DISS. ETH NO. 26118

Nucleotide-resolution Genome-wide Mapping of Oxidative DNA Damage

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(**Dr. sc. ETH Zurich**)

presented by

Junzhou Wu

M.sc., Peking University, China

born on 04.01.1990

citizen of China

accepted on the recommendation of

Prof. Dr. Shana J. Sturla

Prof. Dr. Maureen McKeague

PD Dr. Jochen Klumpp

Dr. Enni Markkanen

2019

Acknowledgements

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Shana Sturla, for the continuous support of my Ph.D study, for her patience, motivation, and immense knowledge. You gave me the freedom and trust to be an independent scientist, always providing me with the help and support needed throughout these years. I could not have imagined having a better advisor and mentor for my Ph.D study.

A special thanks goes to Prof. Maureen McKeague, Dr. Ioannis A. Trantakis and Dr. Xavier Casadevall i Solvas for mentoring me at different period in the past years. Thank you for always finding the time for me and the priceless scientific, personal, and career related advice.

Many thanks to Dr. Jochen Klumpp and Dr. Enni Markkanen for accepting to be my co-examiner and for your insightful comments, encouragement, and extensive professional guidance.

I would like to thank Prof. Barbara van Loon for critical inputs and our collaboration. Thank you for mentoring my student and inviting me to visit your lab. Also a special thanks to Prof. Matthias Altmeyer. Thanks for the insightful experimental suggestions and supports.

Thanks to Dr. Niklaus Zemp, Dr. Aria Maya Minder Pfyl, Silvia Kobel, Catharine Aquino, Andrea Patrignani, all the Genetic Diversity Centre members and all the Functional Genomics Center Zurich members. Thank you for all the supports on sequencing library preparation and bioinformatics in the past years.

Thanks to all of the Sturla lab members. You all made my time as a PhD student so enjoyable with such an amazing and helpful atmosphere in the group. I would especially like to thank Dr. Michael Räz for being a great office neighbor and colleague. Thank you for helping me to deal with German documents. A special thanks also goes to Dr. Todor Angelov, Dr. Arman Nilforoushan, Dr. Stefano Malvezzi, Dr. Florence Berger. Thank you for all the experimental support in the lab when I started in Sturla

group. Special thanks to all the former and present D19 officemates: Dr. Jianbo Zhang, Claudia Aloisi, Emma Sandell, Jeanne Cresson, Nikolai Püllen and Celine Stäuble. Thank you for all the laughter that we share. Thank you to all the present and former lab members: Dr. Hailey Gahlon, Dr. Gabriele Fontana, Dr. Laura Wyss, Dr. Marianna Stamou, Dr. Alessia Stornetta, Dr. Simon Sieber, Dr..Guido Paul Zeegers, Dr. Sabrina Huber, Dr. Melanie Erzinger, Dr. Lea von Moos, Dr. Seiichiro Kizaki, Dr. Claudia Otto, Dr. Katie Hurley, Dr. Nicole Antczak, Cécile Mingard, Nathalie Ziegler, Susanne Geisen, Aida Huber, Daniela Kalbermatter, Mirjam Schneider, Sabine Senger, Fabrice Müller, Jiang Yang, Alejandro Ramirez, Vanessa Mazotti and Anastasia Stanga.

Many thanks also go to Jasmina Büchel and Raphael Kuhn. You were both great students and it was great to mentor you.

I would also like to thank all my friends: Dr. Henwei Huang, Dr. Famin Qiu, Dr. Bin Huang, Prof. Yizhou Li, Prof. Bing Hu, Dr. Xiaobao Cao, Dr. Feng Shao, Dr. Gang Lu, Dr. Dan Zhu and Dr. Shaopu Wang. I am very thankful for all the great time we had.

I would especially like to thank my wife, Dr. Jingjing Sun. Thank you for moving to Zurich with me, for your understanding and love during the past few years. You have always stood by me and been extremely supportive of me throughout my entire PhD.

最后，我要特别感谢我的爸爸、妈妈 、岳父、岳母、姐姐、姐夫和外甥女。你们一直以来的关心和支持让我能够全身心的投入到学业中。我爱你们。

Table of contents

Acknowledgements i

Abstract v

Zusammenfassung vii

Abbreviation ix

Chapter 1: Introduction 1

- 1. DNA structure and genetic information 3
- 2. DNA lesions 4
- 3. DNA repair 5
- 4. Genomic mutations 7
- 5. DNA damage detection 8
- 6. DNA damage sequencing 10
- 7. 8-Oxo-2'-deoxyguanosine 12
  - 7.1 Occurrence and relevance 12
  - 7.2 Repair, mutagenicity, and toxicity 12
  - 7.3 8-oxoG and its repair enzyme OGG1 modulate gene expression 13
  - 7.4 Sequence-Based 8-oxoG Analysis 15
- 8. Oxidative sugar damage 17
  - 8.1 Occurrence and relevance 17
  - 8.2 Repair, mutagenicity, and toxicity 18
  - 8.3 Detection strategies 18
- 9. Overview of thesis work 19
- References 20

Chapter 2: Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq 27

- Abstract 29
- Introduction 29
  - Labelling and amplifying 8-oxoG in DNA 30
  - 8-oxoG distribution in genomic features 33
  - Sequences flanking 8-oxoG in the genome 35
- Conclusions 35

References 36
Supporting Information 41
Chapter 3: Nucleotide-resolution mapping of DNA oxidation in relation to mutagenesis and chromatin state 67
Abstract 68
Introduction 68
Results 70
DNA oxidation mapping of HAP1 cells by click-code-seq 70
8-oxoG damage signature correlates with cancer mutation signatures 72
Crosstalk between DNA damage and epigenetic histone modification 73
8-oxoG levels are related with gene length but not expression 75
Conclusion 75
References 77
Supporting information 79
Chapter 4: Amplification and sequencing of 5’-aldehyde lesions resulting from DNA oxidation 89
Abstract 90
Introduction 90
Results and Discussion 92
Synthesis of modified oligonucleotides 92
Labelling of aldehyde modified oligonucleotide with the code sequence 93
Bypass and amplification of ligated products raised from aldehyde labelling 93
qPCR quantification of 5’-aldehyde terminus in dsDNA 95
Conclusion 97
References 97
Supporting information 99
Chapter 5 Summary and Outlook 127
Appendix A: Impact of DNA oxidation on toxicology: from quantification to genomics 131
Appendix B: Click-code-seq Protocol 137

Abstract

Human cancers arise from mutations due to endogenous processes and exposure to xenobiotic chemicals. Cellular repair pathways are fantastically evolved to avoid adverse effects of DNA damage, particularly in response to the high abundance of various oxidation adducts. Nonetheless, exposures to an increasing variety and amount of chemicals from environment, diet and drugs, together with defects or deficiencies in repair, add to risks of mutagenesis and carcinogenesis. Current understanding of how DNA oxidation drive mutagenesis is advanced, but our ability to predict the mutagenicity of chemicals or disease risks related to these processes remains limited. There exists a mismatch between our low resolution understanding of how DNA adducts are distributed and dynamically altered on a genome-wide level vs. our sophisticated knowledge of intricate mutational landscapes of human cancers. Therefore, understanding the genome-wide distribution of DNA damage is an important aspect of elucidating mutagenesis and carcinogenesis mechanisms.

Chapter 1 contains the scientific background of topics discussed in this thesis. This chapter covers a brief overview about DNA structure, DNA damage & repair and DNA lesion detection methods. In particularly, next-generation sequencing technology and its applications to DNA lesions are described. Finally, two DNA oxidative lesions, 8-oxoG and 5'-aldehyde terminus are introduced, focusing on their occurrence, biological relevance and detection strategies.

In **Chapter 2**, a new sequencing method, click-code-seq, was developed for labeling and amplifying oxidized DNA bases as a way to locate them in the genome. It involves the incorporation of an oligonucleotide code to mark each position of an oxidized guanine in DNA. The biocompatible code enabled high-throughput, base resolution sequencing of the 8-oxoguanine site. By applying click-code-seq in a eukaryotic (yeast) genome, we uncovered thousands of 8-oxoG sites with features and patterns suggesting a potential relationship to chromatin formation and transcription. Click-code-seq overcomes current challenges in DNA damage sequencing and provides a new approach for generating comprehensive, sequence-specific information about 8-oxoG patterns in whole genomes.

In **Chapter 3**, click-code-seq was further applied to human genome. A single nucleotide-resolution genome-wide map of DNA oxidation was achieved in human haploid cells (HAP1). The results revealed a specific damage pattern which is

correlated with mutation signatures in cancer, suggesting the straightforward process from damage to mutation for the first time. Furthermore, the distribution of oxidative DNA damage varies widely across the genome, with distinct patterns related to chromatin architecture, epigenetic modification, DNA damage response and DNA-protein interaction. Sequencing of 8-oxoG paves a powerful approach for studying biological and toxicological questions surrounding DNA oxidation.

In **Chapter 4**, a new strategy to locate the major 2-deoxyribose oxidation, 5’-aldehyde terminus, at single nucleotide resolution was developed. 5’-aldehyde terminus was labelled with an aminooxy-functionalized oligonucleotide, giving rise to an oxime linked biocompatible DNA. The yield DNA form 5’-aldehyde terminus could be amplified by polymerase chain reaction. Meanwhile, abasic sites could also be labelled, but the produced DNA could not be amplified, supporting that various sites are labelled but only those derived from 5’-aldehyde precursors could be bypassed and amplified by a polymerase. This method was validated with synthetic oligonucleotides and site-specific modified plasmids. The results of this work provide a new strategy for 5’-aldehyde terminus detection.

In **Chapter 5**, the most important results of the doctoral work are summarized and critically evaluated. Limitations of our achievements and future directions on DNA damage sequencing are discussed.

The **Appendix A** contains a mini-review focusing on recent emergent 8-oxoG sequencing methods, major biological findings and outlooks on future studies.

The **Appendix B** contains a step-by-step library preparation protocol of Click-code-seq.

Zusammenfassung

Krebs entsteht im Menschen aufgrund endogener Prozesse und Belastung durch xenobiotische Substanzen. Reparaturmechanismen in Zellen sind hervorragend entwickelt, um nachteilige Effekte durch DNA Schäden zu vermeiden, insbesondere von Addukten durch Oxidationsprozesse. Dennoch trägt die Belastung einer ansteigenden Vielfalt und Menge von Chemikalien aus Umwelt, Ernährung und Medikamenten zusammen mit Schäden oder Mängeln im Reparatursystem zu einem erhöhten Risiko der Mutations- und Krebsentstehung bei. Der aktuelle Forschungsgegenstand, inwiefern Oxidation von DNA eine Rolle in der Mutationsbildung spielt, ist bereits weit fortgeschritten, jedoch sind die Voraussagemöglichkeiten bezüglich des Mutationspotentials von Chemikalien oder diesem Prozess zugehörigen Krankheitsrisiken weiterhin eingeschränkt. Es besteht eine Diskrepanz zwischen unserer nicht sehr akkuraten Analyse der dynamischen Verteilung von DNA Addukten im Genom vs. unserem detailreichen Wissen über die komplexen Verteilungsmuster von Mutationen in menschlichen Tumoren. Um die Mechanismen hinter Mutations- und Tumorbildung besser verstehen zu können, ist daher das Verständnis des genomweiten Auftretens von DNA Schäden ein Aspekt von hoher Relevanz.

In **Kapitel 1** werden die wissenschaftlichen Hintergründe der in dieser Arbeit diskutierten Themen dargestellt. Es beinhaltet einen kurzen Überblick der DNA Struktur, DNA Schäden und Reparatur und Methoden zur Detektion von DNA-Schäden. Insbesondere wird die Technologie des *next-generation sequencing* und deren Anwendungen in Bezug auf DNA Schäden beschrieben. Zuletzt werden zwei oxidative DNA-Schäden, 8-oxoG und 5'-Aldehyd Terminus, mit dem Fokus auf deren Auftreten, biologischer Relevanz und Detektionsstrategien vorgestellt.

In **Kapitel 2** wird eine neue Sequenzierungsmethode, *click-code-seq*, eingeführt, die für die Markierung und Vervielfältigung von oxidierten DNA Basen verwendet werden kann, um diese im Genom zu lokalisieren. Die Methode beinhaltet das Einfügen eines Oligonukleotidcodes, welcher jede Position eines oxidierten Guanins in der DNA markiert. Der biokompatible Code ermöglicht die Hochdurchsatzsequenzierung von 8-Oxoguanin-Positionen mit Einzelnukleotidauflösung. Durch das Anwenden von *click-code-seq* haben wir tausende 8-oxoG entdeckt, welche Merkmale und Muster aufzeigen, die auf eine potentielle Verbindung zur Chromatinbildung und Transkription hindeuten. *Click*-code-seq überwindet aktuelle Herausforderungen der Sequenzierung

von DNA-Schäden und bietet eine neue Möglichkeit, um umfassende Sequenz-spezifische Informationen über 8-oxoG in gesamten Genomen zu generieren.

In **Kapitel 3** wurde *click-code-seq* auf das menschliche Genom angewandt. Eine genomweite Karte auf Nukleotidebene von DNA Oxidationsschäden konnte von menschlichen haploiden Zellen (HAP1) erstellt werden. Die Ergebnisse konnten ein spezifisches Schadensmuster aufdecken, das mit Mutationssignaturen in Tumoren korreliert, was erstmalig eine direkte Beziehung zwischen Schäden und Mutationen vermuten lässt. Weiterhin variiert die Verteilung von oxidativen DNA Schäden im Genom mit Mustern, die in Verbindung mit der Chromatinstruktur, epigenetischen Modifikationen, Reaktionen auf DNA-Schäden und DNA-Protein Interaktionen stehen. Die Sequenzierung von 8-oxoG bietet eine leistungsstarke Methode, um biologische und toxikologische Fragestellungen bezüglich DNA Oxidation anzugehen.

Kapitel 4 beschreibt eine neue Strategie, um das Hauptoxidationsprodukt von 2-Deoxyribose, dem 5'-Aldehyd Terminus, auf Einzelnukleotidebene zu lokalisieren. Der 5'-Aldehyd Terminus wurde mit einem Aminooxy-funktionalisiertem Oligonukleotid markiert, was in einer oxim-verknüpften biokompatiblen DNA resultierte. Die Menge an DNA mit einem 5'-Aldehyd Terminus konnte durch Polymerase-Kettenreaktion amplifiziert werden. Abasische Stellen hätten ebenfalls markiert, jedoch die resultierende DNA nicht amplifiziert werden können, wodurch zwar verschiedene Positionen markiert, aber nur die von 5'-Aldehyd stammenden Produkte von einer Polymerase gelesen und amplifiziert werden können. Die Methode wurde mit Hilfe von synthetischen Oligonukleotiden und einem positionsspezifisch modifizierten Plasmid validiert. Die Ergebnisse dieser Arbeit bieten eine neue Strategie zur 5'-Aldehyd Terminusdetektion

In **Kapitel 5** werden die wichtigsten Ergebnisse der Doktorarbeit zusammengefasst und kritisch bewertet. Die Begrenzungen und zukünftige Richtungen unserer Arbeit auf dem Gebiet der Sequenzierung von DNA-Schäden werden diskutiert.

Anhang A enthält ein Mini-Review, das auf kürzlich erschienene 8-oxoG Sequenzierungsmethoden sowie die wichtigsten biologischen Erkenntnisse und Perspektiven eingeht.

Anhang B enthält ein Schritt-für-Schritt Protokoll für das Erstellen einer DNA *library* mit *click-code-seq*.

Abbreviation

•OH	hydroxyl radical
3MeA	3-methyladenine
5fC	5-formyl-2’-deoxycytidine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
6-4PPs	6-4 photoproducts
7MeG	7-methylguanine
8-oxo-dA	8-oxo-2’-deoxyadenosine
8-oxoG	8-oxo-2’-deoxyguanosine
A	adenine
AAG	alkyladenine DNA glycosylase
Adap	adenosine-1,3-diazaphenoxazine
AP	abasic site
APE1	AP-endonuclease
ARID1A	AT-rich interaction domain 1A
ARP	aldehyde reactive probe
ARS	autonomously replicating sequences
ATF4	activating transcription factor 4
Au-NP	gold nanoparticle
BER	base-excision repair
BPDE-dG	BaP diol epoxide-deoxyguanosine
C	cytosine
CPDs	cyclobutane-pyrimidine dimers
CRC	colorectal cancer
CTCF	transcription repressor CTCF
DDR	DNA damage response
dGh	guanidinohydantoin 2’-deoxynucleoside
DHS	DNase I hypersensitive sites
DMT	dimethoxytrityl
DSB	double strand breaks
ELISA	enzyme-linked immunosorbent assay

FANC	Fanconi anaemia complementation group
fapy-dG	2,6-diamino-4-hydroxy-5-formamidopyrimidine-2’-deoxynucleoside
FISH	fluorescence in situ hybridization
G	guanine
G4	G-quadruplex-forming sequences
GEO	Gene Expression Omnibus
HAP1	human haploid cells
HGP	Human genome project
HIF-1a	hypoxia-inducible factor 1a
HR	homologous recombination
IP	immunoprecipitation
LC-MS/MS	liquid chromatography-tandem mass spectrometry
LM-PCR	ligation-mediated polymerase chain reaction
LPS	lipopolysaccharide
MAP	MUTYH-associated polyposis
MEFs	mouse embryonic fibroblasts
MMR	mismatch repair
NER	nucleotide-excision repair
NF-kB	nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next-generation sequencing
NHEJ	non-homologous end joining
$O_2^{\bullet -}$	superoxide anion radical
O^6-BnG	O^6-benzylguanine
O^6-CMG	O^6-carboxymethylguanine
O^6-MeG	O^6-methylguanine
ODN	oligodeoxynucleotide
OGG1	8-oxoguanine glycosylase
PAECs	Rat pulmonary arterial endothelial cells
PAGE	polyacrylamide gel electrophoresis
PARP1	poly(ADP-ribose) polymerase 1
Pol II	phosphorylated-RNA polymerase II

Pols	polymerases
prop-dGTP	*O*-3'-propargyl modified nucleotide
RNA polII	RNA polymerase II
ROS	reactive oxygen species
SBS	sequencing-by-synthesis
SCGE	single cell gel electrophoresis
SMRT	single-molecule real-time
SOD	superoxide dismutase
SP1	specificity protein 1
SSBR	single strand breaks repair
SSBs	single strand breaks
STAT1	signal transducer and activator of transcription 1
T	thymine
TF IID	transcription initiation factor II-D
TFs	transcription factors
TGC	trinucleotide context
TLS	translesion DNA synthesis
TM	transcriptional mutagenesis
TSS	transcription start sites
TTS	transcription terminator sites
TUNEL	terminal deoxynucleotidyl transferase dUTP nick end labeling
UNG	uracil DNA glycosylase
VEGF	vascular endothelial growth factor
WT	wild type

Chapter 1: Introduction

Chapter 1

1. DNA structure and genetic information

Deoxyribonucleic acid (DNA) is essential to the basic cellular functions of all living organism on Earth, recording all necessary information for building and maintaining that organism. The information in DNA is stored as a code composed of the four nucleobases: adenine (A), guanine (G), cytosine (C), and thymine (T).[1] These four nucleobases are polymerized together via a sugar-phosphate backbone to form a DNA sequence. DNA bases are paired up through Watson-Crick hydrogen bonds with each other (A with T, G with C), resulting in two antiparallel long strands that form a spiral double helical structure (Figure 1). Hundreds to several millions of DNA base pairs constitute a gene as the basic functional unit of DNA. A genome consisting of all DNA sequence/genes in an organism is well known to vary enormously in size: the *S. cerevisiae* genome is composed of about 12 million base-pairs (bps) and around 6,000 genes, the human genome has 3.2 billion bps and more than 20,000 genes.[2-3] The decoding of this genetic information in an organism follows the central dogma of molecular biology including transcription from DNA to RNA and translation from mRNA to protein, by which the information flows from genes into functional proteins.[1] DNA damage and mutation threaten the faithful delivery of genetic information to offspring. Meanwhile, the genetic information could be decoded by determining the order of the four bases using DNA sequencing technologies, including DNA damage, chemical

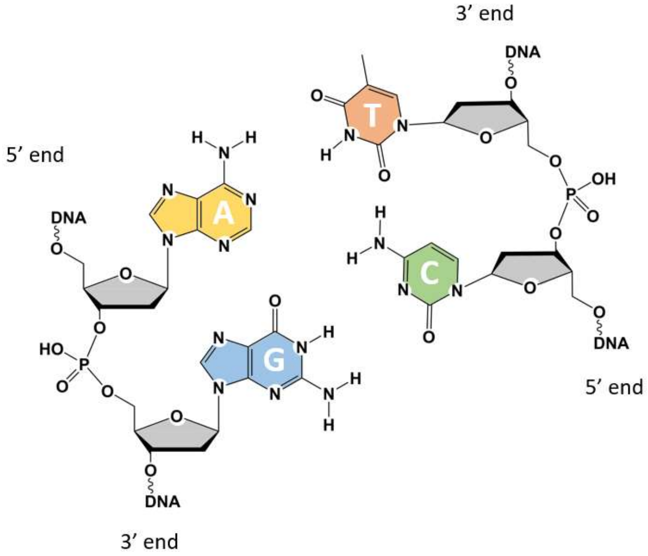


Figure 1. Chemical structure of DNA. DNA molecule is composed of a deoxyribose sugar molecule (grey) to which is attached a phosphate group and one of four bases: A (yellow), T (orange), C (green), G (blue).

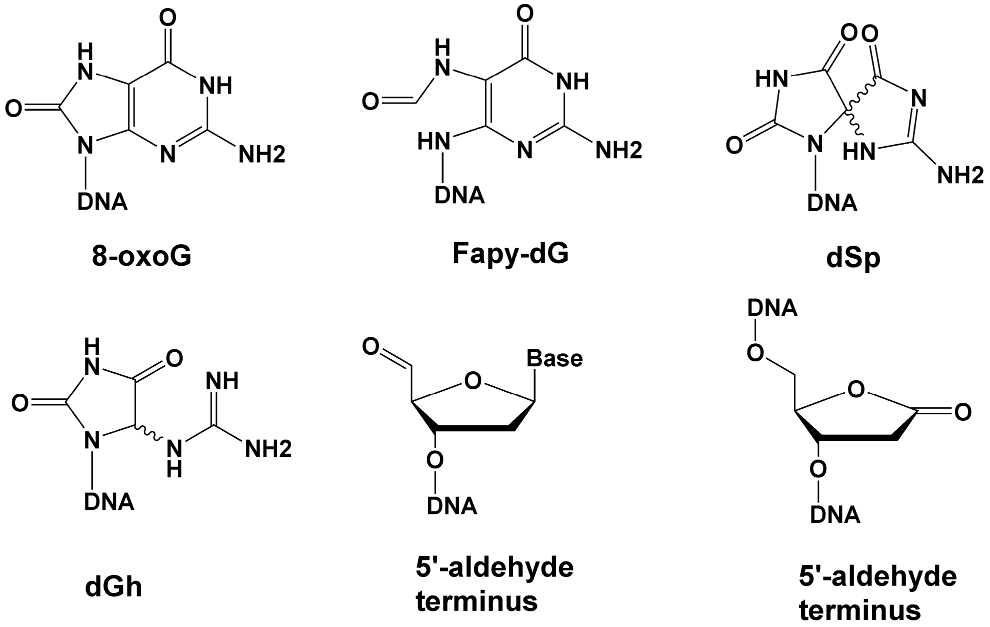
modification and mutation. Advances in understanding the genetic formation behind DNA damage and mutation enable the development of early diagnosis and new treatments.[4]

2. DNA lesions

The reliable transmission of genetic information lies in the sequence and chemical structures of the nucleobases. However, the chemical structure of DNA is not stable. DNA lesions form by the attack of endogenous and exogenous chemicals, such as ionizing radiation, heavy metal ions, reactive oxygen species (ROS), ultraviolet irradiation and alkylating agents.[5] Each healthy human cell is subject to approximately 70,000 endogenous lesions per day, including 55,000 single strand breaks (SSBs), 12,000 abasic sites, 2,800 8-oxoG, 200 cytosine deamination and 25 double strand breaks (DSB).[6] At the cellular level, unrepaired DNA lesions can lead to transient arrest, genomic instability, apoptosis, or senescence as acute effects and permanent DNA mutations or chromosome aberrations as long-term consequences. More importantly, these cellular level disorders predispose the organism to immunodeficiency, neurological disorders and cancer.[7-9] Therefore, it is essential to understand chemical and biological process of DNA lesions as the basic of disease development and aging process.

The mitochondrial respiratory chain is a major source of ROS, leading to the formation of superoxide anion radical ($O^{2\bullet-}$). The proximal superoxide is then transformed to hydrogen peroxide (H_2O_2) by superoxide dismutase (SOD), which is further converted to hydroxyl radical ($^{\bullet}OH$) via Fenton reaction with transition metal ions.[10] Hydroxyl radical can directly react with nucleobases and DNA backbone, yielding a variety of nucleobase lesions and SSBs. Common oxidative nucleobase lesions include 8-oxo-2'-deoxyguanosine (8-oxoG), 8-oxo-2'-deoxyadenosine (8-oxo-dA), 2,6-diamino-4-hydroxy-5-formamidopyrimidine-2'-deoxynucleoside (fapy-dG), guanidinohydantoin 2'-deoxynucleoside (dGh) and spiroiminodihydantoin 2'-deoxynucleoside (dSp) (Figure 2).[10] Oxidation of deoxyribose can also occur at each of the five positions under most biologically relevant conditions, yielding 2' deoxyribonolactone (1'), erythrose abasic site (2'), 3'-keto-2'-deoxynucleotide (3'), 2-deoxypentos-4-ulose abasic site (4'), nucleotide-5'-aldehyde (5') and so on.[11]

Apart from ROS, UV irradiation and alkylating agents damage DNA significantly. Common UV-induced lesions include cyclobutane-pyrimidine dimers (CPDs) and 6-4 photoproducts (6-4PPs).[12] Alkylating agents could attack multiple nucleophilic sites in nucleobases, yielding 3-methyladenine (3MeA), 7-methylguanine (7MeG), O^6-methylguanine (O^6-MeG), benzo[α]pyrene adducts and cisplatin adducts.[13] Accumulation of DNA lesions could lead to deleterious biological consequences, thus, DNA repair, an efficient defense system, is involved to remove DNA lesions in cell.


