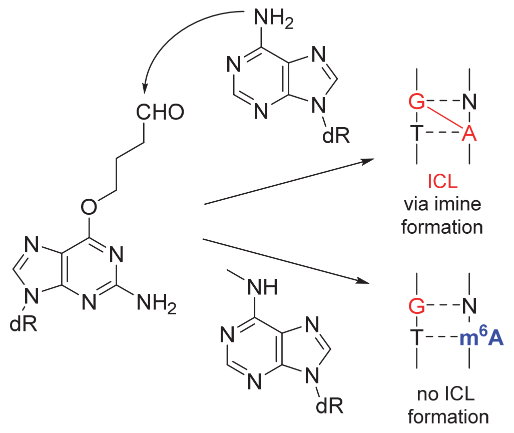
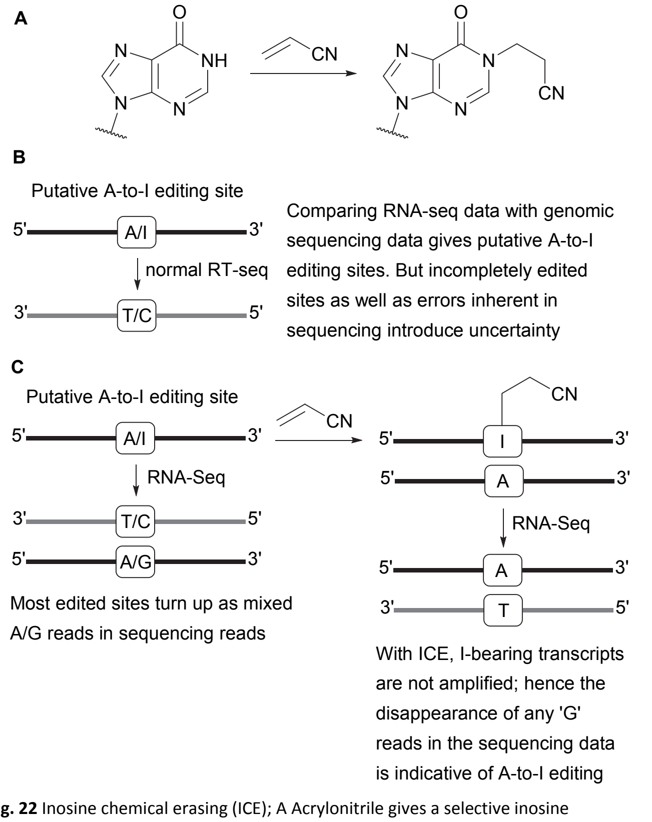


Fig. 21 Concept of m^{6}A sequencing by imine ICL formation

While it has been known for some time that m^{1}A is present in transcriptome isolates, only recently has the technology been invented to sequence this new entry to the 'epitranscriptome'. Indeed two independent reports conclude that m^{1}A occurs in specific regions on mRNAs (such as the first splice site upstream of start codons) and are subject to dynamic regulation.[211, 212] The Yi group exploits anti-m^{1}A antibodies followed by standard sequencing. The m^{1}A sites lead to truncated cDNAs where the end defines the m^{1}A position. A key validation in this technology is an additional sequencing run where the methyl groups are removed by the known *E. coli* DNA and RNA repair protein AlkB. Comparing the truncated positions to the reads of the demethylated sequencing run provides additional confidence in the calling algorithm. In contrast the He group uses a chemical step to identify putative m^{1}A sites. The also begin with an anti-m^{1}A antibody step followed by normal sequencing. Sites that give unusual troughs in the read counts in comparison to neighbouring bases are likely m^{1}A positions since polymerases will lead to truncations and mismatches when this non-canonical base is encountered. A second sequencing is run on samples that have been subjected to high pH. This step induces the well-known Dimroth rearrangement of m^{1}A to m^{6}A.[213] The Watson-Crick face of m^{6}A is the same as normal A and hence the trough will disappear in the sequencing read counts around this position. Comparison of these two runs again allows calling of m^{1}A with high confidence. These two reports are likely the tip of the iceberg for m^{1}A – the hunt is now on to determine how it is regulated and what it regulates.

4.4 Inosine sequencing by acrylonitrile modification

Recently another chemical reaction that was discovered in the sixties has been applied to epigenome sequencing.[214, 215] Yoshida and co-workers determined that inosine and pseudouridine were both effective nucleophiles in a conjugate addition to acrylonitrile, while other bases were untouched. In the presence of magnesium they found that inosine alkylation was preferred over pseudouridine. Acrylonitrile's preference for inosine alkylation has been used to sequence A-to-I editing sites in the transcriptome.[216] In principle, comparing DNA sequencing with reverse transcriptase sequencing reads (*i.e.* the RNA-DNA difference method) should deliver information on A-to-I editing,[217] since inosine is converted to guanine during the reverse transcription and PCR steps. But false positives derived from alignment errors, sequencing bias, PCR errors, and variation in reference genomes,[218-222] mean that orthogonal methods that chemically establish the presence of inosine are important for validation of putative editing sites.[223] Inosine chemical erasing (ICE) has emerged as a complementary technique to the RNA-DNA difference method (see Fig. 22). Here inosine sites are alkylated by acrylonitrile and these are no longer amplified in the PCR reaction because the polymerase cannot bypass the acrylonitrile adduct. In the final sequencing reads I sites are lost and instead replaced by A, as a result of incomplete editing of the original transcript. Comparing this data to a sample that was not treated with acrylonitrile allows the identification of A-to-I editing sites in the original RNA. Although this technique is still relatively new, in combination with the RNA-DNA difference technique it should allow greater confidence is the assignment of A-to-I editing sites in the transcriptome.