8-oxoG

Fapy-dG

dSp

dGh

5'-aldehyde terminus

5'-aldehyde terminus

Figure 2. Structure of common DNA lesions.

3. DNA repair

To survive from the deleterious effects of DNA lesions, a complex network of mechanisms have been evolved to remove DNA lesions from the genome, such as base-excision repair (BER), nucleotide-excision repair (NER), mismatch repair (MMR), homologous recombination (HR) and non-homologous end joining (NHEJ) (Figure 3).[8] The link between DNA lesions and human diseases is long established by studying syndromes arisen from DNA repair-deficiency patients. Defects in DNA repair can lead to accelerated aging, cancer and neurological diseases. For example, congenital defects in uracil DNA glycosylase (UNG) which is a BER associated protein, would result in hyper-IgM syndrome; Mutations in any of at least sixteen Fanconi anaemia complementation group (FANC) family genes cause Fanconi anemia, a disorder characterized by sensitivity to DNA interstrand crosslinks and susceptibility to tumor formation.[14] However, detailed processes from DNA lesions to diseases are still not well understood because the lack of our knowledge on genome-wide distribution of DNA lesions.

In particular, the BER pathway corrects high levels of spontaneous DNA lesions that may result from normal metabolic processes, such as oxidation, deamination and alkylation. The BER pathway is initiated by a family of enzymes called DNA golycosylases that cleave the N-glycosidic bond of DNA lesions. There are at least 11 DNA glycosylases assigned to BER, and each one is responsible for the recognition and removal of a subgroup of DNA lesions, such as 8-oxoguanine glycosylase (OGG1) for 8-oxoG, fapy-dG, fapy-dA; alkyladenine DNA glycosylase (AAG) for 3/7-MeA, 3/7-MeG (Table 1).[15] Most DNA golycosylases employ a base flipping mechanism to sequester damaged nucleobases from dsDNA into an active site pocket and then remove the flipped out base, leaving an abasic site (AP). The AP site could be removed by a major AP-endonuclease (APE1) in mammalian cells or bifunctional glycosylases. Either by a two or one enzyme process, the result is a one nucleotide gap. Completion of BER is accomplished by several additional enzymes responsible for end processing, repair synthesis and ligation.[16] Together with BER, DNA repair pathways could remove

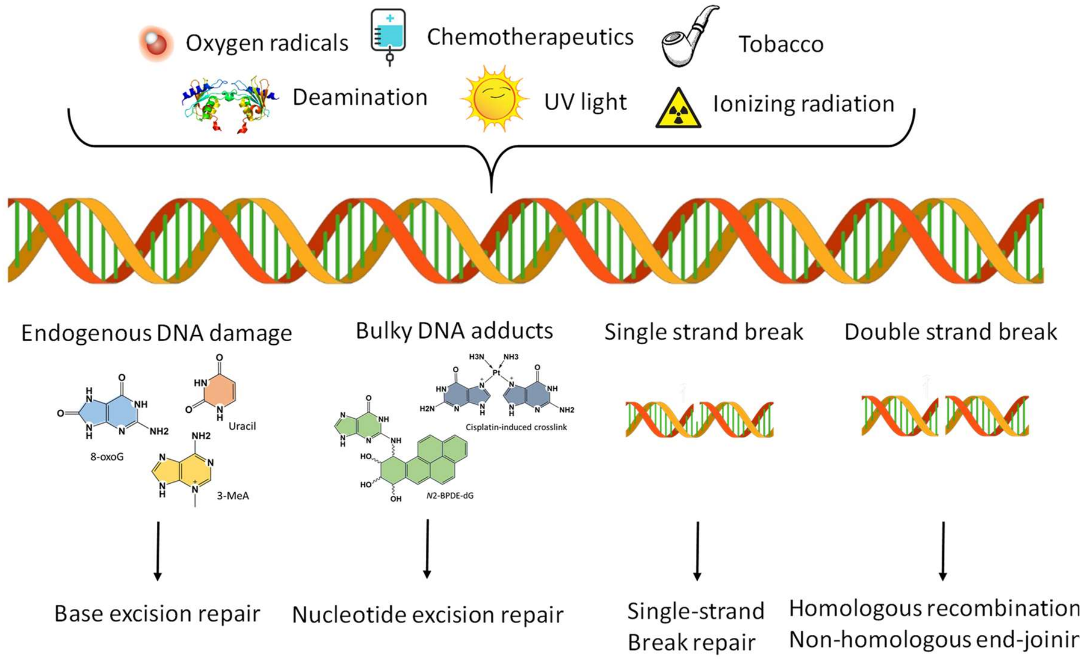


Figure 3. Type of DNA damage resources, DNA damage and repair pathways.

majority of DNA lesions, however, unrepaired DNA lesions could lead to threaten consequences, such as genomic mutations.

4. Genomic mutations

If left unrepaired, damaged dsDNA may be used as a template during DNA replication. The process of copying a template containg a lesion is called translesion DNA synthesis (TLS) and is mediated by several specialized low-fidelity TLS polymerases (Pols), including Y-family Pol η, ι, κ, and Rev1 and B-family Pol ζ.[17] TLS Pols either insert a correct nucleotide opposite a DNA lesion by an error-free pathway or insert a mismatched nucleotide by an error-prone pathway. With error-prone TLS, a transitory DNA lesion is converted to a reproducible DNA mutation. The accumulation of mutations in genomic DNA is believed to contribute to the development of cancer, neurological diseases and aging. In particular, the link between cancer and DNA damage is more clearly defined. Several types of cancers are linked to exogenous carcinogen exposure via this genotoxic mechanism, including sunlight-associated skin cancers,[18] tobacco-associated lung cancers,[19] and aristolochic acid-related urothelial tumors.[20] The exposure to these carcinogens has been very well-studied, and in all cases result in an increased abundance of specific DNA lesions, thus supporting the role of DNA lesions in the initiation of carcinogenesis.[21]

Exposure to environmental mutagens is linked to special mutation pattern called mutation signature. Several mutation signatures induced by DNA oxidation are

Glycosylase	Substrates
OGG1	8oxoG, fapyG, fapyA
MYH	A:G, A:C and A:8oxoG mismatches
NTH1	Tg, Cg, fapyG, DHU, 5-ohU, 5ohC
NEIL1	Tg, 5ohC, fapyA, fapyG, Urea, 8oxoG
NEIL2	NTH1 plus NEIL1 substrates
NEIL3	Oxopurines, fapyG, fapyA
AAG/MPG	3meA,7meA,3meG,7meG, εA
UNG	U, U:G, U:A, 5-FU
TDG	T, U, 5-FU, εC, 5-hmU, 5-fC, 5-caC:G
SMUG1	U, 5-hmU, 5-FU, U:A, U:G
MBD4	U or T in U/TpG:5meCpC

Table 1. Oxidation, alkylation and deamination substrate scope of mammalian glycosylases.

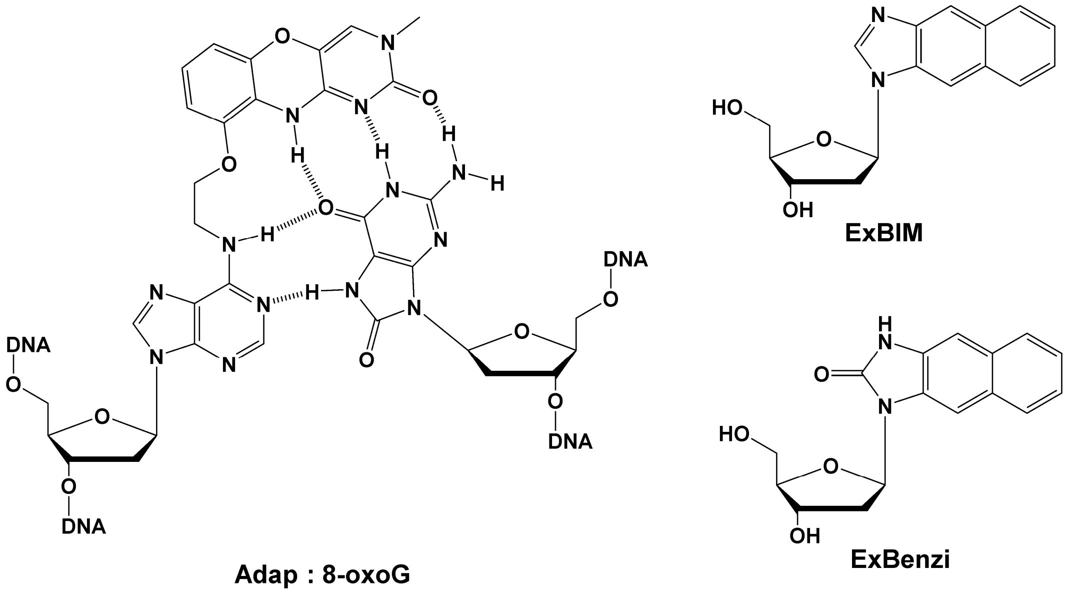
reported. A mutation signature has been extracted from MUTYH-deficient colorectal cancer patients,[22] and the TOY-KO mouse deficient in the three main means of repairing 8-oxoG described above yield a germ-cell-line mutation signature similar to COSMIC signature 18, which is dominated by G>T transversions, particularly in the GCA trinucleotide context (TGC).[23] Although Signature 18 is thought to be caused by oxidative stress the driver mechanisms in patients harboring this signature are not clear. Also of interest, Signature 17, from oesophagus, breast, liver, lung, B-cell, stomach and melanoma, could be the consequence of damaged guanine single nucleotide pool because of its almost single T>G base substitutions.[24] In contrast to our ability to define outcomes of DNA lesions (i.e., mutation, cytotoxicity, carcinogenesis), there is a lag in our capacity to detect DNA damage in the genome.

5. DNA damage detection

A comprehensive understanding about the adverse biological outcomes of DNA lesions requires the capacity to characterize their occurrence and repair throughout genome. There are many well-established strategies for DNA lesion analysis integrated over the whole genome or in particular locations, including ligation-mediated polymerase chain reaction (LM-PCR),[25] single cell gel electrophoresis (SCGE; comet assay),[26] liquid chromatography-tandem mass spectrometry (LC-MS/MS),[21] fluorescence *in situ* hybridization (FISH),[27] terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL) assay,[28] enzyme-linked immunosorbent assay (ELISA)[29] and so on. Recently, DNA adductomics or untargeted DNA adduct profiling based on mass spectrometry screening of nucleoside adducts and their fragmentations (MS/MS) are emerged for the simultaneous detection of hundreds of DNA lesions in genomic DNA.[30] Another significant improvement has taken place with the classical comet assay, a gel electrophoresis assay detects single or double strand breaks. The traditional slide-based comet assay has limitations in reproducibility and throughput, but utilizing microchip technology, a Comet-Chip platform was reported with 200 times increased capacity and excellent reproducibility over the traditional comet assay.[31] However, all these methods could only determine total amount of DNA damage in genomic DNA without location information.

Over the past 10 years, several new methods have been created to determine the exact sequence and position of DNA damage. In particular, several adduct-directed synthetic nucleosides have been developed to form more stable base pairs with DNA lesions than canonical nucleobases, including pyrene nucleoside and 5-napthyl-indolyl nucleoside for abasic site,[32-33] adenosine-1,3-diazaphenoxazine (Adap) derivative for 8-oxoG,[34] and ExBenzi/ExBIM for O^6-alkylG.[35] Two general strategies were used to detect DNA lesions using adduct-directed synthetic nucleosides, namely selective hybridization probe and synthetic nucleoside triphosphate.

The principle of hybridization probe is based on that synthetic nucleoside modified oligonucleotide which shows higher stability or dramatic fluorescent changes when binding to specific DNA lesion containing oligonucleotide. For example, oligonucleotide containing Adap exhibited selective fluorescence quenching when duplexed with 8-oxoG modified oligonucleotide.[34] In our lab, a novel hybridization-mediated aggregation of gold nanoparticle (Au-NP) was designed to detect O^6-MeG with ExBenzi and ExBIM modified oligonucleotides attached. The selective stabilizing effect between ExBenzi/ExBIM and O^6-MeG led to the conjugation of ExBenzi/ExBIM modified Au-NP and O^6-MeG modified Au-NP, resulting in a color change from Au-NP suspension in red to Au-NP aggregation in purple. This Au-NP strategy allowed visual detection of O^6-MeG in the human KRAS gene.[35] However, due to extremely low abundance of DNA lesions in the genome, the amplification-free hybridization probes are not sensitive enough to detect DNA lesions in a specific location at genome-wide level.



Adap : 8-oxoG

ExBIM

ExBenzi

Figure 4. Sturcture of Adap, ExBIM and ExBenzi.

Thus, synthetic nucleoside triphosphates, such as BIMTP and BenziTP, were synthesized and used for polymerase amplification of O^6-MeG to increase sensitivity. A KlenTaq mutant KTqM747K was used to incorporate synthetic triphosphates opposite O^6-MeG more frequently than natural dNTPs. These triphosphates also showed good selectivity to several other O^6-alkylG adducts, such as O^6-benzylguanine (O^6-BnG) and O^6-carboxymethylguanine (O^6-CMG).[36-38] The successful full-length extensions of O^6-alkylG adducts allowed the linear amplification of these adducts by iteratively repeating primer-extension steps.[39] However, compare to exponential amplification during PCR, linear amplification is not power enough to get sufficient signal for detection from genomic DNA. A significant improvement in this area will be a specific base pairing partner for synthetic nucleoside as a third base pair to achieve exponential amplification of a DNA lesion in a gene. To our knowledge, until now, no molecular biology tool could detect a specific DNA lesion in a gene.

6. DNA damage sequencing

Detection of DNA damage in genome is now possible because of the advances with next-generation sequencing (NGS) technology. In the past 40 years, we witnessed the fantastic revolutions of DNA sequencing technologies: from a few hundreds of bases to the first human genome, and now to billions/trillions of bases in one run. In 1976, Sanger and Coulson reported a method to decode hundreds of bases in one sequencing run, involving four extensions of a labelled primer by DNA polymerase with small amounts of a chain-terminating nucleotide in each reaction.[40] In the 1990s, modern Sanger sequencing (first generation sequencing technology) was achieved in an automated, fluorescence-based sequencing machine, using four colour terminators and capillary electrophoresis.[41] The automated Sanger sequencing was used for the Human genome project (HGP) to get the first human genome and is still widely used in molecular clone and variant detection until now.[42-43] After the HGP completion, next generation sequencing (NGS) also known as massively parallel sequencing began to rise with pyrosequencing technology by 454 Life Sciences in 2005 as the first example, followed by SOLiD, Helicos and Illumina platforms.[44] In principle, the concepts behind all second-generation sequencing platforms are similar. First, the dense multiplexing templates for sequencing are prepared by clonal *in vitro* amplification of millions to billions of immobilized templates. Then, all the template clusters are sequenced at the same time by fluorescent microscopy during polymerase extension or DNA ligation, also known as sequencing-by-synthesis (SBS).[45-46] Since 2012, the Illumina platform is dominant because of the extremely high throughput and low cost per Mb. All of the aforementioned first- and second-generation platforms require template amplification during library preparation or at least during clusters generation. This will lead to the loss of chemical modification information, such as DNA damage and methylation. More recently, two new third-generation sequencing platforms given rise to amplification free sequencing, namely single-molecule real-time (SMRT) sequencing and Oxford nanopore sequencing.[47-48] Besides amplification free, long read length is another major advantage of third generation sequencing platforms over others, alleviating numerous computational challenges during genome assembly, transcript reconstruction and metagenomics.[49]

Together with the evolutions of sequencing platforms, a number of modified library construction methods have been developed to achieve special genomic sequencing purposes, including low-frequency chemical modifications to nucleotides. There are two major challenges for DNA lesion sequencing. The first is that these events are very rare and often present at a frequency similar to background mutations of sequencing methods. The second challenge is that chemical modifications cannot be read by polymerases: either the polymerase stalls or inserts an incorrect nucleotide opposite the lesion. As a result, the chemical information is lost. Emerging methods have addressed these obstacles by several different strategies to sequence these rare events specifically and sensitively, including antibody enrichment, polymerases stalling, repair protein excision and chemical direct conversion/labelling, along or as a combination.

By immunoprecipitation (IP) with lesion-specific antibody, lesion-containing DNA fragments are isolated and could be amplified and sequenced directly with several hundred bases resolution, such as 8-oxoG,[50] and acrolein-dA.[51] For bulky lesions that could block the processivity of DNA polymerases, IP could be further combined with polymerase stalling to achieve single nucleotide resolution mapping of specific lesion, such as Damage-Seq and Cisplatin-Seq for Cisplatin-adducts,[52-53] HS-Damage-Seq for UV-induced damage.[54] During DNA damage repair, bulky adduct is removed as a short single-stranded DNA by NER pathway, which could be used as an alternative way than IP to enrich lesion containing DNA fragments, such as XR-Seq for UV damage,[55-56] and BaP diol epoxide-deoxyguanosine (BPDE-dG).[57] DNA glycosylases as the core proteins in BER pathway could recognise and remove DNA lesions specifically, leading to markable free ends. Glycosylase based library preparation methods have led to the successful genome-wide mapping of 8-oxoG,[58] ribonucleotides,[59-62] uracil,[63] cyclobutane pyrimidine dimers (CPDs),[63-64] and alkylation DNA damage.[65] Besides these enzyme/antibody based methods, some of modified nucleotides could be labelled or converted directly by chemical reactions, such as the well-known bisulfite sequencing for cytosine methylation,[66] bisulfite-free methods for 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC) and 5-formylcytosine (5fC),[67-68] abasic site,[69] and 8-oxoG.[70] With the power of newly developed genome-wide approaches, we are able to take first steps to gain a better understanding of DNA damage and repair in a genome scale.

7. 8-Oxo-2'-deoxyguanosine

7.1 Occurrence and relevance

Guanine has the lowest redox potential of the DNA bases, and thus can be easily oxidized to form 8-oxo-7,8-dihydroguanine (8-oxoG) via a one electron transfer reaction mediated by hydroxyl radical.[71] Hydroxyl radicals are generated as by-products of normal metabolic processes or a consequence of exposure to environmental pro-oxidants, such as components of cigarette smoke, alcohol, ionizing and UV radiation, pesticides, and ozone.[72] Normally, the production and scavenging of ROS are well balanced by highly coordinated cellular antioxidant networks, which are essential for cell signaling and homeostasis.[73] However, even under typical physiological ROS levels, 8-oxoG is generated at a frequency of at least several hundred adducts per human cell per day; this rate is further increased under conditions of oxidative overload.[74] Oxidative stress contributes to cancer, atherosclerosis, diabetes, aging, and pathologies of the central nervous system,[75-77] making 8-oxoG an intensely cellular biomarker of pathophysiological processes and an indicator of oxidative stress.

7.2 Repair, mutagenicity, and toxicity

Efficient search and removal of 8-oxoG to maintain cell integrity is performed by the base excision repair (BER) pathway (Figure 5). Three different enzymes cooperate to handle 8-oxoG in the cell, involving a MutT family enzyme (human: MTH1; *S. cerevisiae*: PCD1; *E. coli*: MutT), 8-oxoguanine glycosylase (human: OGG1; *S. cerevisiae*: Ogg1; *E. coli*: Fpg/MutM), and a MutY family glycosylase (human: MUTYH; *S. cerevisiae*: not present; *E. coli*: MutY).[78-80] MutT is an 8-oxo-dGTPase that prevents the incorporation of 8-oxoG into nascent DNA. 8-Oxoguanine glycosylase can directly remove 8-oxoG when paired with cytosine, but if not repaired, 8-oxoG can pair with adenine on the hoogsteen face during replication.[81] In this case, a MutY family glycosylase will excise the incorporated adenine to prevent 8-oxoguanine-related mutagenesis. When all processes are overwhelmed and 8-oxoG persists during replication, it is prone to lead to G:C to T:A transversion mutations. In addition to mutagenesis, unrepaired 8-oxoG can arrest transcription significantly by direct structural interference with transcription components or the repair intermediate of 8-oxoG/OGG1.[82] When 8-oxoG is located on the transcribed DNA strand, other consequences like erroneous bypass of the lesion by transcribing RNA polymerase may occur, termed transcriptional mutagenesis (TM). TM often results in a specific C →A mutation in the RNA transcript and aberrant protein production,[83] which may play a role in protein aggregation and the pathogenesis of neurodegenerative diseases, such as Alzheimer's and Parkinson's diseases.[84]

7.3 8-oxoG and its repair enzyme OGG1 modulate gene expression

Despite the toxicological aspects of 8-oxoG, mounting evidence supports that 8-oxoG may be a cellular friend by facilitating gene activation in response to oxidative stress, countering conventional models of DNA damage effects. The molecular mechanism at the base of 8-oxoG induced gene expression involves several pathways, including

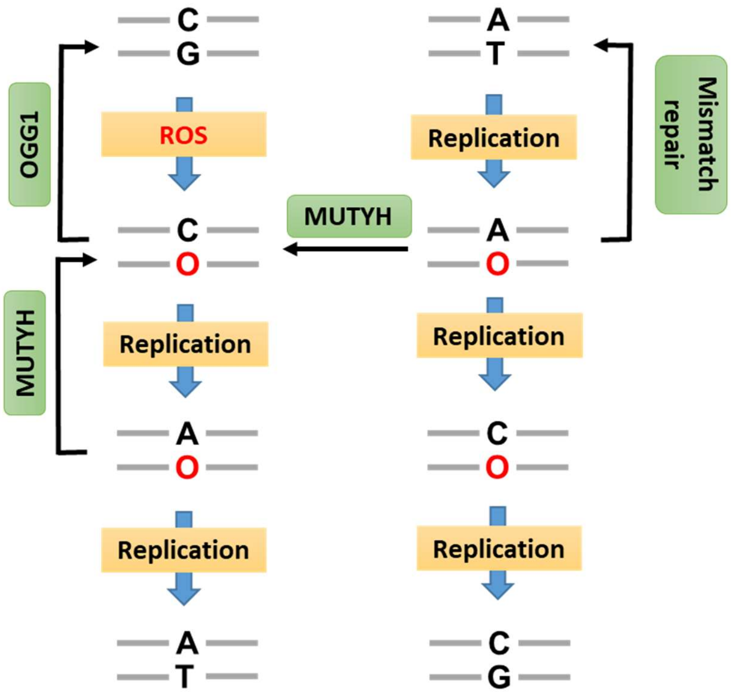


Figure 5. DNA mutation caused by 8-oxoG and its major repair systems. 8-oxoG (O) accumulates in DNA via direct oxidation of DNA or the incorporation of 8-oxo-dGTP. This increases the occurrence of A:T to C:G or G:C to T:A transversion mutations.

direct interactions of OGG1 with transcription factors (TFs) or chromatin remodelers, allosteric transition of G-quadruplex and signal transduction by post-repair OGG1·8-oxoG complex.

When 8-OxoG is located at promoter regions, OGG1 will be recruited to these regions and then enhanced the binding of several TFs to these regions, including hypoxia-inducible factor 1a (HIF-1a),[85] signal transducer and activator of transcription 1 (STAT1)[86] and nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB).[87-88] OGG1 expression decrease in Rat pulmonary arterial endothelial cells (PAECs) strongly reduced the binding of the transcription factor HIF-1α to the vascular endothelial growth factor (VEGF) promoter and reduced VEGF mRNA expression.[85] OGG1 can also act as a coactivator of STAT1 by binding it and induce the transcriptional activation of pro-inflammatory mediators after lipopolysaccharide (LPS) stimulation.[86] In addition, the binding of OGG1 to 8-oxoG in promoter regions enhanced NF-κB/RelA binding to cis-elements and facilitated recruitment of specificity protein 1 (SP1), transcription initiation factor II-D (TF IID) and phosphorylated-RNA polymerase II (Pol II) rapidly, resulting in prompt gene expression upon oxidative exposure.[87-88]

Aside from the interactions between 8-oxoG and TFs, 8-oxoG in DNA G-quadruplex structures can directly up-regulate downstream genes during hypoxia-induced transcription. For example, the VEGF gene harbors three G-rich promoter elements that could adopt a parallel-stranded G4 structure.[89] Binding of SP1 transcription factor to this structure is critical for regulating mRNA synthesis.[90] Upon 8-oxoG accumulated during hypoxia exposure, Sp1 binding was decreased in these G-rich elements. Thus, VEGF transcription was upregulated.[85, 91] These observations highlight the possibility of G4-formation being an activator of transcription activation, particularly when 8-oxoG is present. Recently, Burrows et al. reported that plasmid with 8-oxoG at G4 promoter region produced 2.5-fold more luciferase protein than that same plasmid without 8-oxoG.[92] The data suggest that 8-oxoG in G-rich regions of the VEGF promoter was removed by OGG1, generating an abasic site and destabilizing the G4 structure. This loss of stability led to the formation of a new G4 structure with a fifth G track, and looped out the G track containing the abasic site. This new G4 structure facilitated the binding of APE1 to the abasic site and further stimulated transcription factor binding and activating transcription.[92-93] Collectively, all these findings suggest that 8-oxoG facilitates activation of protective genes in response to oxidative stress, against the conventional knowledge of DNA damage and opening up new and exciting research possibilities. However, there exists a mismatch between our low resolution understanding of how 8-oxoG are distributed and dynamically altered on a genome-wide level vs. our sophisticated knowledge of intricate gene expression changes. System biology-based models linking 8-oxoG location with gene transcription activation are fundamentally limited by this lack of information and methods to map 8-oxoG in a genome wide level.

7.4 Sequence-based 8-oxoG analysis

Given the extremely high biological and health relevance of 8-oxoG, it is highly desired to understand factors that influence the persistence of 8-oxoG in the genome, clarify signatures definitively arising from persistent 8-oxoG and how 8-oxoG modulates gene expression. However, our low resolution understanding of how 8-oxoG are distributed and dynamically altered on a genome-wide level impedes new insights into these questions. Initial strategies to map 8-oxoG involved enrichment by pull-down of fragmented sequences containing 8-oxoG through the use of an 8-oxoG-specific antibody. The antibody-based strategy is analogous to ChIP-seq, where antibodies are used to selectively enrich for DNA sequences bound by a particular protein to map global protein-binding sites in cells. The first genome-wide map of 8-oxoG was constructed using an 8-oxoG-specific monoclonal antibody and immunofluorescence microscopy, resulting in a map of 8-oxoG on human metaphase chromosomes at 1,000 kb resolution from cultured lymphocyte cells.[94] The results revealed 8-oxoG to be unevenly distributed in the human genome, suggesting it as a major cause of SNPs and recombination. The resolution limit of optical microscopy restricted, however, the resolution of the 8-oxoG map and further analysis.

By coupling antibody enrichment and Sanger sequencing, a higher resolution map of 8-oxoG was achieved with mouse renal cortical samples.[51] The results revealed that the distribution of 8-oxoG differed in terms of chromosomes, gene size, and expression, was preferentially formed in highly expressed genes.[51] However, due to the limited throughput of Sanger sequencing, the resulting map only revealed several hundreds of 8-oxoG sites in the mouse genome.

By combining immunoprecipitation with antibodies and microarray analysis, genome-wide mapping of 8-oxoG in normal rat kidney was achieved with higher throughput (244,000 probes) and better resolution (6 kb) than previous efforts.[95] These data revealed that 8-oxoG preferentially located at rat gene deserts which are devoid of protein-coding genes and correlated with lamina-associated domains.[95] The results are partly opposite to the Sanger sequencing results of mouse renal cortical samples, where 8-oxoG preferentially formed in highly expressed genes. These conflicting results reflect the value of a high sensitive and specific method for 8-oxoG sequencing.

Recently, several library preparation methods were reported to afford new opportunities for genome-wide mapping of 8-oxoG with next generation sequencing techniques. OxiDIP-seq, a methodology that combined immunoprecipitation and high throughput sequencing, was developed to sequence 8-oxoG in the human genome at 0.1 kb resolution.[96] The results revealed the accumulation and co-localization of 8-oxoG sites and γH2AX ChIP-seq signals at transcribed regions in MCF10A cells, particularly at long genes and at DNA replication origins. The data also revealed prevalent G4 structures in 8-oxoG enriched peaks. Besides antibody enrichment, a novel chemical enrichment method based on selective oxidation of 8-oxoG to form a biotin-labelled adduct was developed for 8-oxoG mapping (OG-seq).[97] Following high throughput sequencing, a map of 8-oxoG in the mouse genome was constructed at 0.1 kb resolution. These data showed that promoters, 5'-UTRs, 3'-UTRs and G4 structures harbour greater relative levels of 8-oxoG than expected by a random distribution throughout the genome. Another chemical strategy to enrich 8-oxoG is to use aldehyde reactive probe (ARP) to capture abasic sites produced by OGG1 at 8-oxoG sites at approximately 250-bp resolution on a genome-wide scale.[69] After sequencing, the results suggested less damage in functional elements such as promoters, exons, transcription factor binding sites and termination sites in a seemingly GC content-dependent manner in HepG2 cells. However, the resolution of these methods is relatively low to prevent further insight into sequence-specific 8-oxoG occurrence and distribution.

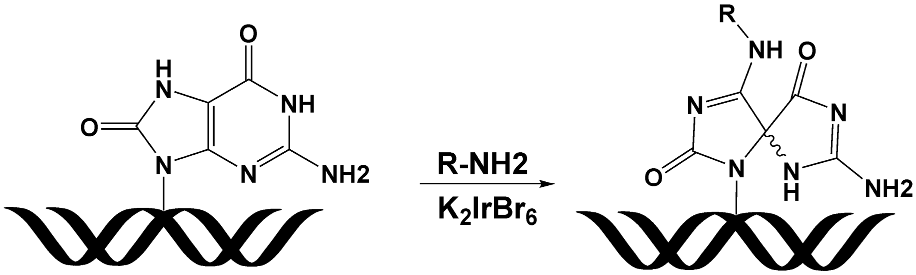


Figure 6. Selective oxidation of OG with a mild one-electron oxidant (K_2IrBr_6) to form a covalent adduct to a primary-amineterminated biotin

Alternatively, 8-oxoG could be sequenced at nucleotide resolution without enrichment using third generation sequencing technologies, e.g. single-molecule real-time sequencing[98] and nanopore technology.[99-100] However, these strategies were only tested on synthetic oligonucleotides as a proof-of-concept rather than biological samples where 8-oxoG occurrence compared to unmodified nucleobases is very low (0.001%).[95] Finally, a number of methods have the potential to detect 8-oxoG at nucleotide resolution, but were designed for one gene/one position, such as DNA hybridization probes containing a non-natural nucleoside specific for 8-oxoG,[101] ligation-mediated polymerase chain reaction (LM-PCR),[102] third base pair based amplification[103] and Hoogsteen base pairing-mediated PCR-sequencing.[104]

8. Oxidative sugar damage

8.1 Occurrence and relevance

Same as nucleobases, 2-deoxyribose is also attacked by a variety of endogenous oxidants and reactive chemicals, leading to a range of single lesions, DNA-DNA cross-links and protein-DNA cross-links. The oxidation of each of the five positions in 2-deoxyribose occurs in different rates which also parallels the solvent exposure of hydrogen atoms, i.e., H5' > H4' > H3' ≈ H2' ≈ H1'.[105] As the major product, the oxidation of 5'-position yields into two branches of products, one path yields a single strand break with a 3'-phosphate and a 5'-aldehyde containing nucleoside, the other results in a strand break with 3'-formylphosphate and 5'-(2-phosphoryl-1,4-dioxo-2-butane)-ended fragments.[106-107] 5'-aldehyde containing nucleoside, the best-characterized C5' nucleotide, undergoes β,δ-elimination reactions to release furfural, exhibiting a half-life of 4 days in a single strand oligonucleotide and varies from 5 - 12 days in dsDNA in phosphate buffer (10 mM, pH 7.2) at 37 °C.[108-109]

The distribution of 2-deoxyribose oxidation in a genome may not random as reflected by previous observations. First, hydroxyl radical DNA footprinting is widely used to study local changes in the solvent accessibility of DNA based on 2-deoxyribose oxidation induced SSBs.[110] DNA footprinting data revealed that the 2-deoxyribose oxidation was not sequence dependent but was affected by DNA-protein interaction, such as nucleosome.[111] Second, a sequence-selective oxidation of 2-deoxyribose was observed by minor groove binding antibiotics, such as bleomycin, neocarzinostatin, calicheamicin and lomaiviticin A.[112] For example, lomaiviticin A penetrate into AT base pair of the duplex d(GCTATAGC) preferentially, leading to SSB.[113]

8.2 Repair, mutagenicity, and toxicity

While studies of oxidative nucleobase damage have dominated this area as reviewed above, there is growing evidence that oxidative sugar damage in DNA poses a serious threat to genetic stability and cell survival. 2-Deoxyribose oxidation is usually accompanied by single strand breaks (SSBs) and by 5'- and/or 3'-termini lesions at the break site. SSB is repaired by a specific SSB repair (SSBR) pathway that include four basic steps: SSB recognition, end processing, gap filling and DNA ligation. First, PARP1 rapily binds to SSB and is then activated with chains of poly(ADP-ribose). Several SSBR proteins, such as XRCC1, Pol β, PNKP and LIG3, are recruited to SSB site by PARP1. In the second step of SSBR, the 3'- and 5'-termini terminus are converted into 3'-hydroxyl and 5'-phosphate moieties by end processing proteins. In the third step, single nucleotide (short-patch repair) or several nucleotides (long-patch repair) are inserted into SSBs gap by multiple DNA polymerases, including Pol β. The final step is DNA nick ligation by ligase.

Unrepaired SSBs can have an impact on cell fate in a number of ways. First, unrepaired SSBs can block DNA replication forks during the S phase of the cell cycle, leading to the formation of DSBs. DSBs can further cause deletions and translocations in the DNA. Moreover, SSBs may stall RNA polymerases during transcription, leading to cell death. Finally, overloading SSBs may induce cell death through excessive activation of PARP1. These evidences suggest that oxidative sugar damage and its accomplished SSB in DNA pose a serious threat to genetic stability and cell survival.

8.3 Detection strategies

Several molecular biology methods were developed to detect SSB, including comet assay, antibody based immunofluorescence assay and immunofluorescence microscopy.[114-115] Moreover, a single-strand break sequencing method was developed for SSB with 3'-hydroxyl group. These sites are labelled by nick translation with digoxigenin labelled dUTP, and enriched with anti-digoxigenin antibody.[116-117] However, SSBs with 3'-hydroxyl group could be artificially formed during genomic DNA extraction, yielding false positive results.

Besides general SSB, a GC/MS based method was reported to quantify 5'-aldehyde terminus specifically. The quantification of 5'-aldehyde termini was achieved by the reaction with *O*-benzylhydroxylamine to form a stable dioxime derivative, the elimination of which yielded the dioxime of trans-1,4-dioxo-2-butene that could be quantified by GC/MS. Based on this strategy, 5'-aldehyde terminus is detected at a frequency of 3.5 per 10^6 nt per µM Fe^{2+} and 57 per 10^6 nt per Gy (G-value 74 nmol/J) respectively.[118] There are limited studies focusing on SSB and 5'-aldehyde terminus. However, considering the vital role of SSB and 5'-aldehyde terminus in genome integrity and diseases. A sensitive and specific detection method is needed to provide valuable insights into SSBR process.

9. Overview of thesis work

The objectives of the work presented in this thesis are to map oxidation damage at genome-wide scale, understand how damage maps are governed by DNA repair and genomic architecture, and relate them with mutational signatures. To achieve genome-wide damage mapping, we came up with a novel strategy to label damage site with a oligonucleotide (code sequence) through bio-conjugation reaction. The key in this method is that the resulting artificial linkage from bio-conjugation reaction could be read through by DNA polymerases. The advantages of this method are: 1) biotin labelled code sequence could be used as a tag for affinity enrichment; 2) code sequence is also an adaptor for PCR amplification; 3) code sequence could be readout during sequencing and be used for marking damage locations. Based on this strategy, two novel methods were developed for DNA oxidation sequencing.

The first method is called click-code-seq that is specific for 8-oxoG. In this approach, the 8-oxoG site is recognised and removed by a DNA glycosylase. Then, a synthetic *O*-3'-propargyl modified nucleotide (prop-dGTP) is incorporated into the resulting gap. After that, the yield 3'-alkynyl DNA is ligated to a 5'-azido-modified code sequence via a copper(I)-catalyzed click reaction, resulting a triazole-linked DNA that could be amplified by DNA polymerases. After adaptor ligation, indexing and amplification, the library is ready for sequencing. In **Chapter 2**, we described the validation of click-code-seq with oligonucleotide and dsDNA models. Then, we applied this method to yeast genome as a proof of concept. In **Chapter 3**, we further applied this method to human genome. Nucleotide-resolution maps of 8-oxoG in both yeast genome and human genome are achieved in these two studies. Both studies revealed distinct patterns of oxidation sites, relating to chromatin architecture, histone modification, DNA-protein interactions and DNA damage response network. More importantly, the 3-bases damage pattern in human genome showed strong correlation with several mutation signatures that were related with increasing oxidative stress or repair protein deficiency. This exciting result provided the first direct observation of the mutagenesis process from DNA damage.

The second method presented in **Chapter 4** is developed for 5'-aldehyde terminus. In this method, the 5'-aldehyde terminus is directly labelled with an aminooxy-functionalized oligonucleotide, giving rise to an oxime linked DNA. This biocompatible altered DNA linkage could be amplified by DNA polymerase. The study with synthetic oligonucleotides showed that this method is able to detect 5'-aldehyde terminus at a frequency as low as 10^{-7} lesions/unmodified bases. This method could be further applied to map 5'-aldehyde terminus at nucleotide-resolution in genomic DNA.

Collectively, this work presented here contributes to understanding the distribution of DNA damage in a genome. The damage pattern is critical for relating early damage profiles with mutation signatures in human cancers. Furthermore, uncovering the mechanisms of mutagenesis process may contributes to a systems biology-based predictive models for early disease diagnosis.

References

1. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P.; Theriot, J.; Morales, M., Molecular biology of the cell, 5th Ed. *Garland Science* 2008.
2. Engel, S. R.; Dietrich, F. S.; Fisk, D. G.; Binkley, G.; Balakrishnan, R.; Costanzo, M. C.; Dwight, S. S.; Hitz, B. C., et al., The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3-Genes Genom. Genet.* **2014,** *4* (3), 389-398.
3. Collins, F. S.; Lander, E. S.; Rogers, J.; Waterston, R. H.; Conso, I. H. G. S., Finishing the euchromatic sequence of the human genome. *Nature* **2004,** *431* (7011), 931-945.
4. Wen, M.; Shen, T.; Wang, Y.; Li, Y. Z.; Shi, X. L.; Dang, X. Q., Next-Generation Sequencing in Early Diagnosis of Dent Disease 1: Two Case Reports. *Front. Med.* **2018,** *5*(7), 347
5. Wells, P. G.; Miller-Pinsler, L.; Bhatia, S.; Drake, D.; Shapiro, A. M., Reactive Oxygen Species (ROS) Formation, Oxidative DNA Damage and Repair in Teratogenesis. *Birth. Defects Res. A* **2015,** *103* (5), 359-359.
6. Tubbs, A.; Nussenzweig, A., Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **2017,** *168* (4), 644-656
7. O'Driscoll, M.; Jeggo, P. A., The role of double-strand break repair - insights from human genetics. *Nat. Rev. Genet.* **2006,** *7* (1), 45-54.
8. Hakem, R., DNA‐damage repair; the good, the bad, and the ugly. *The EMBO Journal* **2008,** *27* (4), 589-605.
9. Rao, K. S., Mechanisms of Disease: DNA repair defects and neurological disease. *Nat. Clin. Pract. Neurol.* **2007,** *3* (3), 162-172.
10. Cadet, J.; Wagner, J. R., DNA Base Damage by Reactive Oxygen Species, Oxidizing Agents, and UV Radiation. *CSH Perspect Biol.* **2013,** *5* (2), a012559
11. Dedon, P. C., The chemical toxicology of 2-deoxyribose oxidation in DNA. *Chem. Res. Toxicol.* **2008,** *21* (1), 206-219.
12. Rastogi, R. P.; Richa; Kumar, A.; Tyagi, M. B.; Sinha, R. P., Molecular Mechanisms of Ultraviolet Radiation-Induced DNA Damage and Repair. *J. Nucleic. Acids* **2010**, *16*, 592980
13. Drablos, F.; Feyzi, E.; Aas, P. A.; Vaagbo, C. B.; Kavli, B.; Bratlie, M. S.; Pena-Diaz, J.; Otterlei, M., et al., Alkylation damage in DNA and RNA - repair mechanisms and medical significance. *DNA Repair* **2004,** *3* (11), 1389-1407.
14. O'Driscoll, M., Diseases Associated with Defective Responses to DNA Damage. *CSH Perspect Biol.* **2012,** *4* (12), a012773

15. Jacobs, A. L.; Schar, P., DNA glycosylases: in DNA repair and beyond. *Chromosoma* **2012,** *121* (1), 1-20.

16. Krokan, H. E.; Bjoras, M., Base Excision Repair. *CSH Perspect Biol.* **2013,** *5* (4), a012583

17. Sale, J. E., Translesion DNA Synthesis and Mutagenesis in Eukaryotes. *CSH Perspect Biol.* **2013,** *5* (3), a012708

18. Jonason, A. S.; Kunala, S.; Price, G. J.; Restifo, R. J.; Spinelli, H. M.; Persing, J. A.; Leffell, D. J.; Tarone, R. E.; Brash, D. E., Frequent clones of p53-mutated keratinocytes in normal human skin. *Proc. Natl. Acad. Sci. U.S.A.* **1996,** *93* (24), 14025-14029.

19. Furrukh, M., Tobacco Smoking and Lung Cancer: Perception-changing facts. *Sultan. Qaboos. Univ. Med. J.* **2013,** *13* (3), 345-358.

20. Chen, C. H.; Dickman, K. G.; Moriya, M.; Zavadil, J.; Sidorenko, V. S.; Edwards, K. L.; Gnatenko, D. V.; Wu, L., et al., Aristolochic acid-associated urothelial cancer in Taiwan. *Proc. Natl. Acad. Sci. U.S.A.* **2012,** *109* (21), 8241-8246.

21. Yu, Y.; Wang, P. C.; Cui, Y. X.; Wang, Y. S., Chemical Analysis of DNA Damage. *Anal. Chem.* **2018,** *90* (1), 556-576.

22. Viel, A.; Bruselles, A.; Meccia, E.; Fornasarig, M.; Quaia, M.; Canzonieri, V.; Policicchio, E.; Urso, E. D., et al., A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **2017,** *20*, 39-49.

23. Ohno, M.; Sakumi, K.; Fukumura, R.; Furuichi, M.; Iwasaki, Y.; Hokama, M.; Ikemura, T.; Tsuzuki, T., et al., 8-oxoguanine causes spontaneous de novo germline mutations in mice. *Sci. Rep.* **2014,** *4*, 4689.

24. Kasai, H., Analysis of a form of oxidative DNA damage, 8-hydroxy-2′-deoxyguanosine, as a marker of cellular oxidative stress during carcinogenesis. *Mutation Research/Reviews in Mutation Research* **1997,** *387* (3), 147-163.

25. Pfeifer, G. P., Measuring the formation and repair of DNA damage by ligation-mediated PCR. *Methods Mol. Biol.* **2006,** *314*, 201-214.

26. Collins, A. R., The comet assay for DNA damage and repair - Principles, applications, and limitations. *Mol. Biotechnol.* **2004,** *26* (3), 249-261.

27. Fernández, J. L.; Gosálvez, J., Application of FISH to Detect DNA Damage. In *In Situ Detection of DNA Damage: Methods and Protocols*, Didenko, V. V., Ed. Humana Press: Totowa, NJ, 2002, 203-216.

28. Mitchell, L. A.; De Iuliis, G. N.; Aitken, R. J., The TUNEL assay consistently underestimates DNA damage in human spermatozoa and is influenced by DNA compaction and cell vitality: development of an improved methodology. *International Journal of Andrology* **2011,** *34* (1), 2-13.

29. Isabel, R. R. M.; Sandra, G. A.; Rafael, V. P.; Carmen, M. V.; Josefina, C. E.; del Carmen, C. E. M.; Rocio, G. M.; Francisco, A. H.; Elena, C. S. M., Evaluation of 8-hydroxy-2 '-deoxyguanosine (8-OHdG) adduct levels and DNA strand breaks in human peripheral blood lymphocytes exposed in vitro to polycyclic aromatic hydrocarbons with or without animal metabolic activation. *Toxicol. Mech. Method* **2012,** *22* (3), 170-183.

30. Villalta, P. W.; Balbo, S., The Future of DNA Adductomic Analysis. *Int. J. Mol. Sci.* **2017,** *18* (9), 1870.

31. Sykora, P.; Witt, K. L.; Revanna, P.; Smith-Roe, S. L.; Dismukes, J.; Lloyd, D. G.; Engelward, B. P.; Sobol, R. W., Next generation high throughput DNA damage detection platform for genotoxic compound screening. *Sci. Rep.* **2018,** *8*, 2771.

32. Matray, T. J.; Kool, E. T., A specific partner for abasic damage in DNA. *Nature* **1999,** *399* (6737), 704-708.

33. Zhang, X. M.; Donnelly, A.; Lee, I.; Berdis, A. J., Rational attempts to optimize non-natural nucleotides for selective incorporation opposite an abasic site. *Biochemistry* **2006,** *45* (44), 13293-13303.

34. Taniguchi, Y.; Kawaguchi, R.; Sasaki, S., Adenosine-1,3-diazaphenoxazine Derivative for Selective Base Pair Formation with 8-Oxo-2'-deoxyguanosine in DNA. *J. Am. Chem. Soc.* **2011,** *133* (19), 7272-7275.

35. Trantakis, I. A.; Nilforoushan, A.; Dahlmann, H. A.; Stäuble, C. K.; Sturla, S. J., In-Gene Quantification of O6-Methylguanine with Elongated Nucleoside Analogues on Gold Nanoprobes. *J. Am. Chem. Soc.* **2016,** *138* (27), 8497-8504.
36. Wyss, L. A.; Nilforoushan, A.; Williams, D. M.; Marx, A.; Sturla, S. J., The use of an artificial nucleotide for polymerase-based recognition of carcinogenic O6-alkylguanine DNA adducts. *Nucleic. Acids. Research* **2016,** *44* (14), 6564-6573.
37. Nilforoushan, A.; Furrer, A.; Wyss, L. A.; van Loon, B.; Sturla, S. J., Nucleotides with Altered Hydrogen Bonding Capacities Impede Human DNA Polymerase eta by Reducing Synthesis in the Presence of the Major Cisplatin DNA Adduct. *J. Am. Chem. Soc.* **2015,** *137* (14), 4728-4734.
38. Wyss, L. A.; Nilforoushan, A.; Eichenseher, F.; Suter, U.; Blatter, N.; Marx, A.; Sturla, S. J., Specific Incorporation of an Artificial Nucleotide Opposite a Mutagenic DNA Adduct by a DNA Polymerase. *J. Am. Chem. Soc.* **2015,** *137* (1), 30-33.
39. Aloisi, C. M. N.; Sturla, S. J.; Gahlon, H. L., A gene-targeted polymerase-mediated strategy to identify O6-methylguanine damage. *ChemComm* **2019,** *55* (27), 3895-3898.
40. Sanger, F.; Nicklen, S.; Coulson, A. R., DNA Sequencing with Chain-Terminating Inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **1977,** *74* (12), 5463-5467.
41. Smith, L. M.; Sanders, J. Z.; Kaiser, R. J.; Hughes, P.; Dodd, C.; Connell, C. R.; Heiner, C.; Kent, S. B. H.; Hood, L. E., Fluorescence Detection in Automated DNA-Sequence Analysis. *Nature* **1986,** *321* (6071), 674-679.
42. Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K., et al., Initial sequencing and analysis of the human genome. *Nature* **2001,** *409* (6822), 860-921.
43. International Human Genome Sequencing, C., Finishing the euchromatic sequence of the human genome. *Nature* **2004,** *431* (7011), 931-945.
44. Heather, J. M.; Chain, B., The sequence of sequencers: The history of sequencing DNA. *Genomics* **2016,** *107* (1), 1-8.
45. Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S., et al., Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005,** *437* (7057), 376-380.
46. Ju, J. Y.; Kim, D. H.; Bi, L. R.; Meng, Q. L.; Bai, X. P.; Li, Z. M.; Li, X. X.; Marma, M. S., et al., Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U.S.A.* **2006,** *103* (52), 19635-19640.
47. Levene, M. J.; Korlach, J.; Turner, S. W.; Foquet, M.; Craighead, H. G.; Webb, W. W., Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **2003,** *299* (5607), 682-686.
48. Deamer, D.; Akeson, M.; Branton, D., Three decades of nanopore sequencing. *Nat. Biotechnol.* **2016,** *34* (5), 518-524.
49. Pollard, M. O.; Gurdasani, D.; Mentzer, A. J.; Porter, T.; Sandhu, M. S., Long reads: their purpose and place. *Hum. Mol. Genet.* **2018,** *27* (R2), R234-R241.
50. Amente, S.; Di Palo, G.; Scala, G.; Castrignano, T.; Gorini, F.; Cocozza, S.; Moresano, A.; Pucci, P., et al., Genome-wide mapping of 8-oxo-7,8-dihydro-2'-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res.* **2018,** *47* (1), 221-236.
51. Akatsuka, S.; Aung, T. T.; Dutta, K. K.; Jiang, L.; Lee, W. H.; Liu, Y. T.; Onuki, J.; Shirase, T., et al., Contrasting genome-wide distribution of 8-hydroxyguanine and acrolein-modified adenine during oxidative stress-induced renal carcinogenesis. *Am. J. Pathol.* **2006,** *169* (4), 1328-1342.
52. Hu, J. C.; Lieb, J. D.; Sancar, A.; Adar, S., Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (41), 11507-11512.
53. Shu, X. T.; Xiong, X. S.; Song, J. H.; He, C.; Yi, C. Q., Base-Resolution Analysis of Cisplatin-DNA Adducts at the Genome Scale. *Angew Chem Int Edit* **2016,** *55* (46), 14244-14247.
54. Hu, J. C.; Adebali, O.; Adar, S.; Sancar, A., Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (26), 6758-6763.
55. Hu, J. C.; Adar, S.; Selby, C. P.; Lieb, J. D.; Sancar, A., Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Gene Dev.* **2015,** *29* (9), 948-960.

56. Adebali, O.; Chiou, Y. Y.; Hu, J. C.; Sancar, A.; Selby, C. P., Genome-wide transcription-coupled repair in Escherichia coli is mediated by the Mfd translocase. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (11), E2116-E2125.

57. Li, W. T.; Hu, J. C.; Adebali, O.; Adar, S.; Yang, Y. Y.; Chiou, Y. Y.; Sancar, A., Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (26), 6752-6757.

58. Wu, J. Z.; McKeague, M.; Sturla, S. J., Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J. Am. Chem. Soc.* **2018,** *140* (31), 9783-9787.

59. Clausen, A. R.; Lujan, S. A.; Burkholder, A. B.; Orebaugh, C. D.; Williams, J. S.; Clausen, M. F.; Malc, E. P.; Mieczkowski, P. A., et al., Tracking replication enzymology in vivo by genome-wide mapping of ribonucleotide incorporation. *Nat. Struct. Mol. Biol.* **2015,** *22* (3), 185-191.

60. Daigaku, Y.; Keszthelyi, A.; Muller, C. A.; Miyabe, I.; Brooks, T.; Retkute, R.; Hubank, M.; Nieduszynski, C. A.; Carr, A. M., A global profile of replicative polymerase usage. *Nat. Struct. Mol. Biol.* **2015,** *22* (3), 192-198.

61. Ding, J.; Taylor, M. S.; Jackson, A. P.; Reijns, M. A. M., Genome-wide mapping of embedded ribonucleotides and other noncanonical nucleotides using emRiboSeq and EndoSeq. *Nat. Protoc.* **2015,** *10* (9), 1433-1444.