Comparing RNA-seq data with genomic sequencing data gives putative A-to-I editing sites. But incompletely edited sites as well as errors inherent in sequencing introduce uncertainty

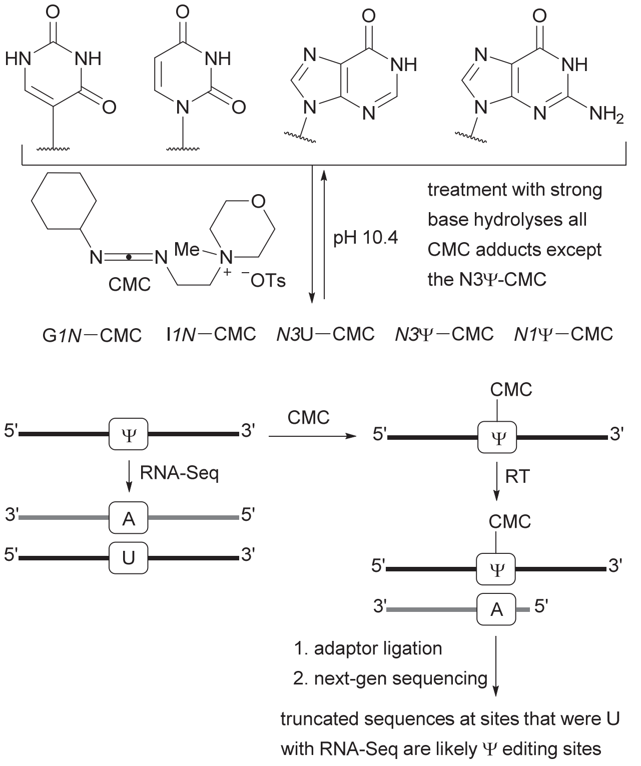
Most edited sites turn up as mixed A/G reads in sequencing reads

With ICE, I-bearing transcripts are not amplified; hence the disappearance of any 'G' reads in the sequencing data is indicative of A-to-I editing

Fig. 22 Inosine chemical erasing (ICE); A Acrylonitrile gives a selective inosine modification in the presence of other bases. B The DNA-RNA difference technique compares RNA-seq with genome sequencing. C ICE-Seq relies on the selective modification of inosine with acrylonitrile, which blocks first-strand cDNA synthesis leading to the disappearance of these strands in the final sequencing data

4.5 Pseudouridine sequencing by carbodiimide modification

Pseudouridine (Ψ) is one of the earliest characterized post-transcriptional modifications. It has been known for some time that uridines are extensively edited to pseudouridines in tRNA, but recent efforts on whole transcriptome pseudouridine sequencing (Ψ-seq) have revealed that pseudouridine occurs broadly in RNA.[224, 225] The Ψ-seq method was enabled again by some simple chemistry that was discovered decades ago. In 1962 Gilham established that of the canonical bases only guanine and uridine react at an appreciable rate with the carbodiimide reagent N-cyclohexyl-N'-B-(4-methylmorpholinium)-ethylcarbodiimide (CMC).[226, 227] Later work established that pseudouridine and inosine also react with CMC, but unlike the G, U, and I adducts the Ψ-adduct is highly resistant to hydrolysis and aminolysis.[228] The relative stability of the Ψ-adduct with CMC and the propensity of these adducts to halt reverse transcriptase polymerization has been exploited for sequencing Ψ sites (see Fig. 23);[229-231] First with single RNAs and reverse transcriptase blockade assays, and more recently in massively parallel format to sequence classes of RNAs or even entire transcriptomes.[224, 225, 232, 233] Many more Ψ sites have been identified than expected and the new challenge is to determine the function of these edited sites.



treatment with strong base hydrolyses all CMC adducts except the N3Ψ-CMC

truncated sequences at sites that were U with RNA-Seq are likely Ψ editing sites

Fig. 23 Psi-Seq or Ψ-Seq uses a selective modification of Ψ to characterize editing sites. Top: The carbodiimide reagent CMC reacts with a number of lactam functions in RNA, but all except the NxY-CMC is removed on treatment with base. Bottom: The CMC adduct halts reverse transcription, leading to truncated sequences in the sequencing data.

5 Conclusions and future outlook

An understanding of the properties and reactivity of DNA and RNA has been crucial in the development of technologies for studying epigenetics and the transcriptome. There are plenty of known modifications in the transcriptome for which we have no sequencing method and whose function we don't understand.[8,

[169] Furthermore, it is likely that there are important modifications waiting to be discovered.[234] New chemistry will be required to sequence modified DNAs and RNAs and to determine their function. While the present review makes it clear how important selective chemistry is in the study of nucleic acids, it should also serve to highlight that continued research in nucleic acid chemistry is vital.