62. Koh, K. D.; Balachander, S.; Hesselberth, J. R.; Storici, F., Ribose-seq: global mapping of ribonucleotides embedded in genomic DNA. *Nat. Methods* **2015,** *12* (3), 251.

63. Bryan, D. S.; Ransom, M.; Adane, B.; York, K.; Hesselberth, J. R., High resolution mapping of modified DNA nucleobases using excision repair enzymes. *Genome Res.* **2014,** *24* (9), 1534-1542.

64. Mao, P.; Smerdon, M. J.; Roberts, S. A.; Wyrick, J. J., Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (32), 9057-9062.

65. Mao, P.; Brown, A. J.; Malc, E. P.; Mieczkowski, P. A.; Smerdon, M. J.; Roberts, S. A.; Wyrick, J. J., Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Res.* **2017,** *27* (10), 1674-1684.

66. Li, Y. Y.; Tollefsbol, T. O., DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis. *Methods Mol. Biol.* **2011,** *791*, 11-21.

67. Xia, B.; Han, D. L.; Lu, X. Y.; Sun, Z. Z.; Zhou, A. K.; Yin, Q. Z.; Zeng, H.; Liu, M. H., et al., Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **2015,** *12* (11), 1047-1050.

68. Liu, Y. B.; Siejka-Zielinska, P.; Velikova, G.; Bi, Y.; Yuan, F.; Tomkova, M.; Bai, C. S.; Chen, L., et al., Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **2019,** *37* (4), 424.

69. Poetsch, A. R.; Boulton, S. J.; Luscombe, N. M., Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.* **2018,** *19, 215*.

70. Ding, Y.; Fleming, A. M.; Burrows, C. J., Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **2017,** *139* (7), 2569-2572.

71. Cadet, J.; Wagner, J. R., DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *CSH Perspect Biol.* **2013,** *5* (2), a012559.

72. Phaniendra, A.; Jestadi, D. B.; Periyasamy, L., Free Radicals: Properties, Sources, Targets, and Their Implication in Various Diseases. *Indian J. Clin. Bioche.* **2015,** *30* (1), 11-26.

73. Espinosa-Diez, C.; Miguel, V.; Mennerich, D.; Kietzmann, T.; Sanchez-Perez, P.; Cadenas, S.; Lamas, S., Antioxidant responses and cellular adjustments to oxidative stress. *Redox Biol.* **2015,** *6*, 183-197.

74. Allgayer, J.; Kitsera, N.; von der Lippen, C.; Epe, B.; Khobta, A., Modulation of base excision repair of 8-oxoguanine by the nucleotide sequence. *Nucleic Acids Res.* **2013,** *41* (18), 8559-8571.

75. Kim, G. H.; Kim, J. E.; Rhie, S. J.; Yoon, S., The Role of Oxidative Stress in Neurodegenerative Diseases. *Exp. Neurobiol.* **2015,** *24* (4), 325-340.

76. Prasad, S.; Gupta, S. C.; Pandey, M. K.; Tyagi, A. K.; Deb, L., Oxidative Stress and Cancer: Advances and Challenges. *Oxid. Med. Cell. Longev.* **2016,** 5010423.

77. Hussain, T.; Tan, B.; Yin, Y. L.; Blachier, F.; Tossou, M. C. B.; Rahu, N., Oxidative Stress and Inflammation: What Polyphenols Can Do for Us? *Oxid. Med. Cell. Longev.* **2016**, 7432797.

78. David, S. S.; O'Shea, V. L.; Kundu, S., Base-excision repair of oxidative DNA damage. *Nature* **2007,** *447* (7147), 941-950.

79. Bauer, N. C.; Corbett, A. H.; Doetsch, P. W., The current state of eukaryotic DNA base damage and repair. *Nucleic Acids Res.* **2015,** *43* (21), 10083-10101.

80. Nunoshiba, T.; Ishida, R.; Sasaki, M.; Iwai, S.; Nakabeppu, Y.; Yamamoto, K., A novel Nudix hydrolase for oxidized purine nucleoside triphosphates encoded by ORFYLR151c (PCD1 gene) in Saccharomyces cerevisiae. *Nucleic Acids Res.* **2004,** *32* (18), 5339-5348.

81. Shibutani, S.; Takeshita, M.; Grollman, A. P., Insertion of Specific Bases during DNA-Synthesis Past the Oxidation-Damaged Base 8-Oxodg. *Nature* **1991,** *349* (6308), 431-434.

82. Allgayer, J.; Kitsera, N.; Bartelt, S.; Epe, B.; Khobta, A., Widespread transcriptional gene inactivation initiated by a repair intermediate of 8-oxoguanine. *Nucleic Acids Res.* **2016,** *44* (15), 7267-7280.

83. Saxowsky, T. T.; Meadows, K. L.; Klungland, A.; Doetsch, P. W., 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **2008,** *105* (48), 18877-18882.

84. Basu, S.; Je, G.; Kim, Y. S., Transcriptional mutagenesis by 8-oxodG in alpha-synuclein aggregation and the pathogenesis of Parkinson's disease. *Exp. Mol. Med.* **2015,** *47*.

85. Pastukh, V.; Roberts, J. T.; Clark, D. W.; Bardwell, G. C.; Patel, M.; Al-Mehdi, A. B.; Borchert, G. M.; Gillespie, M. N., An oxidative DNA "damage" and repair mechanism localized in the VEGF promoter is important for hypoxia-induced VEGF mRNA expression. *Am. J. Physiol-Lung C* **2015,** *309* (11), L1367-L1375.

86. Kim, H. S.; Kim, B. H.; Jung, J. E.; Lee, C. S.; Lee, H. G.; Lee, J. W.; Lee, K. H.; You, H. J., et al., Potential role of 8-oxoguanine DNA glycosylase 1 as a STAT1 coactivator in endotoxin-induced inflammatory response. *Free Radical Bio. Med.* **2016,** *93*, 12-22.

87. Ba, X. Q.; Bacsi, A.; Luo, J. X.; Aguilera-Aguirre, L.; Zeng, X. L.; Radak, Z.; Brasier, A. R.; Boldogh, I., 8-Oxoguanine DNA Glycosylase-1 Augments Proinflammatory Gene Expression by Facilitating the Recruitment of Site-Specific Transcription Factors. *J. Immunol.* **2014,** *192* (5), 2384-2394.

88. Pan, L.; Zhu, B.; Hao, W. J.; Zeng, X. L.; Vlahopoulos, S. A.; Hazra, T. K.; Hegde, M. L.; Radak, Z., et al., Oxidized Guanine Base Lesions Function in 8-Oxoguanine DNA Glycosylase-1-mediated Epigenetic Regulation of Nuclear Factor B-driven Gene Expression. *J. Biol. Chem.* **2016,** *291* (49), 25553-25566.

89. Agrawal, P.; Hatzakis, E.; Guo, K. X.; Carver, M.; Yang, D. Z., Solution structure of the major G-quadruplex formed in the human VEGF promoter in K+: insights into loop interactions of the parallel G-quadruplexes. *Nucleic Acids Res.* **2013,** *41* (22), 10584-10592.

90. Schafer, G.; Cramer, T.; Suske, G.; Kemmner, W.; Wiedenmann, B.; Hocker, M., Oxidative stress regulates vascular endothelial growth factor-A gene transcription through Sp1-and Sp3-dependent activation of two proximal GC-rich promoter elements. *J. Biol. Chem.* **2003,** *278* (10), 8190-8198.

91. Clark, D. W.; Phang, T.; Edwards, M. G.; Geraci, M. W.; Gillespie, M. N., Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free Radical Bio. Med.* **2012,** *53* (1), 51-59.

92. Fleming, A. M.; Ding, Y.; Burrows, C. J., Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (10), 2604-2609.

93. Fleming, A. M.; Burrows, C. J., 8-Oxo-7,8-dihydroguanine, friend and foe: Epigenetic-like regulator versus initiator of mutagenesis. *DNA Repair* **2017,** *56*, 75-83.

94. Ohno, M.; Miura, T.; Furuichi, M.; Tominaga, Y.; Tsuchimoto, D.; Sakumi, K.; Nakabeppu, Y., A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. *Genome Res.* **2006,** *16* (5), 567-675.

95. Yoshihara, M.; Jiang, L.; Akatsuka, S.; Suyama, M.; Toyokuni, S., Genome-wide profiling of 8-oxoguanine reveals its association with spatial positioning in nucleus. *DNA Res.* **2014,** *21* (6), 603-612.

96. Gorini, F.; Di Palo, G.; Scala, G.; Lania, L.; Cocozza, S.; Amente, S.; Castrignanò, T.; Moresano, A., et al., Genome-wide mapping of 8-oxo-7,8-dihydro-2'-deoxyguanosine reveals accumulation of

oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res.* **2018,** *47* (1), 221-236.

97. Ding, Y.; Fleming, A. M.; Burrows, C. J., Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **2017,** *139* (7), 2569-2572.

98. Clark, A. B.; Lujan, S. A.; Kissling, G. E.; Kunkel, T. A., Mismatch repair-independent tandem repeat sequence instability resulting from ribonucleotide incorporation by DNA polymerase epsilon. *DNA Repair* **2011,** *10* (5), 476-482.

99. Schibel, A. E. P.; An, N.; Jin, Q. A.; Fleming, A. M.; Burrows, C. J.; White, H. S., Nanopore Detection of 8-Oxo-7,8-dihydro-2'-deoxyguanosine in Immobilized Single-Stranded DNA via Adduct Formation to the DNA Damage Site. *J. Am. Chem. Soc.* **2010,** *132* (51), 17992-17995.

100. An, N.; Fleming, A. M.; White, H. S.; Burrows, C. J., Nanopore detection of 8-oxoguanine in the human telomere repeat sequence. *ACS nano* **2015,** *9* (4), 4296-307.

101. Taniguchi, Y.; Kawaguchi, R.; Sasaki, S., Adenosine-1,3-diazaphenoxazine Derivative for Selective Base Pair Formation with 8-Oxo-2 '-deoxyguanosine in DNA. *J. Am. Chem. Soc.* **2011,** *133* (26), 10322-10322.

102. Nomoto, M.; Yamaguchi, R.; Kohno, K.; Kasai, H., Relations between clusters of oxidatively damaged nucleotides and active or open nucleosomes in the rat Nth 1 gene. *Oncogene* **2002,** *21* (11), 1649-1657.

103. Riedl, J.; Ding, Y.; Fleming, A. M.; Burrows, C. J., Identification of DNA lesions using a third base pair for amplification and nanopore sequencing. *Nat. commun.* **2015,** *6*, 8807.

104. Park, J.; Park, J. W.; Oh, H.; Maria, F. S.; Kang, J.; Tian, X. C., Gene-Specific Assessment of Guanine Oxidation as an Epigenetic Modulator for Cardiac Specification of Mouse Embryonic Stem Cells. *Plos One* **2016,** *11* (6).

105. Balasubramanian, B.; Pogozelski, W. K.; Tullius, T. D., DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95* (17), 9738-9743.

106. Dedon, P. C.; Goldberg, I. H., Free-radical mechanisms involved in the formation of sequence-dependent bistranded DNA lesions by the antitumor antibiotics bleomycin, neocarzinostatin, and calicheamicin. *Chem. Res. Toxicol.* **1992,** *5* (3), 311-32.

107. Chen, B.; Bohnert, T.; Zhou, X.; Dedon, P. C., 5′-(2-Phosphoryl-1,4-dioxobutane) as a Product of 5′-Oxidation of Deoxyribose in DNA: Elimination as trans-1,4-Dioxo-2-butene and Approaches to Analysis. *Chem. Res. Toxicol.* **2004,** *17* (11), 1406-1413.

108. Kodama, T.; Greenberg, M. M., Preparation and analysis of oligonucleotides containing lesions resulting from C5 '-oxidation. *J. Org. Chem.* **2005,** *70* (24), 9916-9924.

109. Rana, A.; Yang, K.; Greenberg, M. M., Reactivity of the Major Product of C5'-Oxidative DNA Damage in Nucleosome Core Particles. *Chembiochem* **2019,** *20* (5), 672-676.

110. Tullius, T. D.; Dombroski, B. A., Hydroxyl Radical Footprinting - High-Resolution Information About DNA Protein Contacts and Application to Lambda-Repressor and Cro Protein. *Proc. Natl. Acad. Sci. U.S.A.* **1986,** *83* (15), 5469-5473.

111. Shaytan, A. K.; Xiao, H.; Armeev, G. A.; Wu, C.; Landsman, D.; Panchenko, A. R., Hydroxyl-radical footprinting combined with molecular modeling identifies unique features of DNA conformation and nucleosome positioning. *Nucleic Acids Res.* **2017,** *45* (16), 9229-9243.

112. Dedon, P. C.; Goldberg, I. H., Free-Radical Mechanisms Involved in the Formation of Sequence-Dependent Bistranded DNA Lesions by the Antitumor Antibiotics Bleomycin, Neocarzinostatin, and Calicheamicin. *Chem. Res. Toxicol.* **1992,** *5* (3), 311-332.

113. Woo, C. M.; Li, Z. W.; Paulson, E. K.; Herzon, S. B., Structural basis for DNA cleavage by the potent antiproliferative agent (-)-lomaiviticin A. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (11), 2851-2856.

114. Calini, V.; Urani, C.; Camatini, M., Comet assay evaluation of DNA single- and double-strand breaks induction and repair in C3H10T1/2 cells. *Cell Biol. Toxicol.* **2002,** *18* (6), 369-79.

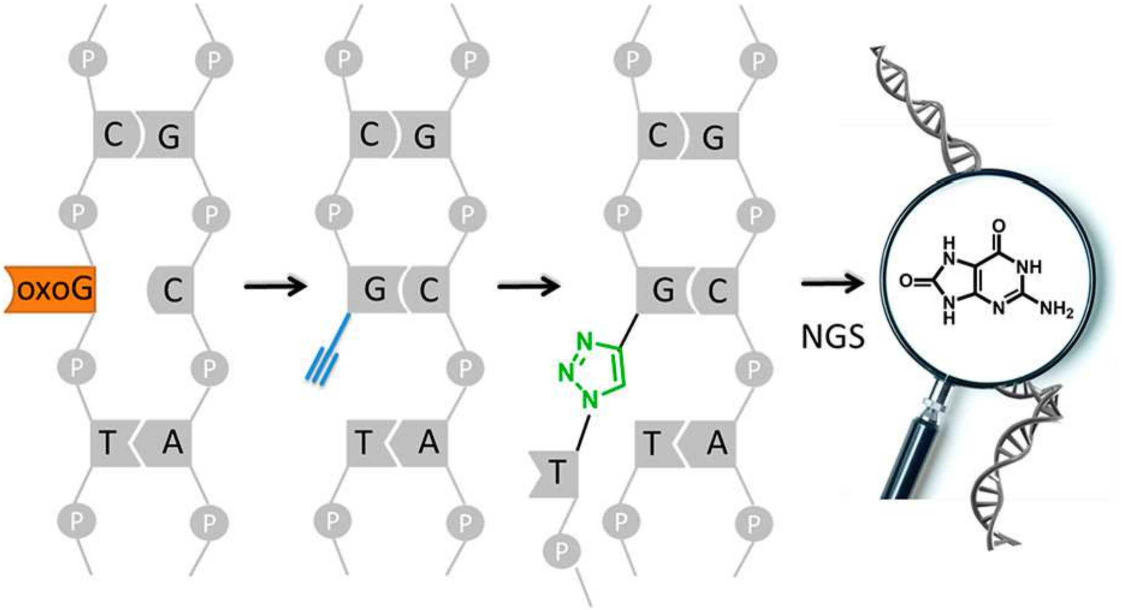
115. Michelena, J.; Lezaja, A.; Teloni, F.; Schmid, T.; Imhof, R.; Altmeyer, M., Analysis of PARP inhibitor toxicity by multidimensional fluorescence microscopy reveals mechanisms of sensitivity and resistance. *Nat. Commun.* **2018,** *9* (1), 2678.

116. Baranello, L.; Kouzine, F.; Wojtowicz, D.; Cui, K. R.; Przytycka, T. M.; Zhao, K. J.; Levens, D., DNA Break Mapping Reveals Topoisomerase II Activity Genome-Wide. *Int. J. Mol. Sci.* **2014,** *15* (7), 13111-13122.

117. Baranello, L.; Kouzine, F.; Wojtowicz, D.; Cui, K.; Zhao, K.; Przytycka, T. M.; Capranico, G.; Levens, D., Mapping DNA Breaks by Next-Generation Sequencing. *Methods Mol. Biol.* **2018,** *1672*, 155-166.

118. Chan, W.; Chen, B.; Wang, L.; Taghizadeh, K.; Demott, M. S.; Dedon, P. C., Quantification of the 2-Deoxyribonolactone and Nucleoside 5′-Aldehyde Products of 2-Deoxyribose Oxidation in DNA and Cells by Isotope-Dilution Gas Chromatography Mass Spectrometry: Differential Effects of γ-Radiation and Fe2+−EDTA. *J. Am. Chem. Soc.* **2010,** *132* (17), 6145-6153.

Chapter 2: Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq



Reprinted with permission from

Junzhou Wu, Maureen McKeague, and Shana J. Sturla. Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq, *J. Am. Chem. Soc.* 2018, 140 (31), 9783-9787

https://doi.org/10.1021/jacs.8b03715

Copyright © 2018 American Chemical Society

J.W. and S.J.S. designed the study. J.W. synthesized compounds and oligonucleotides, validated click-code-seq with model study, prepared sequencing library and analysed sequencing data. M.M. and S.J.S. conceived the research. J.W., M.M. and S.J.S. wrote the manuscript.

Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq

Junzhou Wu, Maureen McKeague, and Shana J. Sturla

Department of Health Science and Technology, Institute of Food, Nutrition, and Health, ETH Zurich, 8092 Zürich, Switzerland

Abstract

Single-nucleotide-resolution sequencing of DNA damage is required to decipher the complex causal link between the identity and location of DNA adducts and their biological impact. However, the low abundance and inability to specifically amplify DNA damage hinders single-nucleotide mapping of adducts within whole genomes. Despite the high biological relevance of guanine oxidation and seminal recent advances in sequencing bulky adducts, single-nucleotide-resolution whole genome mapping of oxidative damage is not yet realized. We coupled the specificity of repair enzymes with the efficiency of a click DNA ligation reaction to insert a biocompatible locator code, enabling high-throughput, nucleotide-resolution sequencing of oxidative DNA damage in a genome. We uncovered thousands of oxidation sites with distinct patterns related to transcription, chromatin architecture, and chemical oxidation potential. Click-code-seq overcomes barriers to DNA damage sequencing and provides a new approach for generating comprehensive, sequence-specific information about chemical modification patterns in whole genomes.

Introduction

Reliable transmission of genetic information underlying all processes of life depends on the structural integrity of DNA. However, DNA is continuously exposed to factors that alter nucleobase structures. As a consequence of metabolism and stimulated by chemical exposures, oxidative DNA damage from reactive oxygen species (ROS) constitutes a major threat to genetic integrity. One of the most frequent oxidative damage forms, 8-oxoguanine (8-oxoG), causes G to T transversion mutation frequently found in tumor suppressor genes and oncogenes[1-4]. Occurrence of 8-oxoG is also correlated with oxidative stress-associated processes, including atherosclerosis, diabetes, accelerated aging, and central nervous system pathologies[5-6]. Exome sequencing of tumors revealed a strong triplet sequence-dependent mutational signature from persistent 8-oxoG:A mismatch,[4] suggesting mutations arising from chemically-induced damage may be context dependent. Therefore, there is growing interest in the distribution of damage in genomes. Bulky adducts such as UV photoproducts[7-9], cisplatin adducts[10] and benzo[α]pyrene adducts[11] recently have been mapped by specific methods, but there are no reports of nucleotide-resolution sequencing of oxidative damage on a genome-wide scale.

Recent strategies have emerged toward addressing the distribution of 8-oxoG in DNA. Using 8-oxoG antibodies or selectively oxidizing 8-oxoG to form biotin-labelled sites, biological samples could be enriched for oxidized DNA fragments, providing 8-oxoG genomic maps with 0.1-1,000 kb-resolution[12-15]. The low resolution, however, prevents insight into sequence-specific distribution. Single-molecule real-time sequencing[16] and nanopore sequencing[17-18] were used for single-nucleotide-resolution sequencing in oligonucleotides or plasmids, but not biological samples where 8-oxoG occurrence is extremely low (~0.001%)[14]. Finally, 8-oxoG can be detected at nucleotide-resolution, but only in one gene/one position contexts[19-22]. Thus, there is no nucleotide-resolution data on the distribution of 8-oxoG within whole genomes.

To address this limitation, we developed a novel strategy that combines the removal of damage with repair proteins, incorporation of a novel alkynylated nucleoside with a DNA polymerase and labelling of damage sites with a code sequence via click chemistry. With this approach, damage sites are replaced by a synthetic oligonucleotide that fulfills three roles: i) tag for affinity enrichment; ii) adaptor for PCR amplification; and iii) code sequence for marking damage locations during sequencing. Herein, we sequenced a yeast genome, generating a nucleotide-resolution map of oxidative DNA damage. Our results provide a first insight on genomic distribution as a composite of nucleosome occupancy, histone modification, DNA-protein interactions and local sequence context.

Results and discussion

Labelling and amplifying 8-oxoG in DNA

The basis of click-code-seq involves the incorporation of an oligonucleotide code to mark each 8-oxoG position in genomic DNA (Figure 1a). First, DNA is treated with the base excision repair (BER) proteins, formamidopyrimidine DNA glycosylase (Fpg) and human apurinic/apyrimidinic endonuclease (APE1), to remove adducts and yield a 1 nt gap with a free 3'-hydroxyl[21]. Second, the resulting gap is filled with a synthetic *O*-3'-propargyl modified nucleotide (prop-dGTP), giving rise to a 3'-alkynyl modified DNA. Third, a 5'-azido-modified code sequence is ligated to the 3'-alkynyl DNA via a copper(I)-catalyzed click reaction (Figure 1b). The resulting biocompatible triazole-linked DNA could be read through by DNA polymerases[23]. Via this process, the sites of oxidative damage are stably labelled with a code sequence suitable for sequencing the location of the original damage site.

We first evaluated the biochemical steps involved in the click-code-seq workflow using synthetic oligonucleotides to confirm and optimize the incorporation of prop-dGTP, the click reaction and the bypass of the triazole linkage. Therminator IX DNA polymerase incorporated prop-dGTP opposite dC efficiently as a sequence terminator (Figure S1). The click reaction was carried out by a CuAAC ligation process, wherein, addition of DMSO and potassium phosphate buffer to the reaction greatly enhanced the ligation efficiency and lead to fewer by-products (Figure S2). For the triazole bypass, Vent exo- polymerase and a primer overhanging the triazole linkage were chosen based on a combination of primer extension efficiency (Figure S3) and fidelity[24]. Finally, potential biases regarding the bypass preference for certain sequences were examined with variations on the sequences of the triazole modified templates, and results suggested that the tested sequences were similarly processed (Figure S3, S4). Applying these optimized conditions to synthetic 30-mer DNA duplexes containing a single 8-oxoG

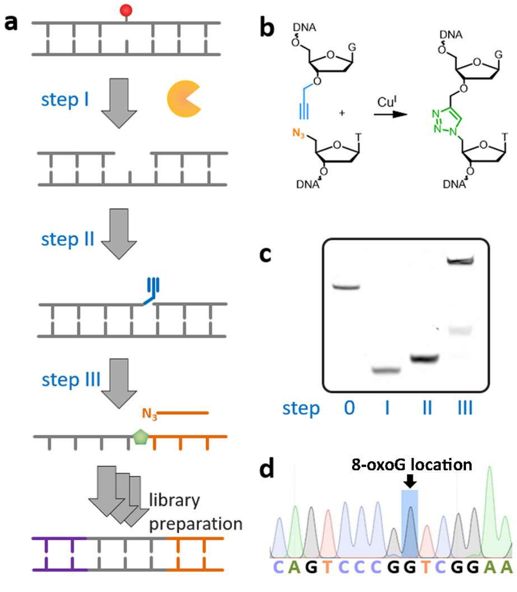


Figure 1: Click-code-seq. a) Steps for labelling 8-oxoG. b) Click reaction to produce ligated-DNA with polymerase-compatible triazole backbone. c) PAGE analysis to monitor labelling reactions in 8-oxoG-modified DNA duplex (fluorescently-labelled). Step 0 represents oligonucleotides before click-code-seq, and Steps I, II, II refer to those indicated in part a. d) Sanger sequencing showing code sequence (5'-TCGGAA-3') marking position of 8-oxoG. Red circle, 8-oxoG; Green hexagon, triazole.

modification indicated that removal of 8-oxo-G (step I) and incorporation of prop-dGTP (step II) were quantitative (Figure 1c), and the click reaction (step III) yield was 80%.

After carrying out studies with oligonucleotides, the same three-step click-code-seq procedure was performed on a specifically-modified 0.9 kb DNA fragment (Figure S5), the DNA was amplified and sequenced using Sanger sequencing. Insertion of the code sequence immediately after the known 8-oxoG location was confirmed (Figure 1d). Furthermore, a titration experiment indicated the capacity to detect 8-oxoG as low as 10^{-6} 8-oxoG/unmodified bases.

Click-code-seq was used to map DNA damage in the yeast genome (strain BY4741). To ensure that free 3' hydroxyl groups analyzed resulted only from excised sites, fragmented DNA was treated with APE1 to remove abasic sites, then in the same pot, the generated free 3' hydroxyl groups, as well as any strand breaks or DNA fragment termini, were all blocked by incorporating dideoxynucleotide phosphates. Biotinylated strands resulting from click-code-seq were harvested using streptavidin beads, 5'-phosphorylated with T4 polynucleotide kinase, and subjected to adapter ligation (Figure S7). After indexing and amplification, the libraries were sequenced by Illumina Miseq (2 million reads depth) and aligned with the *S. cerevisiae* genome R64-2-1[25]. This method allowed us to define the frequency of damage at specific locations with single-nucleotide-resolution (Figure S8). While 8-oxoG is the primary physiological

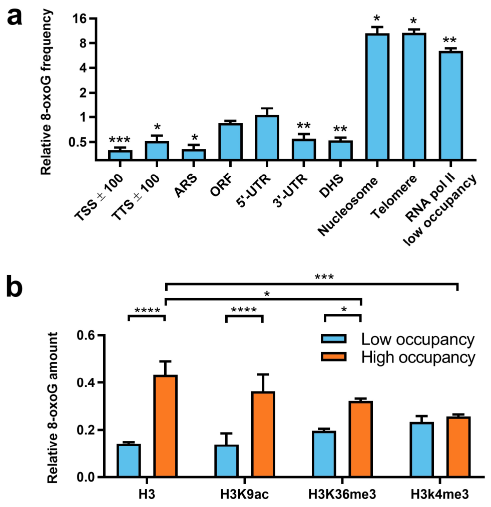


Figure 2: 8-oxoG distribution within the yeast genome. a) Relative 8-oxoG frequency in genomic elements. b) Relative 8-oxoG frequency with low vs high occupancy of histone H3 and histone modifications (n= 41,203). All data normalized to overall frequency of 1 and the size (bp) of the feature.

substrate for Fpg glycosylase in genomic DNA[26], it also removes other DNA adducts. Amongst these, only guanine adducts will be identified due to subsequent reference sequence alignment. Nonetheless, the related guanine oxidation adduct fapy-guanine, and the alkylation adduct methyl-fapy-guanine, are also mapped. Therefore, for further applications of this method, defining relative contributions of these adducts should be addressed[27-35].

8-oxoG distribution in genomic features

We analyzed the distribution and frequency of 8-oxoG within discrete genomic features (Figure 2a). In several features, especially transcription start sites (TSS), transcription terminator sites (TTS), DNase I hypersensitive (DHS) sites and autonomously replicating sequences (ARS), there was less 8-oxoG than the average coverage over the entire genome, however, telomeres, nucleosomes, and positions of low RNA pol II occupancy had higher 8-oxoG frequency. We then further evaluated 8-oxoG within flanking sequences of genomic features (Figures 3a, 3b, S9). We observed clear 8-oxoG minima within the TSS/TTS and maxima at low RNA Pol II occupancy sites. Considering that TSS, TTS and ARS are void of nucleosomes[36], DHS sites have highly remodelled nucleosomes, and low RNA Pol II occupancy sites contain intact

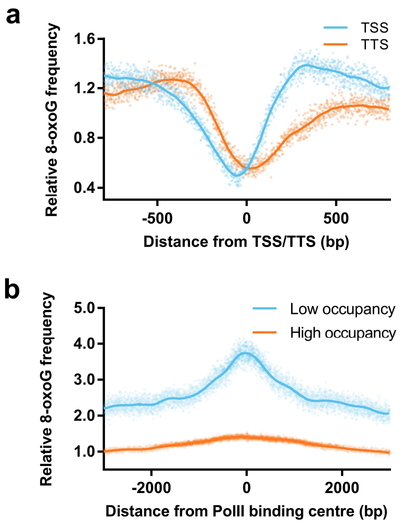


Figure 3: Metaprofiles of 8-oxoG centered on a) transcription start sites (TSS, n=3,736) and transcription termination sites (TTS, n=3,736), and flanking sequences (800 bp), b) Pol II binding sites (low occupancy: n=1,597, high occupancy: n=17,490) and flanking sequences (3 kbp). All data normalized to overall frequency of 1.

nucleosomes[37], the 8-oxoG map reveals a strong nucleosome influence on 8-oxoG distribution.

An 8-oxoG metaprofile centered on the nucleosome dyad suggested that 8-oxoG frequency and nucleosome occupancy was essentially identical in pattern (Figure S9). The nucleosome-free linker region had 16% less 8-oxoG compared to the nucleosome dyad, consistent with biochemical evidence that histone-DNA interactions in the nucleosome core particle impair accessibility to BER enzymes. Thus, damage in the center of nucleosomes is more difficult to remove than near the edge[38-40], consistent with nucleosome-related patterns of the BER substrate 7meG[41].

We then aimed to determine the influence of histone modifications on the distribution of 8-oxoG, which, compared to nucleosomal DNA, is relatively unknown. We compared 8-oxoG in nucleosomes that have either a high (>75%) or low (<25%) occupancy score of histone H3 and three post-translational modifications associated with nucleosome unwrapping: acetylated H3K9, trimethylated H3K36, and trimethylated H3K4 (Figure 2b)[42]. In general, there is a 2.9-fold increase in 8-oxoG abundance for nucleosomes with high occupancy of histone H3 compared to those with low occupancy ($P < 0.0001$), supporting our hypothesis that DNA-nucleosome interactions inhibit BER. When examining the histone modifications, however, the fold change decreases to 2.4 (H3K9ac, $P < 0.0001$), 1.6 (H3K36me3, $P < 0.05$) and 1.1 (H3K4me3, not significant), i.e. the modifications result in a reduction of 8-oxoG damage associated with H3. Furthermore, nucleosomes with higher levels of modifications harboured less 8-oxoG than unmodified nucleosomes (H3K4me3 vs H3: $P<0.001$, H3K36me3 vs H3: $P< 0.05$). These data are consistent with histone modifications being associated with a more open chromatin state, rendering those regions of DNA more accessible to BER[43-44].

The frequency of 8-oxoG over the region 800 bp before and after the TSS had clearly reduced 8-oxoG (Figure 3a). Likewise, similar patterns of reduced occurrence of UV damage[7-9], cisplatin adducts[10], and benzo[α]pyrene adducts[11] in the human genome and 7meG[41] in the yeast genome around the TSS region were recently characterized. We also observed a similar trend around the TF binding sites Abf1 and Reb1 (Figure S9). These findings suggest that in contrast to accumulation of 8-oxoG damage in nucleosomes, DNA–TF interactions may either protect against 8-oxoG occurrence or recruit repair[45].

Sequences flanking 8-oxoG in the genome

Due to a lower ionization potential attributed to π-stacking, oxidation of guanine in dsDNA shows a modest preference for the 5'G in a 5'-GG-3' context[46-47]. We analyzed the sequence context of nucleobases flanking 8-oxoG sites (Figure 4a). The frequency of each base 3' to 8-oxoG was highly variable: 51% G, 20% A, 16% T and 13% C. In contrast, a uniform frequency of all nucleobases was observed 5' to 8-oxoG (<4% variance). The same analysis within different genomic features indicated no significant difference from overall results (Figure S10). The relative frequencies of 8-oxoG in 5'-XGY-3' triplets exhibited a negative linear correlation as a function of the effective energy of a positive charge localized at the middle guanine in 5'-XGY-3'[47-50] ($R^2 = 0.65$, Figure 4b). These results support, on a genome level, the working model that the first G in a 5'-GG-3' dimer is more easily oxidized, but further research is needed to understand any potential biases.

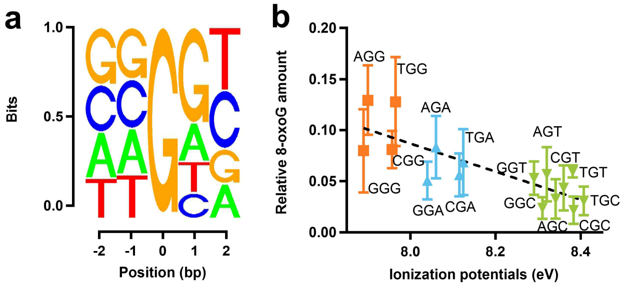


Figure 4: Local sequence context and 8-oxoG distribution. a) Sequence logo plot shows prevalence of bases surrounding 8-oxoG. b) Relative frequency of a triplet bearing 8-oxoG in the middle as a function of ionization potential.

Conclusions

Herein, we developed a method “click-code-seq” involving a novel three-step combination of excision, marking and click-labelling. A key advantage is the ability to map damage in the genome at single-nucleotide-resolution, which is significant for understanding how distribution relates with genomic features, DNA-protein interactions and sequence context.Our analysis suggests that oxidative damage persists within heterochromatin but is more efficiently repaired within euchromatin. These observations are consistent with data from microarray-based profiles of 8-oxoG suggesting that lamina-associated domains, associated with heterochromatin[51], have more 8-oxoG[14]. Meanwhile, 8-oxoG was reduced in regions where histones are acetylated or methylated. These data are consistent with a hypothesis that small ROS can easily penetrate dense, compact chromatin structures, whereas larger repair proteins cannot. The data derived herein suggest a working model involving contributions of both damage and repair to the genomic occurrence of oxidative damage, wherein formation appears to be influenced by local chemistry whereas repair is governed by genomic features and protein interactions.

In summary, click-code-seq is a robust and broadly applicable strategy for generating high-resolution maps of damage distribution in whole genomes. The approach could be adapted to analyze genomes of other organisms and other DNA damage/modifications. Further research will focus on these goals as well as addressing potential biases inherent in damage sequencing. The capacity for efficient nucleotide-resolution damage mapping is anticipated to be a new basis for addressing the etiology of DNA damage-related diseases and therapies.

References

1. Cadet, J.; Wagner, J. R., DNA base damage by reactive oxygen species, oxidizing agents, and UV radiation. *CSH Perspect Biol.* **2013,** *5* (2), a012559.
2. Nakabeppu, Y., Cellular levels of 8-oxoguanine in either DNA or the nucleotide pool play pivotal roles in carcinogenesis and survival of cancer cells. *Int. J. Mol. Sci.* **2014,** *15* (7), 12543-57.
3. Isoda, T.; Nakatsu, Y.; Yamauchi, K.; Piao, J.; Yao, T.; Honda, H.; Nakabeppu, Y.; Tsuzuki, T., Abnormality in Wnt signaling is causatively associated with oxidative stress-induced intestinal tumorigenesis in MUTYH-null mice. *Int. J. Biol. Sci.* **2014,** *10* (8), 940-947.
4. Viel, A.; Bruselles, A.; Meccia, E.; Fornasarig, M.; Quaia, M.; Canzonieri, V.; Policicchio, E.; Urso, E. D.; Agostini, M.; Genuardi, M.; Lucci-Cordisco, E.; Venesio, T.; Martayan, A.; Diodoro, M. G.; Sanchez-Mete, L.; Stigliano, V.; Mazzei, F.; Grasso, F.; Giuliani, A.; Baiocchi, M.; Maestro, R.; Giannini, G.; Tartaglia, M.; Alexandrov, L. B.; Bignami, M., A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **2017,** *20*, 39-49.
5. Bosshard, M.; Markkanen, E.; van Loon, B., Base excision repair in physiology and pathology of the central nervous system. *Int. J. Mol. Sci.* **2012,** *13* (12), 16172-16222.
6. Wells, P. G.; McCallum, G. P.; Chen, C. S.; Henderson, J. T.; Lee, C. J.; Perstin, J.; Preston, T. J.; Wiley, M. J.; Wong, A. W., Oxidative stress in developmental origins of disease: teratogenesis, neurodevelopmental deficits, and cancer. *Toxicological Sciences* **2009,** *108* (1), 4-18.
7. Hu, J.; Adar, S.; Selby, C. P.; Lieb, J. D.; Sancar, A., Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **2015,** *29* (9), 948-960.
8. Adar, S.; Hu, J.; Lieb, J. D.; Sancar, A., Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (15), E2124-E2133.
9. Hu, J.; Adebali, O.; Adar, S.; Sancar, A., Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (26), 6758-6763.
10. Hu, J.; Lieb, J. D.; Sancar, A.; Adar, S., Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (41), 11507-11512.
11. Li, W.; Hu, J.; Adebali, O.; Adar, S.; Yang, Y.; Chiou, Y. Y.; Sancar, A., Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (26), 6752-6757.
12. Ohno, M.; Miura, T.; Furuichi, M.; Tominaga, Y.; Tsuchimoto, D.; Sakumi, K.; Nakabeppu, Y., A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. *Genome Res.* **2006,** *16* (5), 567-575.

13. Akatsuka, S.; Aung, T. T.; Dutta, K. K.; Jiang, L.; Lee, W. H.; Liu, Y. T.; Onuki, J.; Shirase, T.; Yamasaki, K.; Ochi, H.; Naito, Y.; Yoshikawa, T.; Kasai, H.; Tominaga, Y.; Sakumi, K.; Nakabeppu, Y.; Kawai, Y.; Uchida, K.; Yamasaki, A.; Tsuruyama, T.; Yamada, Y.; Toyokuni, S., Contrasting genome-wide distribution of 8-hydroxyguanine and acrolein-modified adenine during oxidative stress-induced renal carcinogenesis. *Am. J. Pathol.* **2006,** *169* (4), 1328-1342.
14. Yoshihara, M.; Jiang, L.; Akatsuka, S.; Suyama, M.; Toyokuni, S., Genome-wide profiling of 8-oxoguanine reveals its association with spatial positioning in nucleus. *DNA Res.* **2014,** *21* (6), 603-612.
15. Ding, Y.; Fleming, A. M.; Burrows, C. J., Sequencing the Mouse Genome for the Oxidatively Modified Base 8-Oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **2017,** *139* (7), 2569-2572.
16. Clark, T. A.; Spittle, K. E.; Turner, S. W.; Korlach, J., Direct detection and sequencing of damaged DNA bases. *Genome integrity* **2011,** *2*, 10.
17. Schibel, A. E.; An, N.; Jin, Q.; Fleming, A. M.; Burrows, C. J.; White, H. S., Nanopore detection of 8-oxo-7,8-dihydro-2'-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site. *J. Am. Chem. Soc.* **2010,** *132* (51), 17992-17995.
18. An, N.; Fleming, A. M.; White, H. S.; Burrows, C. J., Nanopore detection of 8-oxoguanine in the human telomere repeat sequence. *ACS nano* **2015,** *9* (4), 4296-4307.
19. Taniguchi, Y.; Kawaguchi, R.; Sasaki, S., Adenosine-1,3-diazaphenoxazine derivative for selective base pair formation with 8-oxo-2'-deoxyguanosine in DNA. *J. Am. Chem. Soc.* **2011,** *133* (19), 7272-7275.
20. Nomoto, M.; Yamaguchi, R.; Kohno, K.; Kasai, H., Relations between clusters of oxidatively damaged nucleotides and active or open nucleosomes in the rat Nth 1 gene. *Oncogene* **2002,** *21* (11), 1649-1657.
21. Riedl, J.; Ding, Y.; Fleming, A. M.; Burrows, C. J., Identification of DNA lesions using a third base pair for amplification and nanopore sequencing. *Nat. commun.* **2015,** *6*, 8807.
22. Park, J.; Park, J. W.; Oh, H.; Maria, F. S.; Kang, J.; Tian, X., Gene-Specific Assessment of Guanine Oxidation as an Epigenetic Modulator for Cardiac Specification of Mouse Embryonic Stem Cells. *PLoS One* **2016,** *11* (6), e0155792.
23. El-Sagheer, A. H.; Sanzone, A. P.; Gao, R.; Tavassoli, A.; Brown, T., Biocompatible artificial DNA linker that is read through by DNA polymerases and is functional in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.* **2011,** *108* (28), 11338-11343.
24. Litosh, V. A.; Wu, W. D.; Stupi, B. P.; Wang, J. C.; Morris, S. E.; Hersh, M. N.; Metzker, M. L., Improved nucleotide selectivity and termination of 3'-OH unblocked reversible terminators by molecular tuning of 2-nitrobenzyl alkylated HOMedU triphosphates. *Nucleic Acids Res.* **2011,** *39* (6), e39.
25. Database, T. S. G. S288C reference genome sequence and annotation.
26. Tchou, J.; Kasai, H.; Shibutani, S.; Chung, M. H.; Laval, J.; Grollman, A. P.; Nishimura, S., 8-Oxoguanine (8-Hydroxyguanine) DNA Glycosylase and Its Substrate-Specificity. *Proc. Natl. Acad. Sci. U.S.A.* **1991,** *88* (11), 4690-4694.
27. Boysen, G.; Pachkowski, B. F.; Nakamura, J.; Swenberg, J. A., The formation and biological significance of N7-guanine adducts. *Mutat. Res.* **2009,** *678* (2), 76-94.
28. Dizdaroglu, M.; Kirkali, G.; Jaruga, P., Formamidopyrimidines in DNA: Mechanisms of formation, repair, and biological effects. *Free Radical Bio. Med.* **2008,** *45* (12), 1610-1621.
29. Douki, T.; Martini, R.; Ravanat, J. L.; Turesky, R. J.; Cadet, J., Measurement of 2,6-diamino-4-hydroxy-5-formamidopyrimidine and 8-oxo-7,8-dihydroguanine in isolated DNA exposed to gamma radiation in aqueous solution. *Carcinogenesis* **1997,** *18* (12), 2385-2391.
30. Earley, L. F.; Minko, I. G.; Christov, P. P.; Rizzo, C. J.; Lloyd, R. S., Mutagenic Spectra Arising from Replication Bypass of the 2,6-Diamino-4-hydroxy-N-5-methyl Formamidopyrimidine Adduct in Primate Cells. *Chem. Res. Toxicol.* **2013,** *26* (7), 1108-1114.
31. Greenberg, M. M., The Formamidopyrimidines: Purine Lesions Formed in Competition With 8-Oxopurines From Oxidative Stress. *Accounts Chem. Res.* **2012,** *45* (4), 588-597.
32. Jaruga, P.; Kirkali, G.; Dizdaroglu, M., Measurement of formamidopyrimidines in DNA. *Free Radical Bio. Med.* **2008,** *45* (12), 1601-1609.

33. Pande, P.; Haraguchi, K.; Jiang, Y. L.; Greenberg, M. M.; Basu, A. K., Unlike Catalyzing Error-Free Bypass of 8-OxodGuo, DNA Polymerase lambda Is Responsible for a Significant Part of Fapy.dG-Induced G -> T Mutations in Human Cells. *Biochemistry* **2015,** *54* (10), 1859-1862.
34. Patra, A.; Banerjee, S.; Johnson Salyard, T. L.; Malik, C. K.; Christov, P. P.; Rizzo, C. J.; Stone, M. P.; Egli, M., Structural Basis for Error-Free Bypass of the 5-N-Methylformamidopyrimidine-dG Lesion by Human DNA Polymerase eta and Sulfolobus solfataricus P2 Polymerase IV. *J. Am. Chem. Soc.* **2015,** *137* (22), 7011-7014.
35. Pujari, S. S.; Tretyakova, N., Chemical Biology of N(5)-Substituted Formamidopyrimidine DNA Adducts. *Chem. Res. Toxicol.* **2017,** *30* (1), 434-452.
36. Zhang, P.; Du, G.; Zou, H.; Xie, G.; Chen, J.; Shi, Z.; Zhou, J., Genome-wide mapping of nucleosome positions in Saccharomyces cerevisiae in response to different nitrogen conditions. *Sci. Rep.* **2016,** *6*, 33970.
37. Kulaeva, O. I.; Hsieh, F. K.; Chang, H. W.; Luse, D. S.; Studitsky, V. M., Mechanism of transcription through a nucleosome by RNA polymerase II. *Biochim. Biophys. Acta* **2013,** *1829* (1), 76-83.
38. Rodriguez, Y.; Hinz, J. M.; Smerdon, M. J., Accessing DNA damage in chromatin: Preparing the chromatin landscape for base excision repair. *DNA Repair* **2015,** *32*, 113-9.
39. Menoni, H.; Gasparutto, D.; Hamiche, A.; Cadet, J.; Dimitrov, S.; Bouvet, P.; Angelov, D., ATP-dependent chromatin remodeling is required for base excision repair in conventional but not in variant H2A.Bbd nucleosomes. *Mol. Cell Biol.* **2007,** *27* (17), 5949-56.
40. Bilotti, K.; Tarantino, M. E.; Delaney, S., Human Oxoguanine Glycosylase 1 Removes Solution Accessible 8-Oxo-7,8-dihydroguanine Lesions from Globally Substituted Nucleosomes Except in the Dyad Region. *Biochemistry* **2018,** *57* (9), 1436-1439.
41. Mao, P.; Brown, A. J.; Malc, E. P.; Mieczkowski, P. A.; Smerdon, M. J.; Roberts, S. A.; Wyrick, J. J., Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Res.* **2017,** *27* (10), 1674-1684.
42. Pokholok, D. K.; Harbison, C. T.; Levine, S.; Cole, M.; Hannett, N. M.; Lee, T. I.; Bell, G. W.; Walker, K.; Rolfe, P. A.; Herbolsheimer, E.; Zeitlinger, J.; Lewitter, F.; Gifford, D. K.; Young, R. A., Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **2005,** *122* (4), 517-527.
43. Zhang, T.; Cooper, S.; Brockdorff, N., The interplay of histone modifications - writers that read. *EMBO Rep.* **2015,** *16* (11), 1467-1481.
44. Weiner, A.; Hsieh, T. H.; Appleboim, A.; Chen, H. V.; Rahat, A.; Amit, I.; Rando, O. J.; Friedman, N., High-resolution chromatin dynamics during a yeast stress response. *Mol. Cell* **2015,** *58* (2), 371-386.
45. Ba, X. Q.; Bacsi, A.; Luo, J. X.; Aguilera-Aguirre, L.; Zeng, X. L.; Radak, Z.; Brasier, A. R.; Boldogh, I., 8-Oxoguanine DNA Glycosylase-1 Augments Proinflammatory Gene Expression by Facilitating the Recruitment of Site-Specific Transcription Factors. *J. Immunol.* **2014,** *192* (5), 2384-2394.
46. Margolin, Y.; Shafirovich, V.; Geacintov, N. E.; DeMott, M. S.; Dedon, P. C., DNA sequence context as a determinant of the quantity and chemistry of guanine oxidation produced by hydroxyl radicals and one-electron oxidants. *J. Bio. Chem.* **2008,** *283* (51), 35569-35578.
47. Senthilkumar, K.; Grozema, F. C.; Guerra, C. F.; Bickelhaupt, F. M.; Siebbeles, L. D., Mapping the sites for selective oxidation of guanines in DNA. *J. Am. Chem. Soc.* **2003,** *125* (45), 13658-9.
48. Kawanishi, S.; Oikawa, S.; Murata, M.; Tsukitome, H.; Saito, I., Site-specific oxidation at GG and GGG sequences in double-stranded DNA by benzoyl peroxide as a tumor promoter. *Biochemistry* **1999,** *38* (51), 16733-16739.
49. Saito, I.; Nakamura, T.; Nakatani, K.; Yoshioka, Y.; Yamaguchi, K.; Sugiyama, H., Mapping of the hot spots for DNA damage by one-electron oxidation: Efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. *J. Am. Chem. Soc.* **1998,** *120* (48), 12686-12687.
50. Yoshioka, Y.; Kitagawa, Y.; Takano, Y.; Yamaguchi, K.; Nakamura, T.; Saito, I., Experimental and theoretical studies on the selectivity of GGG triplets toward one-electron oxidation in B-form DNA. *J. Am. Chem. Soc.* **1999,** *121* (38), 8712-8719.

51. van Steensel, B.; Belmont, A. S., Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **2017,** *169* (5), 780-791.

52. Miller, G. P.; Kool, E. T., Versatile 5'-functionalization of oligonucleotides on solid support: amines, azides, thiols, and thioethers via phosphorus chemistry. *J. Org. Chem.* **2004,** *69* (7), 2404-10.

53. Jiang, C.; Pugh, B. F., A compiled and systematic reference map of nucleosome positions across the Saccharomyces cerevisiae genome. *Genome Biol.* **2009,** *10* (10), R109.

54. Miura, F.; Kawaguchi, N.; Sese, J.; Toyoda, A.; Hattori, M.; Morishita, S.; Ito, T., A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (47), 17846-51.

55. Hesselberth, J. R.; Chen, X.; Zhang, Z.; Sabo, P. J.; Sandstrom, R.; Reynolds, A. P.; Thurman, R. E.; Neph, S.; Kuehn, M. S.; Noble, W. S.; Fields, S.; Stamatoyannopoulos, J. A., Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **2009,** *6* (4), 283-9.

56. Liu, C. L.; Kaplan, T.; Kim, M.; Buratowski, S.; Schreiber, S. L.; Friedman, N.; Rando, O. J., Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLOS Biol.* **2005,** *3* (10), e328.

57. Kasinathan, S.; Orsi, G. A.; Zentner, G. E.; Ahmad, K.; Henikoff, S., High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **2014,** *11* (2), 203-9.

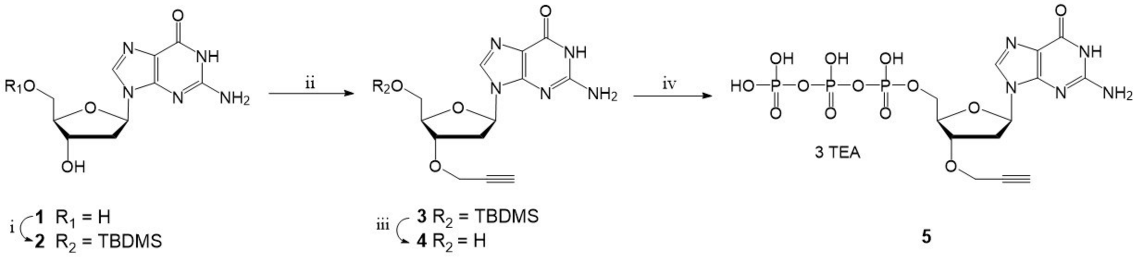
58. Pan, D.; Zhou, Q.; Rong, S. Z.; Zhang, G. T.; Zhang, Y. N.; Liu, F. H.; Li, M. J.; Chang, D.; Pan, H. Z., Electrochemical immunoassay for the biomarker 8-hydroxy-2'-deoxyguanosine using a glassy carbon electrode modified with chitosan and poly(indole-5-carboxylic acid). *Microchim. Acta* **2016,** *183* (1), 361-368.

59. Fan, J.; Liu, Y.; Xu, E.; Zhang, Y.; Wei, W.; Yin, L.; Pu, Y.; Liu, S., A label-free ultrasensitive assay of 8-hydroxy-2'-deoxyguanosine in human serum and urine samples via polyaniline deposition and tetrahedral DNA nanostructure. *Anal. Chim. Acta* **2016,** *946*, 48-55.