References

1. A. Razin, *EMBO J.*, 1998, **17**, 4905-4908.
2. G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y.-G. Yang and C. He, *Nat. Chem. Biol.*, 2011, **7**, 885-887.
3. 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 and C. He, *Mol. Cell*, 2013, **49**, 18-29.
4. n. prize, The Nobel Prize in Chemistry 2015 - Press Release". Nobelprize.org. Nobel Media AB 2014. Web. 6 Nov 2015. <http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2015/press.html>.
5. R. K. Hartmann, A. Bindereif, A. Sch_n and E. Westhof, *Handbook of RNA biochemistry*, John Wiley & Sons, 2015.
6. J. B. Opalinska and A. M. Gewirtz, *Nat. Rev. Drug Discov.*, 2002, **1**, 503-514.
7. V. K. Sharma, P. Rungta and A. K. Prasad, *RSC Advances*, 2014, **4**, 16618-16631.
8. T. Carell, C. Brandmayr, A. Hienzsch, M. Müller, D. Pearson, V. Reiter, I. Thoma, P. Thumbs and M. Wagner, *Angew. Chem. Int. Ed.*, 2012, **51**, 7110-7131.
9. R. M. Izatt, J. J. Christensen and J. H. Rytting, *Chem. Rev.*, 1971, **71**, 439-481.
10. P. Legault and A. Pardi, *J. Am. Chem. Soc.*, 1997, **119**, 6621-6628.
11. T. K. Harris and G. J. Turner, *IUBMB Life*, 2002, **53**, 85-98.
12. J. K. Lanyi, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2006, **1757**, 1012-1018.
13. K. Ramanarayanan, *Acc. Chem. Res.*, 2012, **45**, 2035-2044.
14. M. Skilandat and R. K. O. Sigel, *J. Biol. Chem.*, 2014, **289**, 20650-20663.
15. M. R. Stahley and S. A. Strobel, *Science*, 2005, **309**, 1587-1590.
16. D. M. J. Lilley, *Mechanisms of RNA catalysis*, 2011.
17. L. Sharmeen, M. Kuo, G. Dinter-Gottlieb and J. Taylor, *Journal of Virology*, 1988, **62**, 2674-2679.
18. B. Gong, J.-H. Chen, E. Chase, D. M. Chadalavada, R. Yajima, B. L. Golden, P. C. Bevilacqua and P. R. Carey, *J. Am. Chem. Soc.*, 2007, **129**, 13335-13342.
19. G. Mata and N. W. Luedtke, *J. Am. Chem. Soc.*, 2015, **137**, 699-707.
20. S. Kendrick, Y. Akiyama, S. M. Hecht and L. H. Hurley, *J. Am. Chem. Soc.*, 2009, **131**, 17667-17676.
21. A. Rajendran, S.-i. Nakano and N. Sugimoto, *Chem. Commun.*, 2010, **46**, 1299-1301.
22. T. A. Brooks, S. Kendrick and L. Hurley, *FEBS Journal*, 2010, **277**, 3459-3469.
23. H.-J. Kang, S. Kendrick, S. M. Hecht and L. H. Hurley, *J. Am. Chem. Soc.*, 2014, **136**, 4172-4185.
24. S. Kendrick, H.-J. Kang, M. P. Alam, M. M. Madathil, P. Agrawal, V. Gokhale, D. Yang, S. M. Hecht and L. H. Hurley, *J. Am. Chem. Soc.*, 2014, **136**, 4161-4171.
25. J. Watson, *The double helix*, Hachette UK, 2012.
26. J. Černý, M. Kabeláč and P. Hobza, *J. Am. Chem. Soc.*, 2008, **130**, 16055-16059.
27. V. R. Cooper, T. Thonhauser, A. Puzder, E. Schröder, B. I. Lundqvist and D. C. Langreth, *J. Am. Chem. Soc.*, 2008, **130**, 1304-1308.
28. R. A. DiStasio, O. A. von Lilienfeld and A. Tkatchenko, *Proc. Natl. Acad. Sci. USA*, 2012, **109**, 14791-14795.
29. J. Sponer, A. Mladek, J. E. Sponer, D. Svozil, M. Zgarbova, P. Banas, P. Jurecka and M. Otyepka, *Physical Chemistry Chemical Physics*, 2012, **14**, 15257-15277.
30. P. Jurecka, J. Sponer, J. Cerny and P. Hobza, *Physical Chemistry Chemical Physics*, 2006, **8**, 1985-1993.
31. J. Sponer, K. E. Riley and P. Hobza, *Physical Chemistry Chemical Physics*, 2008, **10**, 2595-2610.
32. I. C. Lin and U. Rothlisberger, *Physical Chemistry Chemical Physics*, 2008, **10**, 2730-2734.
33. I. C. Lin, O. A. von Lilienfeld, M. D. Coutinho-Neto, I. Tavernelli and U. Rothlisberger, *The Journal of Physical Chemistry B*, 2007, **111**, 14346-14354.
34. T. M. Parker, E. G. Hohenstein, R. M. Parrish, N. V. Hud and C. D. Sherrill, *J. Am. Chem. Soc.*, 2013, **135**, 1306-1316.
35. O. A. von Lilienfeld, I. Tavernelli, U. Rothlisberger and D. Sebastiani, *Physical Review Letters*, 2004, **93**, 153004.
36. C. D. M. Churchill, L. Navarro-Whyte, L. R. Rutledge and S. D. Wetmore, *Physical Chemistry Chemical Physics*, 2009, **11**, 10657-10670.
37. W. Koch and M. C. Holthausen, *A chemist's guide to density functional theory*, 2001.
38. J. P. Perdew, *Physical Review B*, 1986, **33**, 8822-8824.
39. P. Stephens, F. Devlin, C. Chabalowski and M. J. Frisch, *The Journal of Physical Chemistry*, 1994, **98**, 11623-11627.
40. C. Campos and F. Jorge, *International Journal of Quantum Chemistry*, 2009, **109**, 285-293.
41. A. Alparone, *Chemical Physics*, 2013, **410**, 90-98.
42. T. Yanai, D. P. Tew and N. C. Handy, *Chemical Physics Letters*, 2004, **393**, 51-57.
43. C. Párkányi, C. Boniface, J.-J. Aaron, M. Bulaceanu-MacNair and M. Dakkouri, *Collection of Czechoslovak chemical communications*, 2002, **67**, 1109-1124.
44. X. Lu, J. M. Heilman, P. Blans and J. C. Fishbein, *Chem. Res. Toxicol.*, 2005, **18**, 1462-1470.
45. T. B. Phan, M. Breugst and H. Mayr, *Angew. Chem. Int. Ed.*, 2006, **45**, 3869-3874.
46. S.-I. Ninomiya, Kawazoe, Y., *Chem. Pharm. Bull.*, 1985, **33**, 5207-5213.
47. A. E. Pegg, *Cancer Investigation*, 1984, **2**, 223-231.
48. G. Klopman, *J. Am. Chem. Soc.*, 1968, **90**, 223-234.
49. L. Salem, *J. Am. Chem. Soc.*, 1968, **90**, 543-552.
50. M. Breugst, H. Zipse, J. P. Guthrie and H. Mayr, *Angew. Chem. Int. Ed.*, 2010, **49**, 5165-5169.
51. U. Galm, M. H. Hager, S. G. Van Lanen, J. Ju, J. S. Thorson and B. Shen, *Chem. Rev.*, 2005, **105**, 739-758.
52. W. C. Tse and D. L. Boger, *Chem. Biol.*, 2004, **11**, 1607-1617.
53. S. E. Wolkenberg and D. L. Boger, *Chem. Rev.*, 2002, **102**, 2477-2496.

54. M.-Y. Kim, H. Vankayalapati, K. Shin-ya, K. Wierzba and L. H. Hurley, *J. Am. Chem. Soc.*, 2002, **124**, 2098-2099.
55. N. Zein, A. Sinha, W. McGahren and G. Ellestad, *Science*, 1988, **240**, 1198-1201.
56. S. M. Hecht, *Journal of Natural Products*, 2000, **63**, 158-168.
57. J. P. Parrish, D. B. Kastrinsky, S. E. Wolkenberg, Y. Igarashi and D. L. Boger, *J. Am. Chem. Soc.*, 2003, **125**, 10971-10976.
58. D. L. Boger and R. M. Garbaccio, *Bioorg. Med. Chem.*, 1997, **5**, 263-276.
59. D. L. Boger and R. M. Garbaccio, *Acc. Chem. Res.*, 1999, **32**, 1043-1052.
60. R. C. FusoN, *Chem. Rev.*, 1935, **16**, 1-27.
61. S. E. Denmark and G. L. Beutner, *J. Am. Chem. Soc.*, 2003, **125**, 7800-7801.
62. T. C. McMorris, M. J. Kelner, W. Wang, L. A. Estes, M. A. Montoya and R. Taetle, *J. Org. Chem.*, 1992, **57**, 6876-6883.
63. Y.-H. HSU, A. HIROTA, S. SHIMA, M. NAKAGAWA, T. ADACHI, H. NOZAKI and M. NAKAYAMA, *J. Antibiot.*, 1989, **42**, 223-229.
64. K. Yamada, M. Ojika and H. Kigoshi, *Natural Product Reports*, 2007, **24**, 798-813.
65. K. Yamada, M. Ojika and H. Kigoshi, *Angew. Chem. Int. Ed.*, 1998, **37**, 1818-1826.
66. M. Tomasz, *Chem. Biol.*, 1995, **2**, 575-579.
67. P. D. Bass, D. A. Gubler, T. C. Judd and R. M. Williams, *Chem. Rev.*, 2013, **113**, 6816-6863.
68. H. W. Moore, *Science*, 1977, **197**, 527-532.
69. A. Turner, *Quarterly Reviews, Chemical Society*, 1964, **18**, 347-360.
70. M. Tomasz, D. Chowdary, R. Lipman, S. Shimotakahara, D. Veiro, V. Walker and G. L. Verdine, *Proc. Natl. Acad. Sci. USA*, 1986, **83**, 6702-6706.
71. M. Tomasz, R. Lipman, D. Chowdary, J. Pawlak, G. Verdine and K. Nakanishi, *Science*, 1987, **235**, 1204-1208.
72. R. Bizanek, B. F. McGuinness, K. Nakanishi and M. Tomasz, *Biochemistry*, 1992, **31**, 3084-3091.
73. A. D. Patil, A. J. Freyer, R. Reichwein, B. Carte, L. B. Killmer, L. Faucette, R. K. Johnson and D. J. Faulkner, *Tetrahedron Lett.*, 1997, **38**, 363-364.
74. S. Dutta, H. Abe, S. Aoyagi, C. Kibayashi and K. S. Gates, *J. Am. Chem. Soc.*, 2005, **127**, 15004-15005.
75. R. S. Coleman, R. J. Perez, C. H. Burk and A. Navarro, *J. Am. Chem. Soc.*, 2002, **124**, 13008-13017.
76. R. W. Armstrong, M. E. Salvati and M. Nguyen, *J. Am. Chem. Soc.*, 1992, **114**, 3144-3145.
77. J. H. Williams, T. D. Phillips, P. E. Jolly, J. K. Stiles, C. M. Jolly and D. Aggarwal, *The American Journal of Clinical Nutrition*, 2004, **80**, 1106-1122.
78. J. M. Essigmann, R. G. Croy, A. M. Nadzan, W. F. Busby, V. N. Reinhold, G. Büchi and G. N. Wogan, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 1870-1874.
79. M. Hara, K. Asano, I. Kawamoto, T. Takiguchi, S. Katsumata, K. I. Takahashi and H. Nakano, *Journal of Antibiotics*, 1989, **42**, 1768-1774.
80. K. S. Gates, *Chem. Res. Toxicol.*, 2000, **13**, 953-956.
81. H. Zang and K. S. Gates, *Chem. Res. Toxicol.*, 2003, **16**, 1539-1546.
82. L. Breydo, H. Zang, K. Mitra and K. S. Gates, *J. Am. Chem. Soc.*, 2001, **123**, 2060-2061.
83. V. Viswesh, K. Gates and D. Sun, *Chem. Res. Toxicol.*, 2010, **23**, 99-107.