60. Ma, B.; Jing, M.; Villalta, P. W.; Kapphahn, R. J.; Montezuma, S. R.; Ferrington, D. A.; Stepanov, I., Simultaneous determination of 8-oxo-2'-deoxyguanosine and 8-oxo-2'-deoxyadenosine in human retinal DNA by liquid chromatography nanoelectrospray-tandem mass spectrometry. *Sci. Rep.* **2016,** *6*, 22375.

Supporting Information

Synthesis of prop-dGTP



i) TBDMSCl, Imidazole, DMF; ii) Propargyl bromide, NaH, THF; iii) TBAF, THF;
iv) $POCl_3$, Proton sponge, TMP; Tributylammonium pyrophosphate, Tributylamine, DMF; TEAB buffer

General synthesis information

All chemical reagents were purchased from Sigma Aldrich and used without further purification. All reactions were monitored by TLC using commercial Merck Plates coated with silica gel GF_{254} (0.24 mm thick). Flash column chromatography was performed on a Biotage SP4 system with pre-packed cartridges. ^{1}H, ^{13}C, and ^{31}P NMR spectra were recorded on a Bruker Biospin 400 MHz NMR instrument at 25 ^{o}C. Chemical shifts (δ, ppm) are reported relative to the residual solvent peaks, together with coupling constants (J). The mass spectrometry analysis was measured on Velos Ion Trap Mass spectrometer (Thermo Scientific).

5’-*O*-(*tert*-butyldimethylsilyl)-2’-deoxyguanosine (**2**)

tert-Butyldimethylsilyl chloride (*t*-BDMSCl) (750 mg, 5 mmol) and imidazole (410 mg, 6mmol) were added to a solution of 2’-deoxyguanosine monohydrate (co-evaporated with anhydrous pyridine (2 × 10 mL)) (1.07 g, 4.0 mmol) in dimethylformamide (DMF) (15 mL). The reaction mixture was stirred at room temperature for 5 h, then poured into ethyl acetate (300 mL), and washed with saturated aq NaCl (3 × 300 mL), and dried over Na_2SO_4. The organic layer was concentrated under reduced pressure, and the resulting residue was purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (0-6% MeOH 3CV’s, 6-10% MeOH 5 CV’s, 10% MeOH 10 CV’s, 1 CV= 33 mL) yielding product **2** (1.32 g, 87%) as a white solid. ^{1}H NMR (400 MHz, DMSO-*d6*) δ 10.64 (s, 1H), 7.83 (s, 1H), 6.49 (s, 2H), 6.11 (dd, *J* = 7.4, 6.1 Hz, 1H), 5.33 (d, *J* = 4.1 Hz, 1H), 4.32 (dd, *J* = 6.0, 3.1 Hz, 1H), 3.84–3.81 (m, 1H), 3.75–3.65 (m, 2H), 2.48–2.44 (m, 1H), 2.26–2.20 (m, 1H), 0.85 (s, 9H), 0.02 (s, 6H). ^{13}C NMR (100 MHz, DMSO-*d6*) δ 156.80, 153.74, 150.93, 134.87, 116.62, 86.98, 82.42, 70.38, 63.37, 39.60, 25.86, 18.05, -5.39, -5.40. LTQ-MS: Cal: 381.18, Found: $[M+H]^+$: 382.27.

3’-*O*-propargyl-5’-*O*-(*tert*-butyldimethylsilyl)-2’-deoxyguanosine (**3**)

To a solution of compound **2** (1.2 g, 3.2 mmol) in tetrahydrofuran (THF) (10 mL), sodium hydride (60 % dispersion in mineral oil, 400 mg; 10 mmol) was added. The suspension was stirred at 0 °C for 15 min. Propargyl bromide (80% in toluene, 377 µL, 3.5 mmol) was added and the reaction was stirred at room temperature for another 6 h. Saturated aqueous $NaHCO_3$ (10 mL) was added to quench the reaction. Then THF in the mixture was removed under reduced presure and the resulting mixture was poured into ethyl acetate (150 mL). The solution was then washed with saturated aqueous $NaHCO_3$ (150 mL) and saturated aq NaCl (2 × 150 mL), and dried over anhydrous Na_2SO_4. The organic layer was concentrated under reduced presure, and the resulting residue was subjected to flash chromatography on a biotage system using a DCM: MeOH gradient (0-6% MeOH 8 CV’s, 6% MeOH 8 CV’s, 1 CV= 33 mL) yielding product **3** (960 mg, 72%) as a white foamy solid. ^{1}H NMR (400 MHz, DMSO-*d6*) δ 10.57 (s, 1H), 7.81 (s, 1H), 6.42 (s, 2H), 6.06 (dd, *J* = 8.4, 5.7 Hz, 1H), 4.34 (dt, *J* = 5.3, 2.3 Hz, 1H), 4.24 (d, *J* = 2.4 Hz, 2H), 3.99 (td, *J* = 5.0, 2.0 Hz, 1H), 3.70 (dd, *J* = 5.0, 1.4 Hz, 2H), 3.46 (t, *J* = 2.4 Hz, 1H), 3.25 (s, 1H), 2.60 (m, 1H), 2.44 (m, 1H), 0.81 (s, 9H), 0.00 (s, 6H). ^{13}C NMR (100 MHz, DMSO-*d6*) δ 157.17, 154.23, 151.46, 135.20, 117.11, 84.58, 82.92, 80.57, 78.89, 77.74, 63.59, 56.28, 36.36, 26.26, 18.40, -4.98, -5.03. LTQ-MS: Cal: 419.20, Found: $[M+H]^+$: 420.25.

3’-*O*-propargyl-2’-deoxyguanosine (**4**)

Tetrabutylammonium fluoride trihydrate (790 mg, 2.5 mmol) was added to a solution of compound **3** (720 mg, 1.7 mmol) in THF (3 mL). After stirring for 4 h at RT, the mixture was poured into ethyl acetate (200 mL), and washed with saturated aq NaCl (3 × 200 mL), and dried over Na_2SO_4. The organic layer was concentrated in *vacuo*, and the residue was subjected to flash chromatography on a biotage system using a DCM: MeOH gradient (0-8% MeOH 8 CV’s, 8% MeOH 8 CV’s, 1 CV= 33 mL) yielding product **4** (470 mg, 91%) as a white solid. ^{1}H NMR (400 MHz, DMSO-*d6*) δ 10.64 (s, 1H), 7.92 (s, 1H), 6.46 (s, 2H), 6.05 (dd, *J* = 8.7, 5.7 Hz, 1H), 5.06 (t, *J* = 5.5 Hz, 1H), 4.32 (dt, *J* = 5.5, 2.0 Hz, 1H), 4.23 (d, *J* = 2.4 Hz, 2H), 3.97 (td, *J* = 4.7, 1.8 Hz, 1H), 3.53 (m, 2H), 3.46 (t, *J* = 2.4 Hz, 1H), 2.58 (m, 1H), 2.42 (m, 1H). ^{13}C NMR (100 MHz, DMSO-*d6*) δ 156.68, 153.66, 150.90, 135.22, 116.61, 84.71, 82.58, 80.24, 78.94, 77.19, 61.64, 55.74, 35.97. LTQ-MS: Cal: 305.11, Found: $[M-H]^-$: 304.16.

3’-*O*-propargyl-2’-deoxyguanosine triphosphate (prop-dGTP) (**5**)

Phosphorous oxychloride (20 µL , 0.20 mmol) was added dropwise under nitrogen atmosphere to a cooled solution (0 ^oC) of compound **4** (66 mg, 0.15 mmol) and 1,8-bis(dimethylamino) naphthalene (proton sponge, 64 mg, 0.30 mmol) in trimethyl phosphate (2 mL). The reaction mixture was stirred for 30 min at 0 ^oC. A solution of tributylammonium pyrophosphate (110 mg, 0.20 mmol) in dry DMF (1.0 mL) and tributylamine (50µL, 0.20 mmol) was added. The reaction mixture was stirred for another 30 min, followed by quenching with triethylammonium bicarbonate (TEAB, 0.1 M, 10 mL). The mixture was allowed to warm to RT while stirring for 2 h. Solvents then were removed by cryodesiccation. The resulting residue was dissolved in H_2O and subject to RP-HPLC, (A, 0.05 M TEAA in H_2O; B, acetonitrile; B, 5-15% in 20 min, 80% in 8 min 80-5% in 2 min) yielding compound **5** as the triethylammonium salt. ^1H NMR (400 MHz, *D_2O*) δ 8.11 (s, 1H), 6.25 (dd, J = 6.0, 9.2 Hz , 1H), 4.63 (m, 1H), 4.37 (m, 1H), 4.22 (m, 1H), 4.16 (m, 2H), 3.20 (q, J = 7.4 Hz, 18H, TEA), 2.94 (t, J = 2.5 Hz, 1H), 2.83 (m, 1H), 2.65 (m, 1H), 1.27 (t, J = 7.4 Hz, 27H, TEA). ^{13}C NMR (100 MHz, *D_2O*) δ 158.84, 153.83, 151.46, 137.62, 116.09, 83.82, 80.04, 79.36, 76.14, 65.82, 59.35, 56.71, 46.59, 36.22, 8.19. ^{31}P NMR (162 MHz, *D_2O*) δ -11.91 (d, J = 19.9 Hz), -13.11 (d, J = 19.6 Hz), -24.85 (d, J = 20.0 Hz) LTQ-MS: Cal 545.01, Found: $[M+H]^+$: 546.01.

Experimental section

Oligonucleotide synthesis

5'-Azido modified oligonucleotides were synthesized on a MerMade 4 Oligonucleotide synthesizer (BioAutomation Corporation, USA). The modified nucleotides were site-specifically incorporated at the desired positions (Supplementary Table 1) with reagents obtained from Glen Research (Sterling, VA, USA). Post-synthesis modifications were carried out on solid phase directly[52]. Subsequent purification was carried out with high-performance liquid chromatography (HPLC). Purified oligonucleotides were characterized by electrospray linear ion trap mass spectrometry.

Click-code-seq steps I-III: 8-oxoG excision, single-nucleotide incorporation and click labelling

The removal of 8-oxoG from oligonucleotides was carried out with 1 µM pre-annealed IL1-oxoG and IL1-T30 duplex, 1 µL Fpg (New England Biolabs (NEB), 8 U/ µL) and 1 µL APE 1 (NEB, 10 U/µL) in 50 µL 1 × NEBuffer 2.1 at 37 °C for 1 h. The reaction mixture was then heated (60 °C, 10 min) to inactive Fpg and APE1. After allowing to return to room temperature, 1 µL of prop-dGTP (10 mM) and 0.2 µL of therminator IX (NEB, 10 U/ µL) was added and the mixture was heated (60°C, 10 min).The resulting DNA was purified with a micro bio-spin 6 column (Bio-Rad), completely dried on a vacuum concentrator and re-suspended in 2 µL water. Next, 3 µL L19-azido (200 µM), 1 µL potassium phosphate buffer (1 M, pH 7.0), 1 µL aminoguanidine hydrochloride solution (50 mM), 1 µL DMSO, 1 µL sodium ascorbate (25 mM) and 1 µL premixed $CuSO_4$:THPTA (1:6, 5 mM in concentration of Cu^{2+}) were add and incubated at room temperature (30 min). Aliquots were withdrawn after each step, quenched with 90 % formamide, and analyzed by 20 % (wt/vol) polyacrylamide, 8 M urea gel electrophoresis (urea-PAGE). Gels were imaged with a ChemiDoc XRS^{+} System (Bio-Rad). Band intensities were quantified with Image Lab (Bio-Rad).

Confirmation of code insertion by sanger sequencing

DNA fragments containing a site-specific 8-oxoG modification were prepared by ligation of short 8-oxoG modified oligonucleotides and a 0.9-kb gapped DNA duplex. Oligonucleotides and primers are presented in Supplementary Table 1. In brief, a pEGFP-N1 plasmid was constructed to contain two Nb.BbvCI, a nicking endonuclease, cleavage sites (Fig. S4). A 0.9 kb DNA duplex (sequence is listed below) was amplified from the plasmid with primers GFP-Pr409 and GFP-Pr1296 and Taq DNA polymerase

(NEB), and then subjected to Nb.BbvCI digestion to generate a gapped duplex. The 20-mer cleaved single-stranded DNA was removed by annealing with a 20-mer 8-oxoG modified complementary oligonucleotide (GFP-oxoG639) in large excess. The duplex with two nick sites was ligated with T4 DNA ligase at 16 °C for 4 h and purified with Monarch Nucleic Acid Purification Kit (NEB). The 8-oxoG excision, single-nucleotide incorporation and click-labelling of resulting DNA were carried out according the protocol described above. Next, PCR amplification was carried out with primers GFP-Pr1296 and IL1-P20 and Taq DNA polymerase. PCR products were purified and sequenced by Sanger sequencing (Microsynth AG) with primer GFP-Pr1296.

Click-code-seq library construction to map 8-oxoG in S. cerevisiae

Yeast strain BY4741 (genotype: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) was purchased from Dharmacon. Genomic DNA from *S. cerevisiae* cells was grown overnight in liquid rich medium containing yeast extract, peptone and dextrose (YPD), then extracted with a DNA Isolation Kit for Cells and Tissues (Roche Holding AG) according to the manufacturer's instructions. To minimize adventitious 8-oxoG formation, antioxidants (100 mM deferoxamine, 100 mM butylated hydroxytoluene) were used in all reactions until the 8-oxoG excision step. Genomic DNA (5 µg) was sheared in 130 µL Tris-EDTA buffer with a Covaris S220 ultrasonicator (Covaris Inc.) using the following parameters: peak incident power 140 W, cycles/burst 200, duty factor 10% and time 100 s. DNA concentration and distribution were estimated using the Nanodrop 8000 and Agilent 2200 Tapestation with high sensitivity D1000 screen tape.

The fragments were treated with APE 1 (final concentration (fc): 0.2U/µL) in 1 × NEBuffer 2.1 at 37 °C for 1 h to remove abasic sites. Dideoxynucleotides (Jena Bioscience GmbH, fc: 200 µM) and therminator IX (NEB, fc: 0.03 U/µL) were added and heated (60 °C, 10 min) to block gaps and terminal nucleotides in DNA fragments. The product was purified using the Monarch nucleic acid purification kit and subjected to steps I-III of click-code seq (8-oxoG excision, single-nucleotide incorporation and click labelling steps) according to the protocol described above. Biotin-modified L-P5-azido was used for click labelling in strand L19-azido as described above. After purification, DNA (50 µL in TE buffer) was heated (95 °C, 2 min) and the tube was quickly transfered to an ice bath. DNA was subjected to slow rotation with Dynabeads MyOne C1 (Invitrogen Corp.) in bead-binding buffer (5 µL, 5.0 mM Tris-HCl (pH 7.5), 1.0 mM EDTA, 1 M NaCl) at room temperature. for 30 min. Afterwards, the beads were washed twice with bead-washing buffer (200 μL, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.2 M NaCl) and resuspended in 49 μL reaction mixture (5 μL T4 PNK reaction buffer (10 ×), 5 μL ATP (10 mM) and 39 μL ddH_2O). After adding 1 μL T4 PNK (NEB), the mixture was warmed at 37 °C for 30 min. Next, the beads were washed two times with 200 μL bead-washing buffer and resuspended in 49 μL reaction mixture (5 μL T4 ligase reaction buffer (10 ×), 5 μL PEG-4000 (50 %), 0.5 μL TWEEN-20 (1 %), 3 μL pre-annealed dsDNA adapter (40 μM)). After adding 1 μL T4 ligase (NEB), the mixture was incubated at 16 °C for 2 h. After adaptor ligation, the beads were washed twice with 200 μL bead-washing buffer, resuspended in 20 μL ddH_2O and heated at 95 °C for 2 min. The beads were pelleted immediately using a magnetic rack, and the supernatant, which contained the single-stranded library molecules, was transferred to a fresh tube.

The entire library was used for indexing with one of the barcoded primers, P7-index01-03, and P5-uni using Vent exo$^-$ polymerase (NEB) in 1 × Thermopol buffer (NEB) for 5 cycles, involving denaturation for 20 s at 95 °C, annealing for 20 s at 64 °C and primer extension for 60 s at 72 °C, final extension for 5 min at 72 °C, and hold at 4 °C. Use of vent exo$^-$ polymerase ensured efficient bypass of the triazole produced by click labelling. (Figure S2) The indexed library was purified using an SPRIselect kit (Beckman Coulter) according to the manufacturer's instructions. After indexing, the library was amplified with Pr-P5-2nd and Pr-P7-2nd using Q5 high fidelity DNA polymerase (NEB) and purified with an SPRIselect kit. The library size was verified using the Agilent 2200 tapestation with high sensitivity D1000 screen tape before being loaded onto the Illumina Miseq platform (Illumina) with a single-end 150 bp sequencing mode. Three biological replicates of yeast genomic DNA were prepared and sequenced according to the same procedure.

Sequence alignment and data analysis

Raw reads were aligned to the *S. cerevisiae* genome (S288C, *Saccharomyces* Genome Database) with Bowtie 2 with end-to-end mode. Reads in this study had an eight-base code sequence incorporated during click labelling, corresponding to the first eight cycles of raw FASTQ sequence. To remove these sequences, --trim5 8 option was used during bowtie 2 alignment with other default settings (bowtie2 --end-to-end --threads 1 -x *reference_genome* -U *input.fastq* --trim5 8 -S *output.sam*). The produced SAM files were sorted and converted to BAM files with Sambamba (sambamba view -S *input.sam* -f bam -l 0 -o /dev/stdout -t 1|sambamba sort /dev/stdin -o /dev/stdout t 1

-l 0 -m 5GB --tmpdir=TMPDIR > *output.bam*). Then, PCR duplicates in BAM files were removed with samtools using rmdup command (samtools rmdup *input.bam output.bam*). Next, bam files were converted to bed files directly (bedtools bamtobed -i *input.bam* > *output.bed*). With the bedtools genomecov command, the depth at each genome position with 1-based coordinates was reported in bed files, which were used to generate the genome wide map of 8-oxoG (genomeCoverageBed -d -i *input.bed* -g *reference_genome* > *output.bed*). Based on the strand tag (+/-), 8-oxoG sites were calculated using the following rules: for read with '+' tag, the 8-oxoG was located at the start site of this read; for reads with '–' tag, the 8-oxoG was located at the end site of this read. The bed files with single resolution data were generated with the above rules, and used for further analysis.

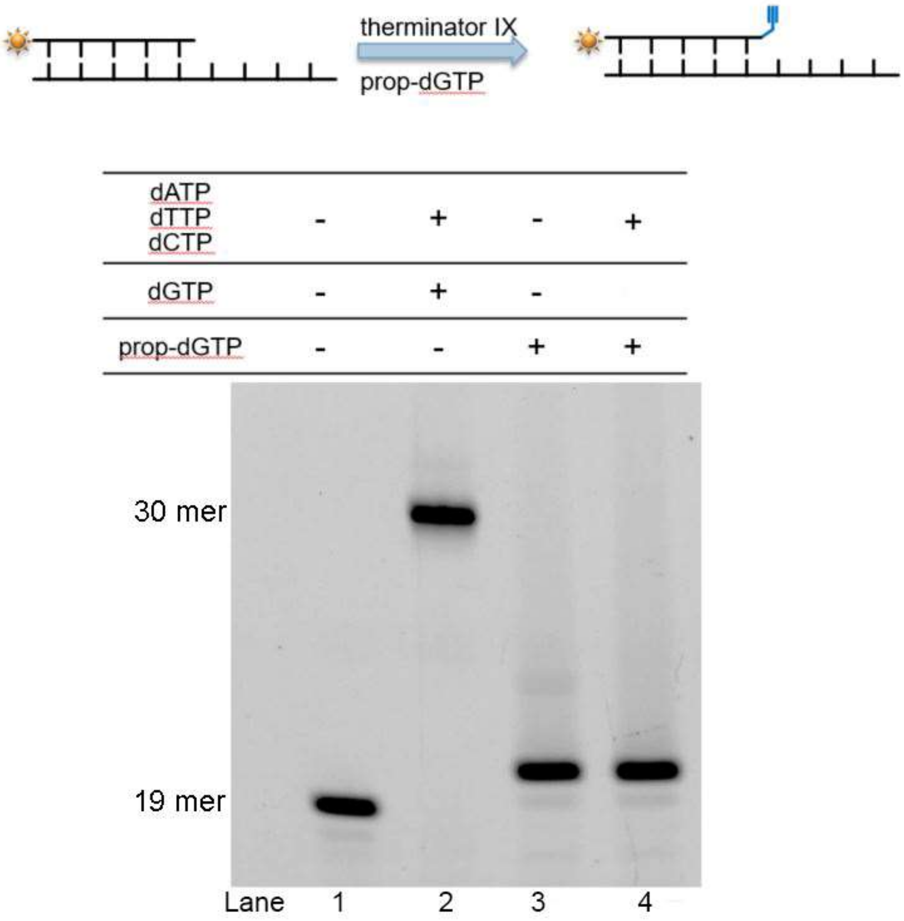
The 8-oxoG frequencies present in each genomic feature were calculated with bedtools intersect command (bedtools intersect -a *input.bed* -b *genomic_feature.bed* > *output.bed*). The 8-oxoG frequencies at each site of the genome feature were called with the bedtools coverage command and Excel power pivot (bedtools coverage -a *genomic_feature.bed* -b *input.bed* -d > *output.bed*). Sequence logo analysis was carried out with bedtools (bedtools getfasta -fi *reference_genome* -bed *inpout.bed* -s -fo *output.bed*) and Weblogo (https://github.com/WebLogo) (seqlogo -f *input.bed* -k 1 -o *output.eps*). All of the data described as "relative 8-oxoG frequency" in the manuscript were calculated based on following equation: Relative 8-oxoG frequency = (8-oxoG counts in genome feature/genome feature size(bp))/(8-oxoG counts in whole genome/whole genome size(bp)). Graphs were made using GraphPad Prism 7.03 (GraphPad Software), Origin 9.1 (OriginLab Corp.) or Rstudio 1. All error bars in this study represent standard deviation. Statistical significance was performed with one-way ANOVA for Figure 2a and two-way ANOVA for Figure 2b. Statistical symbol meaning: **** $P < 0.0001$, *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$.

External data

The TSS and TTS data sets were obtained from Jiang et al.[53]. The 5'-UTR, 3'-UTR, ORF and gene transcription level data sets were obtained from Miura et al.[54]. The nucleosome data set was obtained from Jing et al.[53] and Miura et al.[54]. The DNase I hypersensitive site data set was obtained from Hesselberth et al.[55]. The telomere data set was obtained from the *Saccharomyces* genome database. The RNA pol II occupancy was obtained from Liu et al.[56]. The histone modification data set was obtained from Pokholok et al.[42] The transcription binding site data set was obtained from Kasinathan et al.[57]

Sequence of the dsDNA fragment used for Sanger sequencing

TGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTT
TTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATT
GACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTG
GTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGACTCAGATCTCGAGCTCAA
GCTTCCTCAGCCCGCCCGGGACTGCCTCAGCGGATCCACCGGTCGCCACCATG
GTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGC
TGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGG
CGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGC
TGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGC
TTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATG
CCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTA
CAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATC
GAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCT
GGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAA
CGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGC
AGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCT
GCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCA
ACGAGAA



dATP dTTP dCTP	-	+	-	+
dGTP	-	+	-	
prop-dGTP	-	-	+	+

Figure S1. Selective incorporation of prop-dGTP opposite C with therminator IX polymerase. Lane 1: Marker band, no dNTP was included, band corresponds to primer (L19-Fam). Lane 2: Positive control, all dNTPs (dATP, dTTP, dCTP, dGTP), band corresponds to full-length product. Lane 3: Only prop-dGTP, band corresponds to primer with the addition of prop-dG. Lane 4: prop-dGTP, dATP,dTTP, dCTP were added, band is the same as lane 3. These results indicate that prop-dGTP was selectively added opposite C by therminator IX polymerase even in the presence of other dNTPs. prop-dGTP acts as a chain terminator. Products were analyzed by 20 % (wt/vol) polyacrylamide, 8 M urea gel electrophoresis (urea-PAGE). After electrophoresis, a gel image was obtained with ChemiDoc XRS$^+$ System (Bio-Rad).

F-L15-prop: 5’-FAM-GCCACATCTTCAATGprop
F-T34-Tri: 5’-FAM-GCCACATCTTCAATG[triazole]TCGGAAGAGCACACGTCTG
P19: 5’-CAGACGTGTGCTCTTCCGA

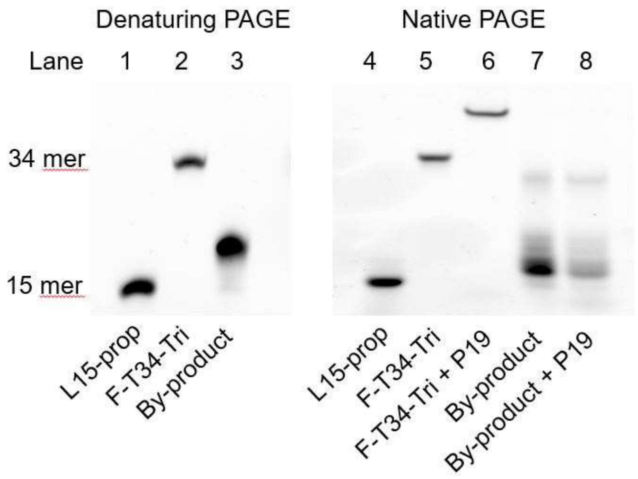


Figure S2. Impact of by-product on library preparation. The by-product from the click reaction did not hybridize to the complementary primer of the full-length product and thus does not impact the library preparation. The full-length product (F-T34-Tri) and by-product were synthesized through the click reaction using the F-L15-prop and an azido modified sequences. The products were purified and isolated with denaturing PAGE (results not shown). Following purification, the length and purity were re-checked by denaturing PAGE gel (left gel, lane 1 and lane 2) confirming the presence of the full-length product (F-T34-Tri) and the same click by-product. Next, 1 μM of the purified F-T34-Tri or the by-product were annealed with 2 μM of the complementary primer P19 in 1 x Thermopol buffer. Native PAGE gel was used to show hybridization of complementary sequences. As expected, hybridization of F-T34-Tri and P19 resulted in a slower migrating band (Lane 6) compared to the full length product F-T34-Tri alone (Lane 5). However, no notable migration changes were observed for by-product with or without P19 (Lane 7 and 8), which indicated that the byproduct is not complementary to P19 and thus the primer used in library presentation could not bind to the by-product and alter the library preparation and sequencing results. Denaturing gel was performed with 20 % (wt/vol) polyacrylamide with 8 M urea (urea-PAGE). Native gel was performed with 15 % (wt/vol) polyacrylamide gel. After electrophoresis at R.T., gel images was obtained with ChemiDoc MP imaging system (Bio-Rad).