84. M. I. Fekry, J. Szekely, S. Dutta, L. Breydo, H. Zang and K. S. Gates, *J. Am. Chem. Soc.*, 2011, **133**, 17641-17651.
85. M. A. M. M. P. Doyle, T. Ye, *Modern Catalytic Methods for Organic Synthesis with Diazo Compounds*, Wiley, New York, 1998.
86. C. C. Nawrat and C. J. Moody, *Natural Product Reports*, 2011, **28**, 1426-1444.
87. S. B. Herzon and C. M. Woo, *Natural Product Reports*, 2012, **29**, 87-118.
88. L. C. Colis, C. M. Woo, D. C. Hegan, Z. Li, P. M. Glazer and S. B. Herzon, *Nature Chem.*, 2014, **6**, 504-510.
89. R. Catane, D. D. Von Hoff, D. L. Glaubiger and F. M. Muggia, *Cancer Treat Rep*, 1979, **63**, 1033-1038.
90. D. Cervantes-Madrid, Y. Romero, Due, #xf1, as-Gonz, #xe1 and A. lez, *BioMed Research International*, 2015, **2015**, 13.
91. R. R. Ellison, D. A. Karnofsky, S. S. Sternberg, M. L. Murphy and J. H. Burchenal, *Cancer*, 1954, **7**, 801-814.
92. Y. Katoh, M. Maekawa and Y. Sano, *Mutation Research/Genetic Toxicology*, 1995, **342**, 37-41.
93. H. S. Lilja, E. Hyde, D. S. Longnecker and J. D. Yager, *Cancer research*, 1977, **37**, 3925-3931.
94. H. S. Lilja, D. S. Longnecker, T. J. Curphey, D. S. Daniel and W. E. Adams, *Cancer Letters*, 1981, **12**, 139-146.
95. B. Sedgwick, *Carcinogenesis*, 1997.
96. M. H. Lewin, N. Bailey, T. Bandaletova and R. Bowman, *Cancer research*, 2006, DOI: 10.1158/0008-5472.CAN-05-2237.
97. J. F. McGarrity and T. Smyth, *J. Am. Chem. Soc.*, 1980, **102**, 7303-7308.
98. J. J. Vavra, C. Deboer, A. Dietz, L. J. Hanka and W. T. Sokolski, *Antibiotics annual*, 1959, **7**, 230-235.
99. P. Lawley, *Nature*, 1968, **218**, 580-581.
100. H. Yamamoto, Y. Uchigata and H. Okamoto, *Nature*, 1981, **294**, 284-286.
101. S. Lenzen, *Diabetologia*, 2008, **51**, 216-226.
102. T. Szkudelski, *Physiological research*, 2001, **50**, 537-546.
103. Y. Pommier, G. Kohlhagen, C. Bailly, M. Waring, A. Mazumder and K. W. Kohn, *Biochemistry*, 1996, **35**, 13303-13309.
104. L. H. Hurley and R. Petrusek, 1979.
105. K. E. Rao and J. W. Lown, *Chem. Res. Toxicol.*, 1990, **3**, 262-267.
106. D. Kanne, K. Straub, H. Rapoport and J. E. Hearst, *Biochemistry*, 1982, **21**, 861-871.
107. S. S. Sastry, H. P. Spielmann, T. J. Dwyer, D. E. Wemmer and J. E. Hearst, *Journal of Photochemistry and Photobiology B: Biology*, 1992, **14**, 65-79.
108. M. J. Ashwood-Smith, G. A. Poulton, M. Barker and M. Mildenberger, *Nature*, 1980, **285**, 407-409.
109. R. Edelson, C. Berger, F. Gasparro, B. Jegasothy, P. Heald, B. Wintroub, E. Vonderheid, R. Knobler, K. Wolff and G. Plewig, *N. Engl. J. Med.*, 1987, **316**, 297-303.
110. R. L. Edelson, *The Yale Journal of Biology and Medicine*, 1989, **62**, 565-577.
111. J. A. Zic, *Dermatologic Therapy*, 2003, **16**, 337-346.
112. T. Henseler, H. Hönigsmann, K. Wolff and E. Christophers, *The Lancet*, 1981, **317**, 853-857.
113. A. K. Gupta and T. F. Anderson, *Journal of the American Academy of Dermatology*, 1987, **17**, 703-734.
114. S. R. Rajski and R. M. Williams, *Chem. Rev.*, 1998, **98**, 2723-2796.

115. B. Fisher, B. Sherman, H. Rockette, C. Redmond, R. Margolese and E. R. Fisher, *Cancer*, 1979, **44**, 847-857.

116. M. Hefazi, *Archives of Iranian Medicine*, 2005, **8**, 162-179.

117. M. Pesonen, K. Vähäkangas, M. Halme, P. Vanninen, H. Seulanto, M. Hemmilä, M. Pasanen and T. Kuitunen, *Frontiers in Pharmacology*, 2010, **1**.

118. B. Singer, *Nature*, 1976, **264**, 333-339.

119. B. Neog, S. Sinha and P. K. Bhattacharyya, *Computational and Theoretical Chemistry*, 2013, **1018**, 19-25.

120. K. W. Kohn, J. A. Hartley and W. B. Mattes, *Nucleic Acids Res.*, 1987, **15**, 10531-10549.

121. G. P. Wheeler, *Cancer Research*, 1962, **22**, 651-688.

122. P. D. Lawley and D. H. Phillips, *Mutat. Res.*, 1996, **355**, 13-40.

123. A. Polavarapu, J. A. Stillabower, S. G. W. Stubblefield, W. M. Taylor and M.-H. Baik, *J. Org. Chem.*, 2012, **77**, 5914-5921.

124. E. Newlands, M. Stevens, S. Wedge, R. Wheelhouse and C. Brock, *Cancer treatment reviews*, 1997, **23**, 35-61.

125. Z. Jihong, F. G. S. Malcolm and D. B. Tracey, *Current Molecular Pharmacology*, 2012, **5**, 102-114.

126. Y. Ramirez, J. Weatherbee, R. Wheelhouse and A. Ross, *Pharmaceuticals*, 2013, **6**, 1475.

127. S. A. Kyrtopoulos, L. M. Anderson, S. K. Chhabra, V. L. Souliotis, V. Pletsa, C. Valavanis and P. Georgiadis, *Cancer detection and prevention*, 1996, **21**, 391-405.

128. G. P. Margison and M. F. Santibáñez-Koref, *BioEssays*, 2002, **24**, 255-266.

129. F.-X. Chen, W. J. Bodell, G. Liang and B. Gold, *Chem. Res. Toxicol.*, 1996, **9**, 208-214.

130. T. Kawasaki, F. Nagatsugi, M. M. Ali, M. Maeda, K. Sugiyama, K. Hori and S. Sasaki, *J. Org. Chem.*, 2004, **70**, 14-23.

131. C.-X. Song and C. He, *Acc. Chem. Res.*, 2011, **44**, 709-717.

132. Q. Zhou and S. E. Rokita, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 15452-15457.

133. Y. Liu and S. E. Rokita, *Biochemistry*, 2012, **51**, 1020-1027.

134. K. Onizuka, Y. Taniguchi and S. Sasaki, *Bioconjug. Chem.*, 2009, **20**, 799-803.

135. N. Sato, G. Tsuji, Y. Sasaki, A. Usami, T. Moki, K. Onizuka, K. Yamada and F. Nagatsugi, *Chem. Commun.*, 2015, **51**, 14885-14888.

136. J. A. Chao, K. Czaplinski and R. H. Singer, in *Probes and Tags to Study Biomolecular Function*, Wiley-VCH Verlag GmbH & Co. KGaA, 2008, DOI: 10.1002/9783527623099.ch9, pp. 163-174.

137. J. M. Halstead, T. Lionnet, J. H. Wilbertz, F. Wippich, A. Ephrussi, R. H. Singer and J. A. Chao, *Science*, 2015, **347**, 1367-1671.

138. S. Hocine, P. Raymond, D. Zenklusen, J. A. Chao and R. H. Singer, *Nat Meth*, 2013, **10**, 119-121.

139. C. Wilson and J. W. Szostak, *Nature*, 1995, **374**, 777-782.

140. A. K. Sharma, J. J. Plant, A. E. Rangel, K. N. Meek, A. J. Anamisis, J. Hollien and J. M. Heemstra, *ACS Chem. Biol.*, 2014, **9**, 1680-1684.

141. R. I. McDonald, J. P. Guilinger, S. Mukherji, E. A. Curtis, W. I. Lee and D. R. Liu, *Nat. Chem. Biol.*, 2014, **10**, 1049-1054.

142. T. Inoue and T. R. Cech, *Proc. Natl. Acad. Sci. USA*, 1985, **82**, 648-652.

143. D. Moazed and H. F. Noller, *Cell*, 1986, **47**, 985-994.

144. D. Moazed, S. Stern and H. F. Noller, *J Mol Biol*, 1986, **187**, 399-416.

145. S. Stern, D. Moazed and H. F. Noller, in *Methods in Enzymology*, Academic Press, 1988, vol. Volume 164, pp. 481-489.

146. T. D. Tullius and J. A. Greenbaum, *Curr. Op. Chem. Biol.*, 2005, **9**, 127-134.

147. D. J. Galas and A. Schmitz, *Nucleic Acids Res.*, 1978, **5**, 3157-3170.

148. J. D. Gralla, *Proc. Natl. Acad. Sci. USA*, 1985, **82**, 3078-3081.

149. E. J. Merino, K. A. Wilkinson, J. L. Coughlan and K. M. Weeks, *J. Am. Chem. Soc.*, 2005, **127**, 4223-4231.

150. R. C. Spitale, P. Crisalli, R. A. Flynn, E. A. Torre, E. T. Kool and H. Y. Chang, *Nat. Chem. Biol.*, 2013, **9**, 18-20.

151. J. Talkish, G. May, Y. Lin, J. L. Woolford and C. J. McManus, *RNA*, 2014, **20**, 713-720.

152. 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 and H. Y. Chang, *Nature*, 2015, **519**, 486-490.