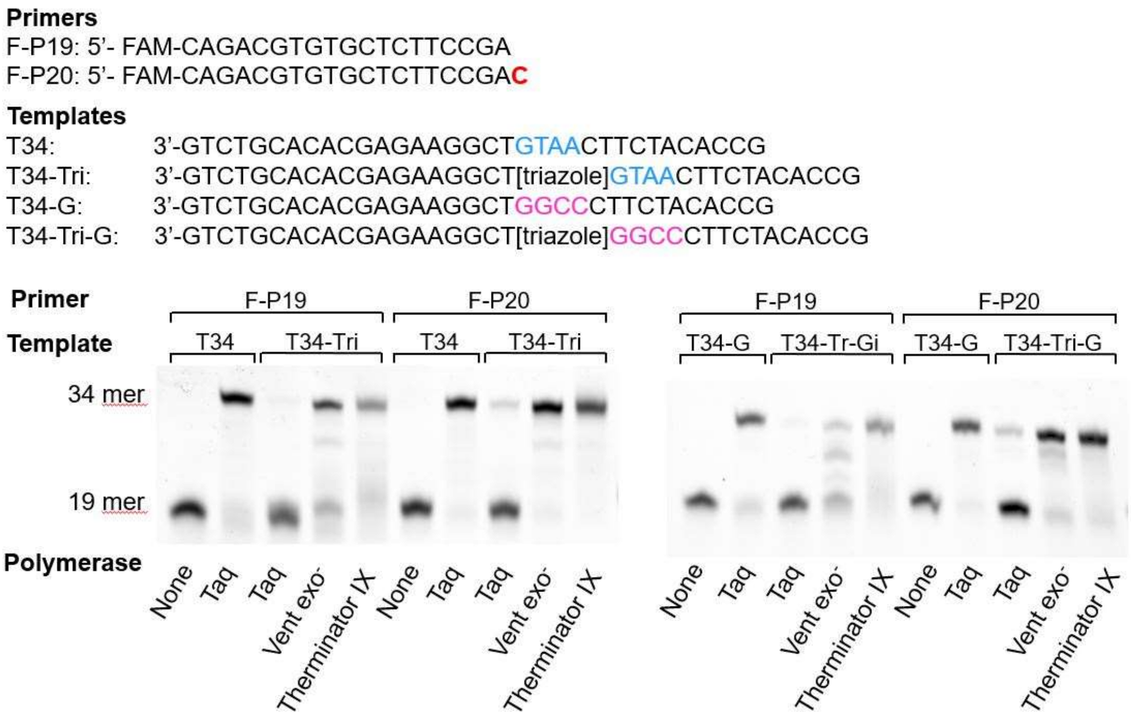

Primers
F-P19: 5'- FAM-CAGACGTGTGCTCTTCCGA
F-P20: 5'- FAM-CAGACGTGTGCTCTTCCGAC
Templates
T34: 3'-GTCTGCACACGAGAAGGCTGTAACTTCTACACCG
T34-Tri: 3'-GTCTGCACACGAGAAGGCT[triazole]GTAACTTCTACACCG
T34-G: 3'-GTCTGCACACGAGAAGGCTGGCCCTTCTACACCG
T34-Tri-G: 3'-GTCTGCACACGAGAAGGCT[triazole]GGCCCTTCTACACCG
Primer
F-P19
F-P20
F-P19
F-P20
Template
T34
T34-Tri
T34
T34-Tri
T34-G
T34-Tr-Gi
T34-G
T34-Tri-G
34 mer
19 mer
Polymerase
None
Taq
Taq
Vent exo-
Therminator IX
None
Taq
Taq
Vent exo-
Therminator IX
None
Taq
Taq
Vent exo-
Therminator IX
None
Taq
Taq
Vent exo-
Therminator IX

Figure S3. Bypass of triazole linkage. Primers with different lengths and two triazole-modified templates with different sequences (GTAA and GGCC) were tested with three different polymerases: Taq, Vent exo$^-$ and Therminator IX. These results indicate that the bypass efficiency through the triazole linkage varied depending on which polymerase was used and primer length but not sequence. Bypass was more efficient for primers overhanging the triazole linkage (F-P20). Both Vent exo- and Therminator IX bypassed the triazole linkage without significant difference between two different templates, especially with one base overhang. Products were analyzed by 20 % (wt/vol) polyacrylamide, 8 M urea gel electrophoresis (urea-PAGE). After electrophoresis, a gel image was obtained with a ChemiDoc MP imaging system (Bio-Rad). Based on these results, we decided to use Vent exo$^-$ polymerase and a primer with one dC overhang (i.e. P5-uni in Table 1) for further library preparation.

Primer

F-P20: 5'- FAM-CAGACGTGTGCTCTTCCGAC

Templates

T34: 3'-GTCTGCACACGAGAAGGCTGTAACTTCTACACCG
T34-Tri: 3'-GTCTGCACACGAGAAGGCT[triazole]GTAACTTCTACACCG
T34-Tri-G: 3'-GTCTGCACACGAGAAGGCT[triazole]GGCCCTTCTACACCG

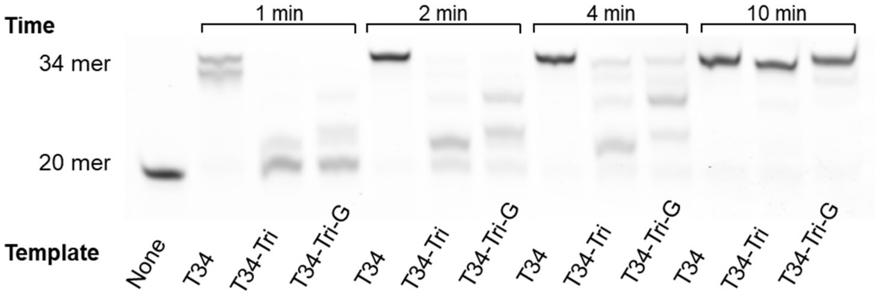


Figure S4. Bypass of the triazole linkage by Vent exo$^-$ polymerase. Aliquots were removed from the extension reaction at varying time points, quenched and analysed by electrophoresis on a 20 % (wt/vol) polyacrylamide, 8 M urea gel electrophoresis. After electrophoresis at R.T., gel images were obtained with ChemiDoc MP imaging system (Bio-Rad). Several truncated products were observed after 1, 2 or 4 min extension time for bypass of triazole-linked templates, but after 10 min, all truncated products were fully extended. These data indicate that the shorter by-products during Vent exo- extension were mainly caused by polymerase pausing rather than DNA slippage. The truncations due to pausing would not introduce bias as in the case of slippage products.

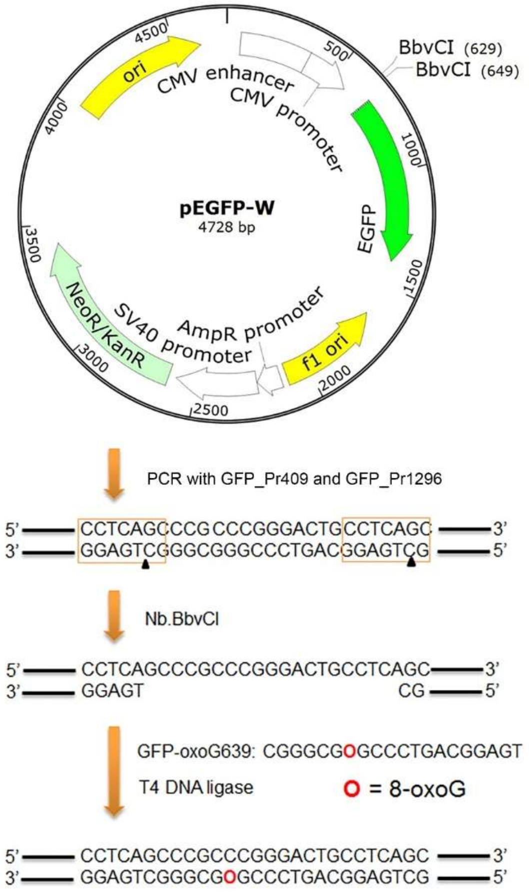

BbvCI (629)
BbvCI (649)
500
1000
1500
2000
2500
3000
3500
4000
4500
ori
CMV enhancer
CMV promoter
EGFP
f1 ori
AmpR promoter
SV40 promoter
NeoR/KanR
pEGFP-W
4728 bp
PCR with GFP_Pr409 and GFP_Pr1296
5' CCTCAGCCCGCCCGGGACTGCCTCAGC 3'
3' GGAGTCGGGCGGGCCCTGACGGAGTCG 5'
Nb.BbvCI
5' CCTCAGCCCGCCCGGGACTGCCTCAGC 3'
3' GGAGT CG 5'
GFP-oxoG639: CGGGCGOGCCCTGACGGAGT
T4 DNA ligase
O = 8-oxoG
5' CCTCAGCCCGCCCGGGACTGCCTCAGC 3'
3' GGAGTCGGGCGOGCCCTGACGGAGTCG 5'

Figure S5. The plasmid map of pEGFP-W used in this study to generate the 8-oxoG site-specific modified dsDNA. The plasmid was constructed based on pEGFP-N1 (Addgene, 6085-1). BbvCI endonuclease cutting sites are located at positions 629 and 649. A 0.9 kb DNA fragment was amplified from this plasmid, containing these two BbvCI sites. Nb.BbvCI was used to generate two nick sites on the same strand, leading to a 20-mer gap in dsDNA. The gap was filled and ligated with a 8-oxoG modified oligonucleotides (GFP-oxoG639) and T4 ligase, generating a 8-oxoG site-specific modified dsDNA.

Input	A	B	C
8-oxoG modified dsDNA	60 ng	6 ng	0
native dsDNA	540 ng	594 ng	600 ng
8-oxoG/unmodified bases	10^{-5}	10^{-6}	0

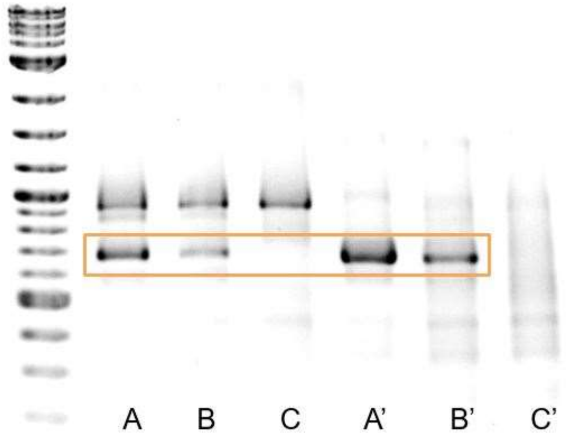


Figure S6. Detection limit study with click-code-seq. To determine whether click-code-seq could be used to detect low amounts of 8-oxoG in a DNA sample, various amounts of 8-oxoG-modified dsDNA (60 ng, 6 ng and 0) were mixed with native dsDNA as listed on the table. Click-code-seq (Steps I-III) was applied as described. The resulting DNA was separated into two: half was PCR amplified with click-code specific primers (A, B, C), and the other half was PCR amplified after biotin enrichment to remove excess native DNA (A', B', C'). The target products are marked by the frame in the figure. The calculation is based on dividing through by all bases (i.e. 1818 nt) of the 0.9 kb dsDNA. The 8oxoG/nt was calculated based on the following equation: 8-oxoG/nt = $= \frac{8-oxoG\ (nmol)}{total\ bases\ (nmol)} = \frac{8-oxoG\ dsDNA\ (ng) / M.W.\ dsDNA}{\left[total\ dsDNA\ (ng) / M.W.\ dsDNA\right] * total\ nt} = \frac{8-oxoG\ dsDNA\ (ng)}{total\ dsDNA\ (ng) * total\ nt}$. For 1818nt, the results are 5.5×10^{-5} and 5.5×10^{-6} for 60 ng or 6 ng of 8-oxoG-modified dsDNA in 600 ng total dsDNA, respectively. These results indicate that the click-code-seq is sufficiently sensitive to label 8-oxoG that is present in native DNA at a frequency as low as 10^{-6} 8-oxoG/ unmodified bases or 10 fmol/200 pM of 8-oxoG in a dsDNA sample. The detection limit of our method is comparable to other methods including an electrochemical immunosensor for 8-oxoG detection (LOD = 105.9 pM) [58] and an aptamer-based electrochemical detection (LOD = 1 pM) [59]. However, highly sensitive nanoLC-NSI-MS/MS systems that lose sequencing information are able to detect 8-oxo-dG orders of magnitude lower (LOD = 0.01 fmol, LOQ = 0.1 fmol) [60].

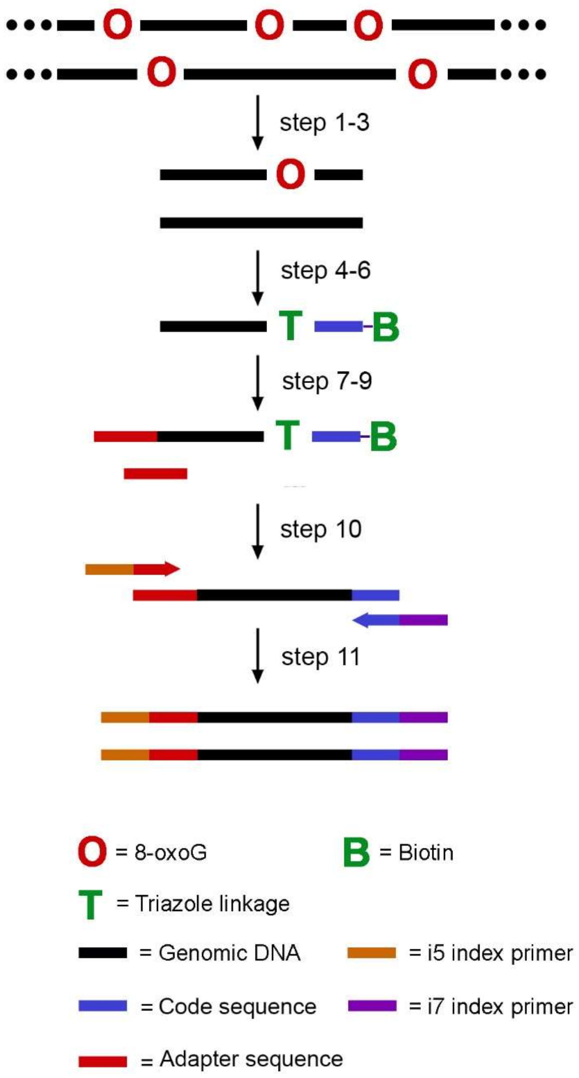


Figure S7. Workflow for click-code-seq library preparation with genomic DNA. Step 1: DNA fragmentation by sonication; Step 2: APE1 treatment to remove abasic sites; Step3: ddNTP blocking; Step 4: Fpg and APE1 treatment; Step 5: prop-dGTP incorporation, Step 6: click reaction with code sequence; Step 7: affinity enrichment by biotin-avidin interaction; Step 8: 5'-phosphorylation; Step 9: adaptor ligation; Step 10: indexing for sequencing; Step 11: final amplification.

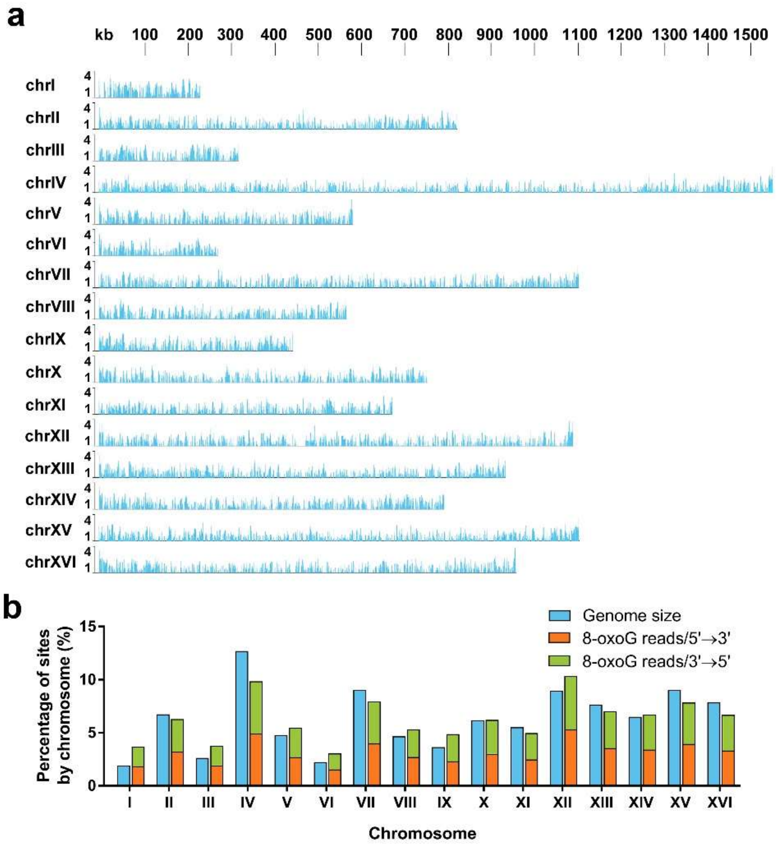


Figure S8. Distribution of 8-oxoG in the yeast genome. a) The click-code-seq map of 8-oxoG in genomic DNA from BY4741 cells (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0). The peaks of 8-oxoG reads are shown for the individual nuclear chromosomes (Chr I–XVI). The height of each peak corresponds to the number of reads (log ratio). b) Chromosomal distribution of 8-oxoG compared to the size of each nuclear chromosome. A comparison of nuclear reads for 5'→3' and 3'→5' strands is also displayed.

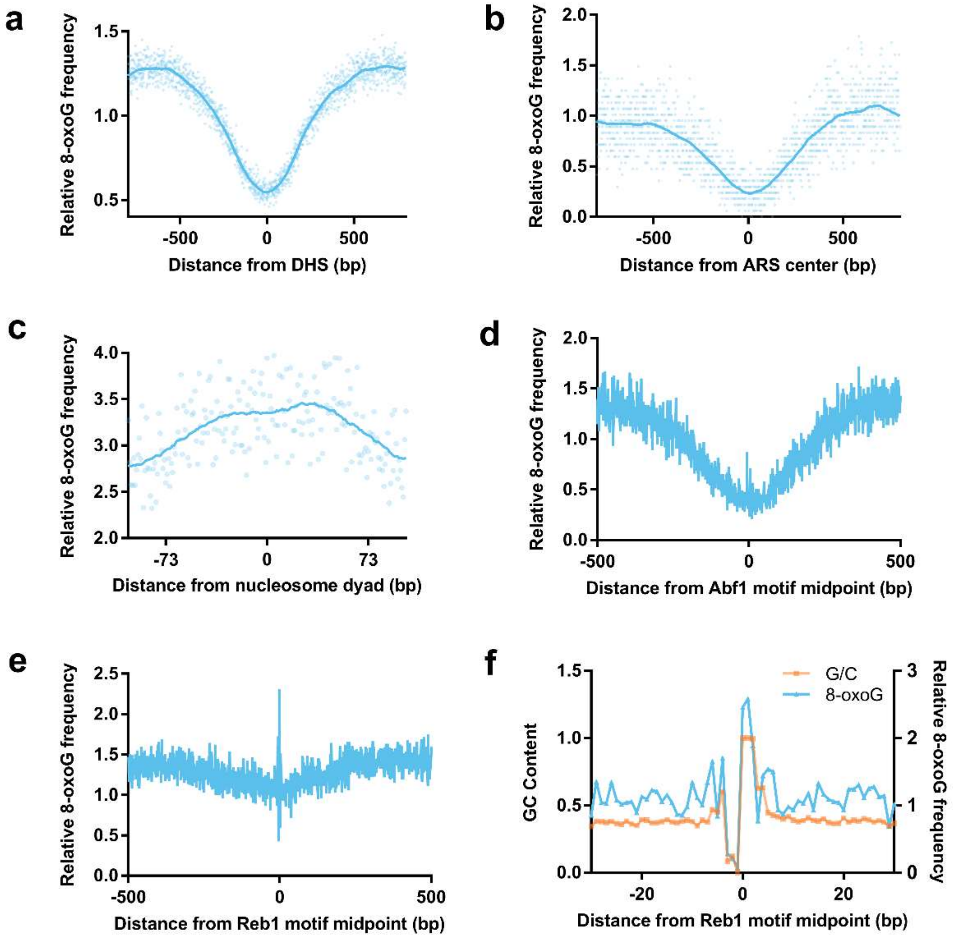


Figure S9. Distribution of 8-oxoG within different genome features. Distribution of 8-oxoG on a) DNase I hypersensitive sites (DHS, n=6,108) and their flanking sequences (800 bp), b) autonomously replicating sequence (ARS, n=348) and their flanking sequences (800 bp), c) nucleosomes (n=4,508). d) Abf1 binding site (n=1068) and e) Reb1 binding site (n=1943). f) Comparison of the distribution of 8-oxoG frequency and GC content around Reb1 binding site. The abnormal fluctuations in the middle of Abf1 and Reb1 binding site are related with the GC context fluctuation.

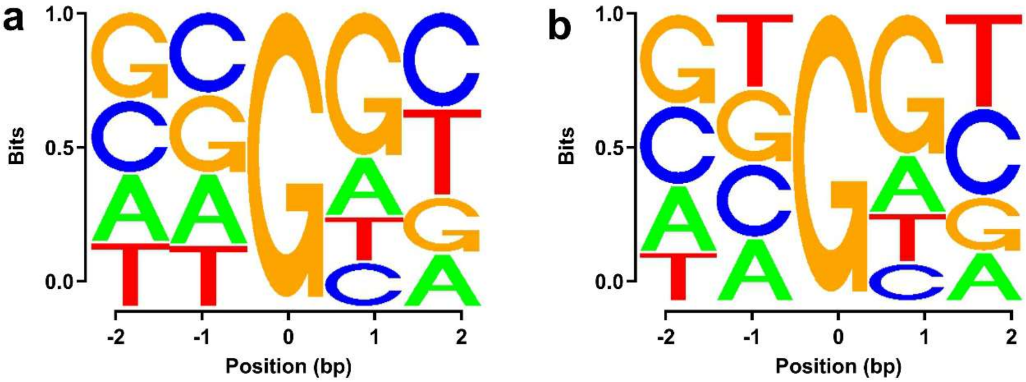
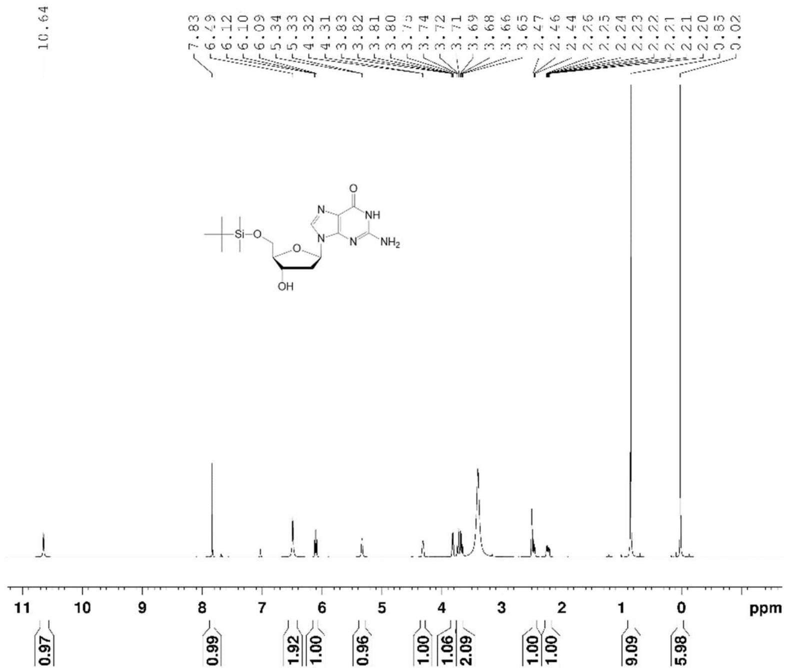


Figure S10. The local sequence preference surrounding 8-oxoG is similar within differing genomic features. Sequence logo plot surrounding 8-oxoG (position 0) within a) transcription start sites (with 100 bp flanking sequence, n = 7,495) and b) nucleosomes (n = 4,545).

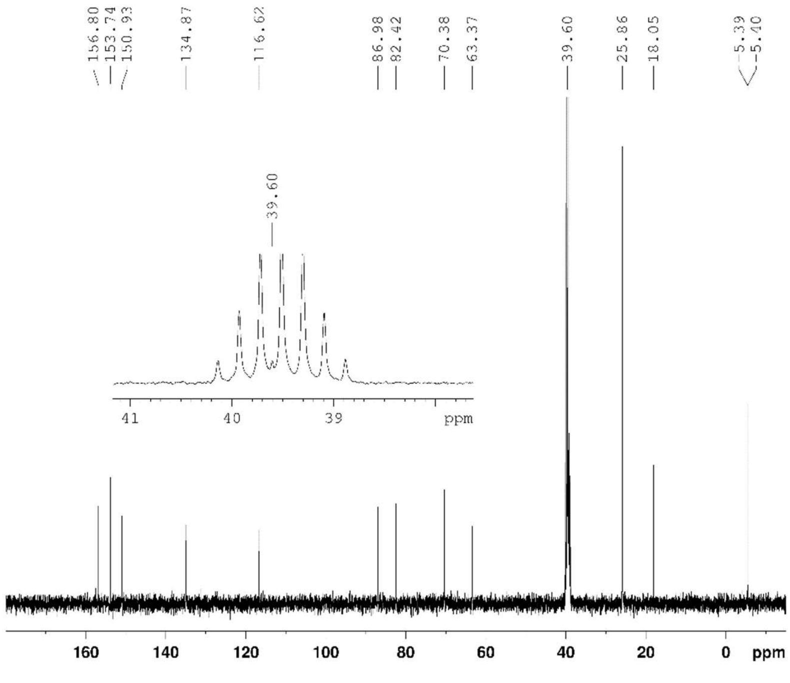
Table S1: **Sequences used in this study**.