153. N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson and K. M. Weeks, *Nat Meth*, 2014, **11**, 959-965.

154. K. E. Watters, T. R. Abbott and J. B. Lucks, *Nucleic Acids Res.*, 2015, DOI: 10.1093/nar/gkv879.

155. Y. Ding, C. K. Kwok, Y. Tang, P. C. Bevilacqua and S. M. Assmann, *Nat. Protocols*, 2015, **10**, 1050-1066.

156. C. K. Kwok, Y. Ding, Y. Tang, S. M. Assmann and P. C. Bevilacqua, *Nat. Commun.*, 2013, **4**.

157. D. Gillingham and R. Shahid, *Curr. Op. Chem. Biol.*, 2015, **25**, 110-114.

158. J. M. Holstein, D. Schulz and A. Rentmeister, *Chem. Commun.*, 2014, **50**, 4478-4481.

159. J. M. Holstein, D. Stummer and A. Rentmeister, *Chem. Sci.*, 2015, **6**, 1362-1369.

160. D. Schulz, J. M. Holstein and A. Rentmeister, *Angew. Chem. Int. Ed.*, 2013, **52**, 7874-7878.

161. D. M. Patterson, L. A. Nazarova and J. A. Prescher, *ACS Chem. Biol.*, 2014, **9**, 592-605.

162. M. Blackman, M. Royzen and J. Fox, *J. Am. Chem. Soc.*, 2008, **130**, 13518-13519.

163. N. K. Devaraj, R. Weissleder and S. A. Hilderbrand, *Bioconjug. Chem.*, 2008, **19**, 2297-2299.

164. F. Li, J. Dong, X. Hu, W. Gong, J. Li, J. Shen, H. Tian and J. Wang, *Angew. Chem. Int. Ed.*, 2015, **54**, 4597-4602.

165. S. C. Alexander, K. N. Busby, C. M. Cole, C. Y. Zhou and N. K. Devaraj, *J. Am. Chem. Soc.*, 2015, **137**, 12756-12759.

166. B. E. Bernstein, A. Meissner and E. S. Lander, *Cell*, 2007, **128**, 669-681.

167. P. A. Callinan and A. P. Feinberg, *Hum. Mol. Genet.*, 2006, **15**, R95-R101.

168. J.-K. Zhu, *Cell*, 2008, **133**, 395-397.

169. I. Behm-Ansmant, M. Helm and Y. Motorin, *Journal of Nucleic Acids*, 2011, **2011**.

170. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 560-564.

171. R. Shapiro, R. E. Servis and M. Welcher, *J. Am. Chem. Soc.*, 1970, **92**, 422-424.

172. H. Hayatsu, Y. Wataya, K. Kai and S. Iida, *Biochemistry*, 1970, **9**, 2858-2865.

173. H. Hayatsu, Y. Wataya and K. Kai, *J. Am. Chem. Soc.*, 1970, **92**, 724-726.

174. R. Shapiro, B. Braverman, J. B. Louis and R. E. Servis, *J. Biol. Chem.*, 1973, **248**, 4060-4064.

175. A. McLaren, E. S. Gonos, T. Carr and J. P. Goddard, *FEBS Letters*, 1993, **330**, 177-180.

176. J. E. Squires, H. R. Patel, M. Nousch, T. Sibbritt, D. T. Humphreys, B. J. Parker, C. M. Suter and T. Preiss, *Nucleic Acids Res.*, 2012, **40**, 5023-5033.

177. Y. Motorin, F. Lyko and M. Helm, *Nucleic Acids Res.*, 2010, **38**, 1415-1430.

178. T. B. Johnson and R. D. Coghill, *J. Am. Chem. Soc.*, 1925, **47**, 2838-2844.

179. R. Holliday, *Science*, 1987, **238**, 163-170.

180. T. Lindahl and B. Nyberg, *Biochemistry*, 1974, **13**, 3405-3410.

181. M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy and C. L. Paul, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 1827-1831.

182. J. C. Susan, J. Harrison, C. L. Paul and M. Frommer, *Nucleic Acids Res.*, 1994, **22**, 2990-2997.

183. C. Grunau, S. Clark and A. Rosenthal, *Nucleic Acids Res.*, 2001, **29**, e65-e65.

184. I. R. Henderson and S. E. Jacobsen, *Nature*, 2007, **447**, 418-424.

185. R. Lister, R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker, *Cell*, 2008, **133**, 523-536.

186. H. Wu, X. Wu, L. Shen and Y. Zhang, *Nat. Biotech.*, 2014, **32**, 1231-1240.

187. S. Kriaucionis and N. Heintz, *Science (New York, N.y.)*, 2009, **324**, 929-930.

188. M. Tahiliani, K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind and A. Rao, *Science*, 2009, **324**, 930-935.

189. S. Ito, L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He and Y. Zhang, *Science*, 2011, **333**, 1300-1303.

190. T. Pfaffeneder, B. Hackner, M. Truß, M. Münzel, M. Müller, C. A. Deiml, C. Hagemeier and T. Carell, *Angew. Chem. Int. Ed.*, 2011, **50**, 7008-7012.

191. S. Liu, J. Wang, Y. Su, C. Guerrero, Y. Zeng, D. Mitra, P. J. Brooks, D. E. Fisher, H. Song and Y. Wang, *Nucleic Acids Res.*, 2013, **41**, 6421-6429.

192. H. Hashimoto, S. Hong, A. S. Bhagwat, X. Zhang and X. Cheng, *Nucleic Acids Res.*, 2012, **40**, 10203-10214.

193. M. Mellén, P. Ayata, S. Dewell, S. Kriaucionis and N. Heintz, *Cell*, **151**, 1417-1430.

194. E.-A. Raiber, D. Beraldi, G. Ficz, H. E. Burgess, M. R. Branco, P. Murat, D. Oxley, M. J. Booth, W. Reik and S. Balasubramanian, *Genome Biol*, 2012, **13**, R69.

195. 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 and C. He, *Cell*, 2013, **153**, 678-691.

196. Y. Huang, W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu and A. Rao, *PLoS ONE*, 2010, **5**, e8888.

197. M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik and S. Balasubramanian, *Science*, 2012, **336**, 934-937.

198. M. J. Booth, G. Marsico, M. Bachman, D. Beraldi and S. Balasubramanian, *Nature Chem.*, 2014, **6**, 435-440.

199. H. Wu and Y. Zhang, *Nat Struct Mol Biol*, 2015, **22**, 656-661.

200. M. Yu, Gary C. Hon, Keith E. Szulwach, C.-X. Song, L. Zhang, A. Kim, X. Li, Q. Dai, Y. Shen, B. Park, J.-H. Min, P. Jin, B. Ren and C. He, *Cell*, 2012, **149**, 1368-1380.

201. D. Wion and J. Casadesús, *Nature Reviews. Microbiology*, 2006, **4**, 183-192.

202. D. Ratel, J.-L. Ravanat, F. Berger and D. Wion, *BioEssays*, 2006, **28**, 309-315.

203. R. Desrosiers, K. Friderici and F. Rottman, *Proc. Natl. Acad. Sci. USA*, 1974, **71**, 3971-3975.

204. R. C. Desrosiers, K. H. Friderici and F. M. Rottman, *Biochemistry*, 1975, **14**, 4367-4374.

205. K. D. Robertson and A. P. Wolffe, *Nat Rev Genet*, 2000, **1**, 11-19.

206. X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan and C. He, *Nature*, 2014, **505**, 117-120.

207. A. Bart, M. W. J. van Passel, K. van Amsterdam and A. van der Ende, *Nucleic Acids Res.*, 2005, **33**, e124.

208. K. Chen, Z. Lu, X. Wang, Y. Fu, G.-Z. Luo, N. Liu, D. Han, D. Dominissini, Q. Dai, T. Pan and C. He, *Angewandte Chemie (International ed. in English)*, 2015, **54**, 1587-1590.

209. D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio and G. Rechavi, *Nat. Protocols*, 2013, **8**, 176-189.

210. C. Dohno, T. Shibata and K. Nakatani, *Chem. Commun.*, 2010, **46**, 5530-5532.

211. D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M. S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, W. C. Clark, G. Zheng, T. Pan, O. Solomon, E. Eyal, V. Hershkovitz, D. Han, L. C. Doré, N. Amariglio, G. Rechavi and C. He, *Nature*, 2016, **advance online publication**.

212. X. Li, X. Xiong, K. Wang, L. Wang, X. Shu, S. Ma and C. Yi, *Nat. Chem. Biol.*, 2016, **advance online publication**.

213. J. B. Macon and R. Wolfenden, *Biochemistry*, 1968, **7**, 3453-3458.

214. M. Yoshida, Y. Furuichi, Y. Kaziro and T. Ukita, *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis*, 1968, **166**, 636-645.

215. M. Yoshida and T. Ukita, *The Journal of Biochemistry*, 1965, **57**, 818-821.

216. M. Sakurai, T. Yano, H. Kawabata, H. Ueda and T. Suzuki, *Nat. Chem. Biol.*, 2010, **6**, 733-740.

217. J. B. Li, E. Y. Levanon, J.-K. Yoon, J. Aach, B. Xie, E. LeProust, K. Zhang, Y. Gao and G. M. Church, *Science*, 2009, **324**, 1210-1213.

218. M. Li, I. X. Wang and V. G. Cheung, *Science*, 2012, **335**, 1302.

219. M. Li, I. X. Wang, Y. Li, A. Bruzel, A. L. Richards, J. M. Toung and V. G. Cheung, *Science*, 2011, **333**, 53-58.

220. C. L. Kleinman and J. Majewski, *Science*, 2012, **335**, 1302.

221. W. Lin, R. Piskol, M. H. Tan and J. B. Li, *Science*, 2012, **335**, 1302.

222. J. K. Pickrell, Y. Gilad and J. K. Pritchard, *Science*, 2012, **335**, 1302.

223. C.-X. Song, C. Yi and C. He, *Nat. Biotech.*, 2012, **30**, 1107-1116.

224. T. M. Carlile, M. F. Rojas-Duran, B. Zinshteyn, H. Shin, K. M. Bartoli and W. V. Gilbert, *Nature*, 2014, **515**, 143-146.

225. A. F. Lovejoy, D. P. Riordan and P. O. Brown, *PLoS ONE*, 2014, **9**, e110799.

226. P. T. Gilham, *J. Am. Chem. Soc.*, 1962, **84**, 687-688.

227. N. W. Y. Ho and P. T. Gilham, *Biochemistry*, 1967, **6**, 3632-3639.

228. P. T. Gilham and N. W. Y. Ho, *Biochemistry*, 1971, **10**, 3651-3657.

229. A. Bakin and J. Ofengand, in *Protein Synthesis*, ed. R. Martin, Springer New York, 1998, vol. 77, ch. 22, pp. 297-309.

230. A. Bakin and J. Ofengand, *Biochemistry*, 1993, **32**, 9754-9762.
231. J. Ofengand and A. Bakin, *J Mol Biol*, 1997, **266**, 246-268.
232. S. Schwartz, Sudeep D. Agarwala, Maxwell R. Mumbach, M. Jovanovic, P. Mertins, A. Shishkin, Y. Tabach, Tarjei S. Mikkelsen, R. Satija, G. Ruvkun, Steven A. Carr, Eric S. Lander, Gerald R. Fink and A. Regev, *Cell*, 2013, **155**, 1409-1421.
233. X. Li, P. Zhu, S. Ma, J. Song, J. Bai, F. Sun and C. Yi, *Nat. Chem. Biol.*, 2015, **11**, 592-597.
234. H. Cahova, M.-L. Winz, K. Hofer, G. Nubel and A. Jaschke, *Nature*, 2015, **519**, 374-377.