Name	Sequence (5′–3′)	Working dilution (µM)
IL1-oxoG	[Fam]GCCACATCTTCAAT[oxoG]TATGCTACAAAAGAT	10
IL1-T30	ATCTTTTGTAGCATACATTGAAGATGTGGC	10
L19-azido	[Azido]TCGGAAGAGCACACGTCTG	200
P19-Fam	[Fam]CAGACGTGTGCTCTTCCGA	10
P19	CAGACGTGTGCTCTTCCGA	10
P20	CAGACGTGTGCTCTTCCGAC	10
P21	CAGACGTGTGCTCTTCCGACA	10
GFP-Pr409	TGACTCACGGGGATTTCCAAGTCT	10
GFP-Pr1296	TTCTCGTTGGGGTCTTTGCTCA	10
GFP-oxoG639	[PHO]TGAGGCAGTCCCG[oxoG]GCGGGC	100
IL1-P20	CAGACGTGTGCTCTTCCGAC	10
L-P5-azido	[Azido]TCGACGGCAGATCGGAAGAGCGTTT[TEG-Biotin] (TEG = triethylene glycol spacer)	200
L-P7	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	40
L-P7-c	NNNNNAGATCGGAAGAG[Phosphate] (N=[A,C,G,T])	40
P5-uni	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCGTCGAC	10
P7-01	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTT CAGACGTGTGCTCTTCCGATCT	10
P7-02	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10
P7-03	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10
Pr-P5-2nd	AATGATACGGCGACCACCG	10
Pr-P7-2nd	CAAGCAGAAGACGGCATACG	10

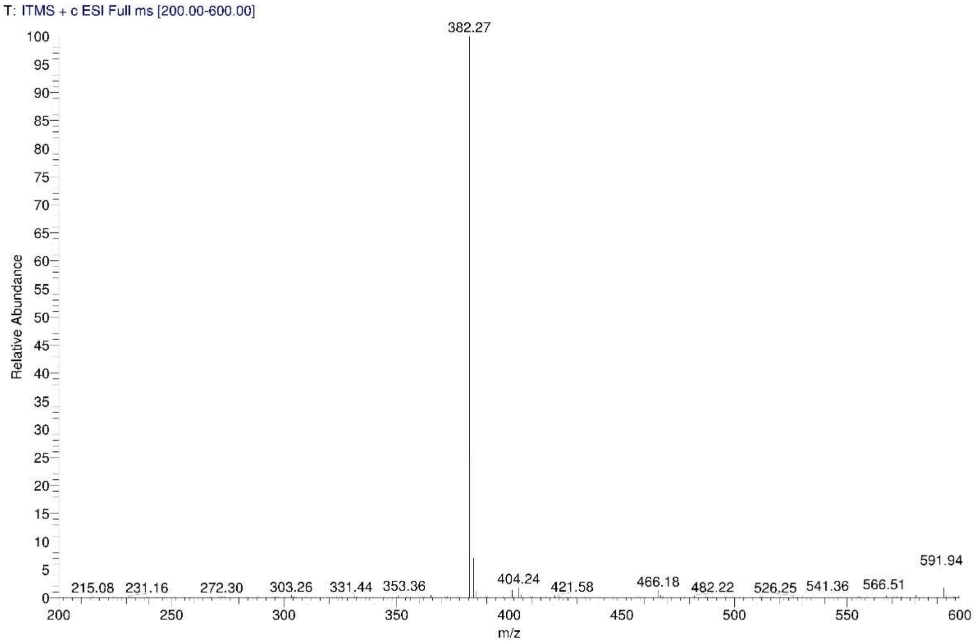
NMR and Mass Spectroscopy Characterization



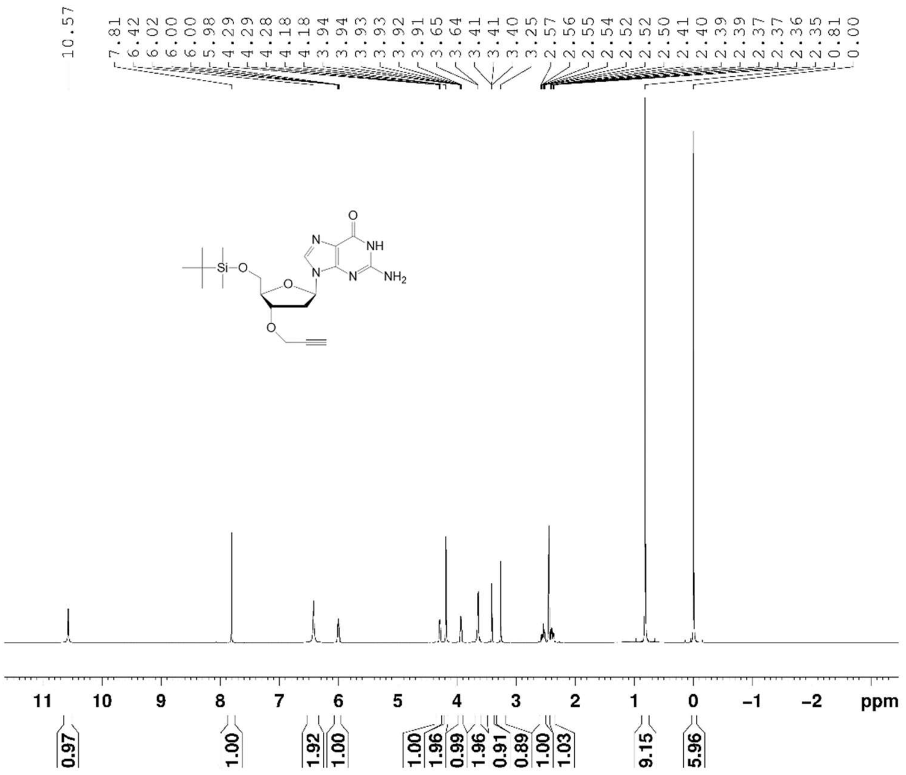
^{1}H NMR of compound 2



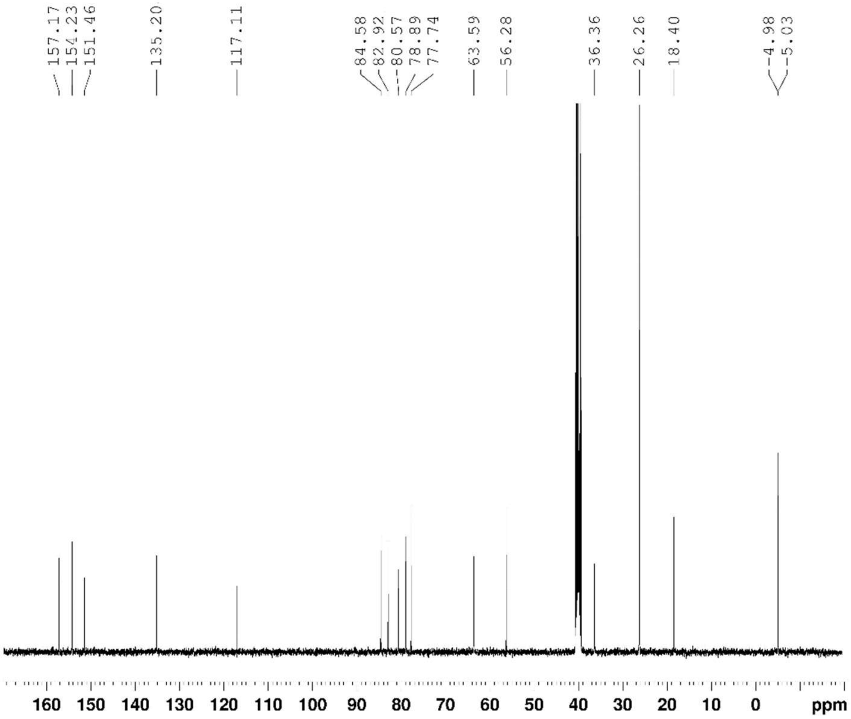
^{13}C NMR of compound 2



Mass spectrometry of Compound **2**

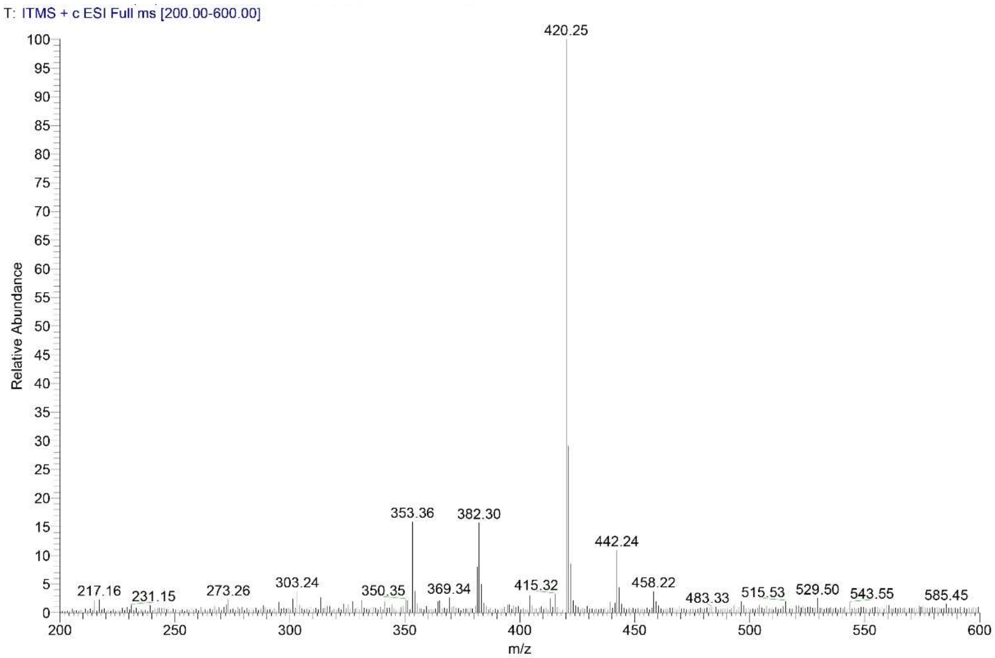


^{1}H NMR of compound 3

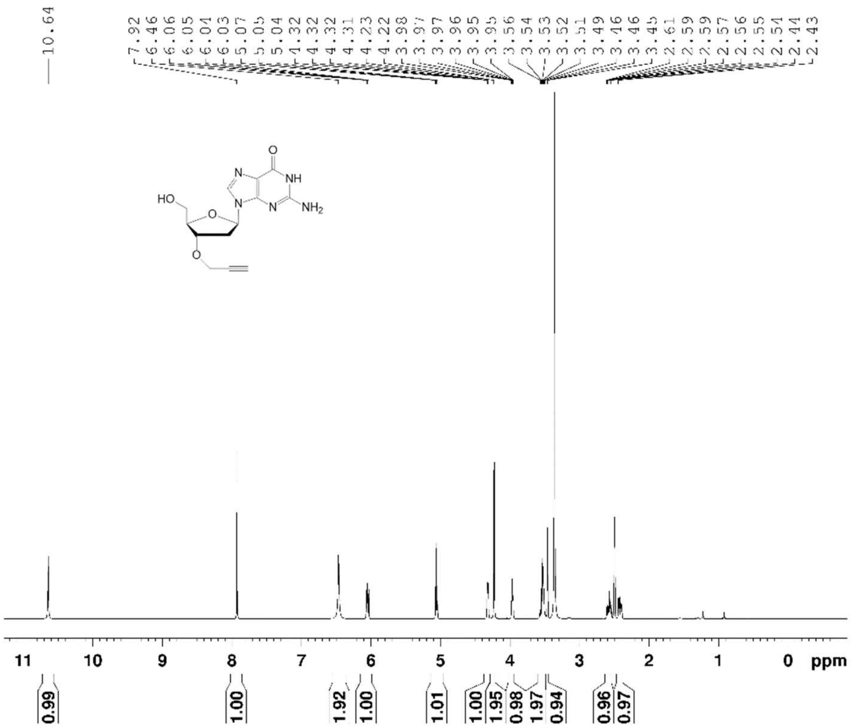


^{13}C NMR of compound 3

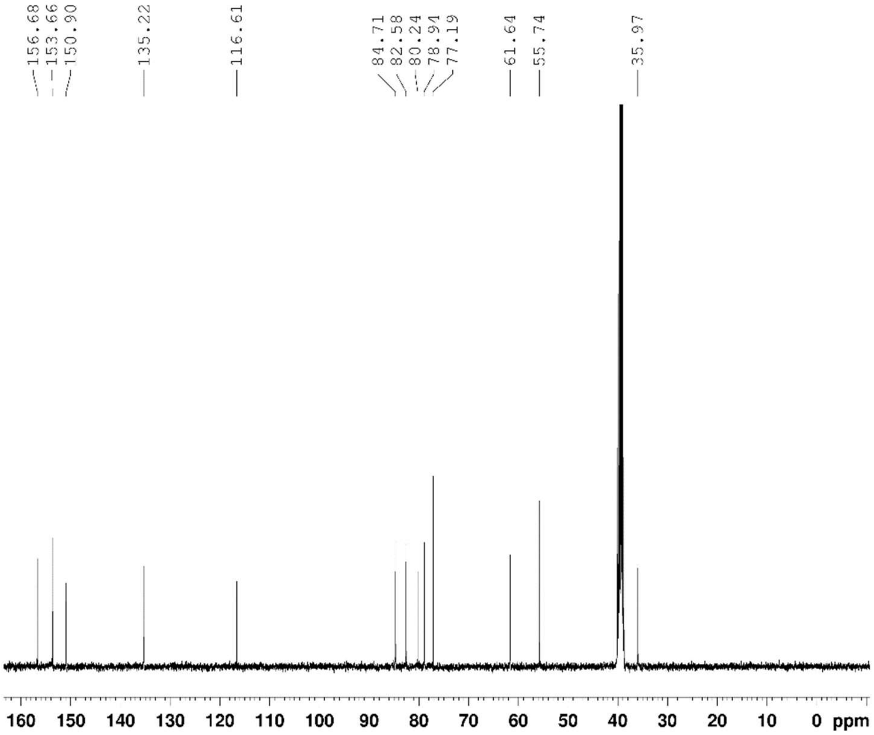
Mass spectrometric analysis of Compound **3**



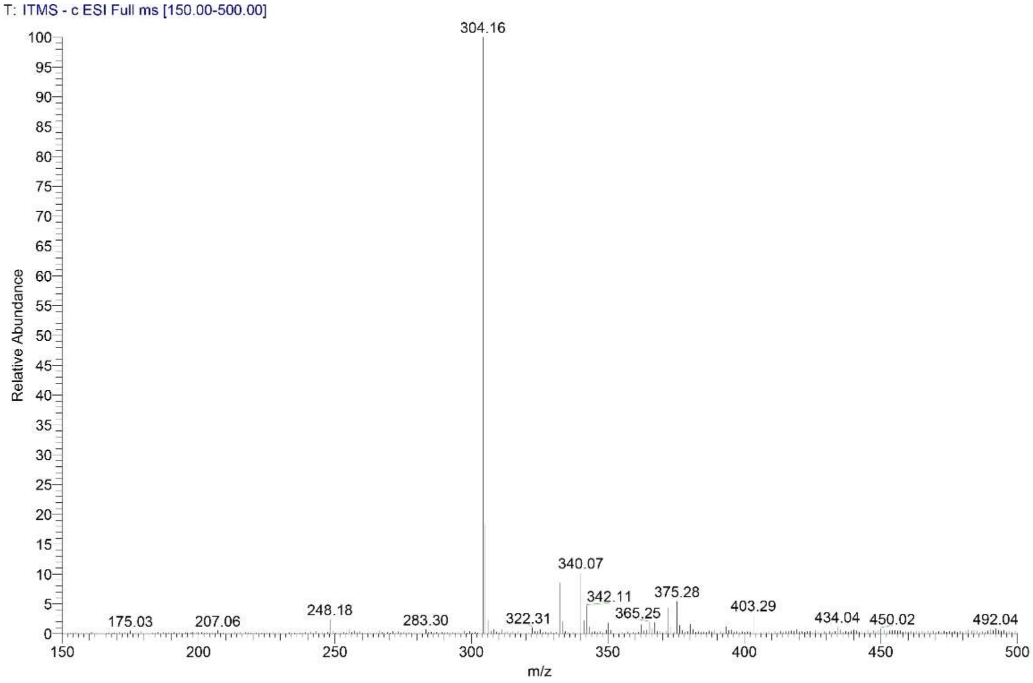
Mass spectrometry of Compound **3**



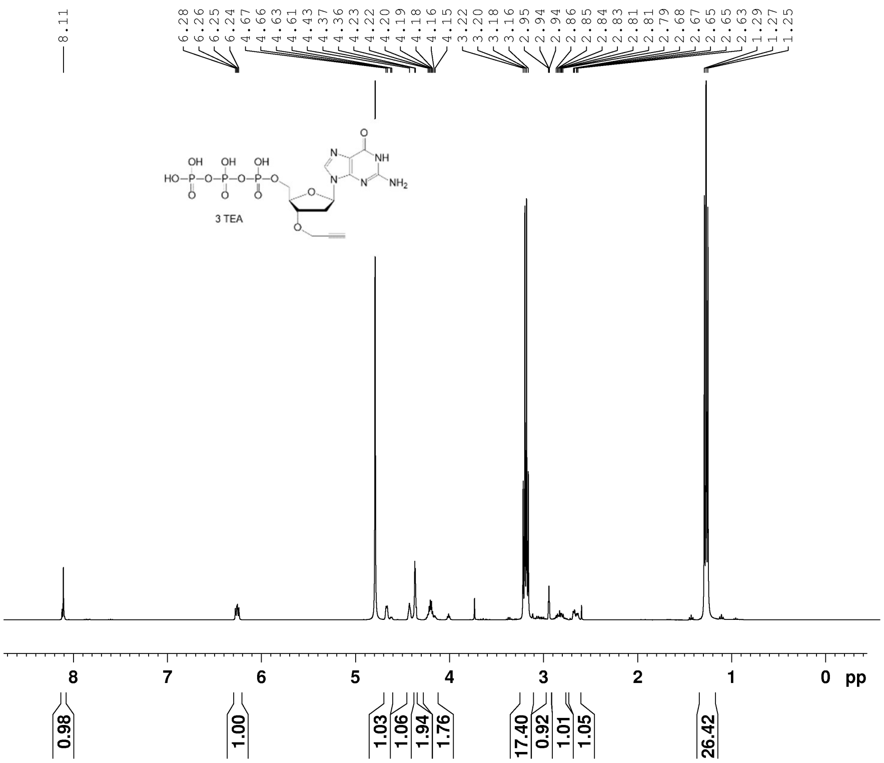
^{1}H NMR of compound 4



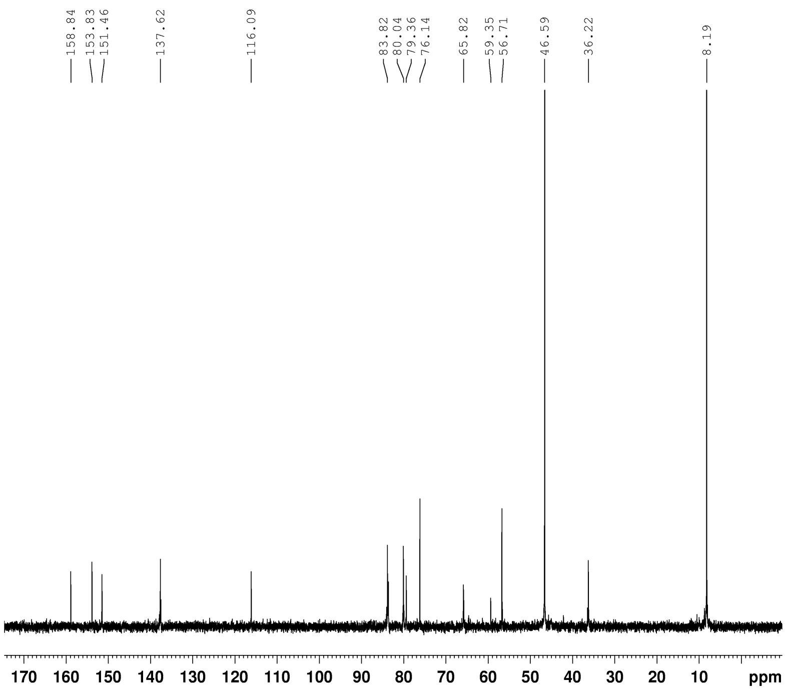
^{13}C NMR of compound 4



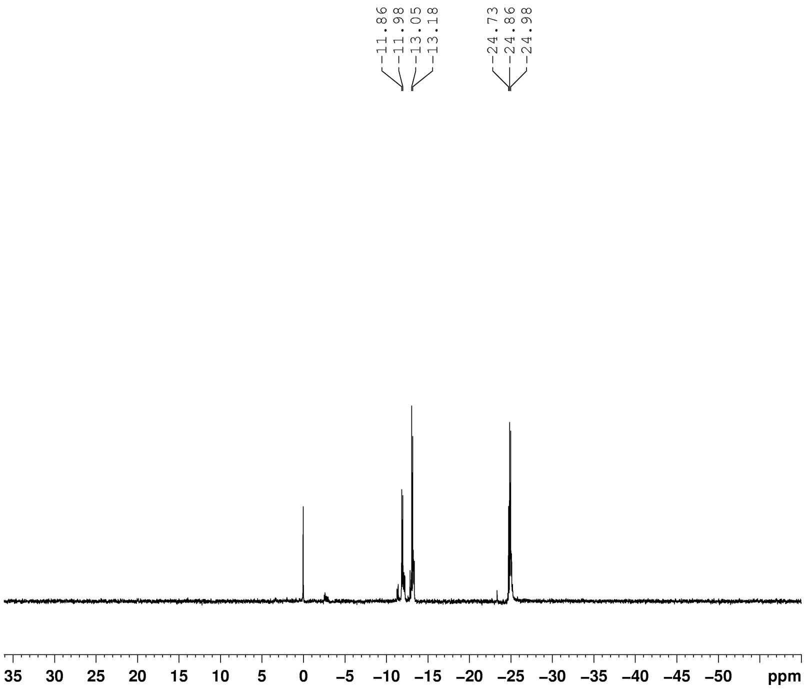
Mass spectrometry of Compound **4**



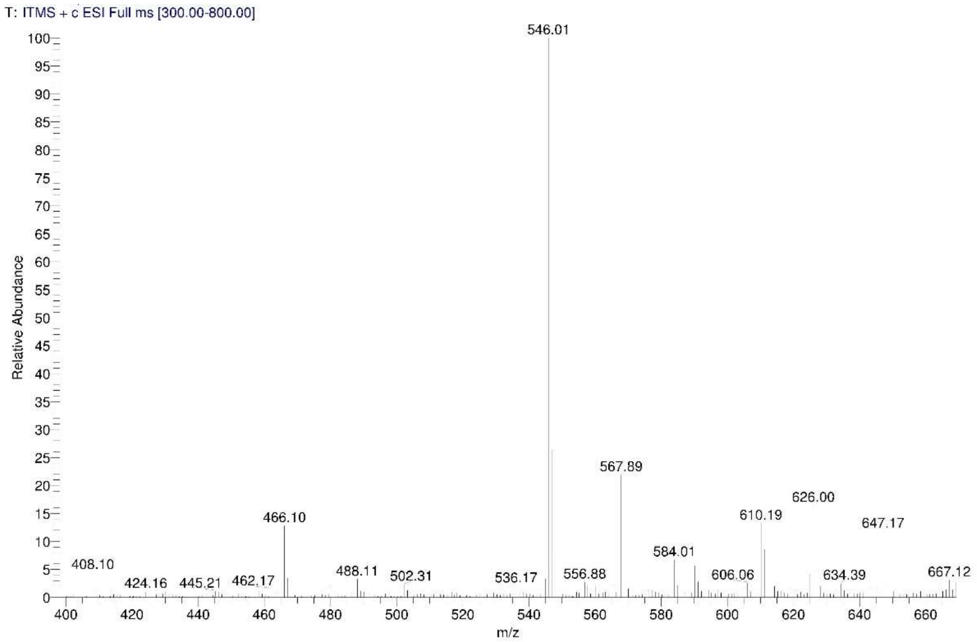
^{1}H NMR of compound **5**



^{13}C NMR of compound **5**



^{31}P NMR of compound **5**



Mass spectrometry of Compound **5**

References

1. Miller, G. P.; Kool, E. T. *J. Org. Chem.* **2004,** *69*, 2404-10.

2. Jiang, C.; Pugh, B. F. *Genome Biol.* **2009,** *10*, R109.

3. Miura, F.; Kawaguchi, N.; Sese, J.; Toyoda, A.; Hattori, M.; Morishita, S.; Ito, T. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103*, 17846-51.

4. Hesselberth, J. R.; Chen, X.; Zhang, Z.; Sabo, P. J.; Sandstrom, R.; Reynolds, A. P.; Thurman, R. E.; Neph, S.; Kuehn, M. S.; Noble, W. S.; Fields, S.; Stamatoyannopoulos, J. A. *Nat. Methods* **2009,** *6*, 283-9.

5. Liu, C. L.; Kaplan, T.; Kim, M.; Buratowski, S.; Schreiber, S. L.; Friedman, N.; Rando, O. J. *PLOS Biol.* **2005,** *3*, e328.

6. Pokholok, D. K.; Harbison, C. T.; Levine, S.; Cole, M.; Hannett, N. M.; Lee, T. I.; Bell, G. W.; Walker, K.; Rolfe, P. A.; Herbolsheimer, E.; Zeitlinger, J.; Lewitter, F.; Gifford, D. K.; Young, R. A. *Cell* **2005,** *122*, 517-27.

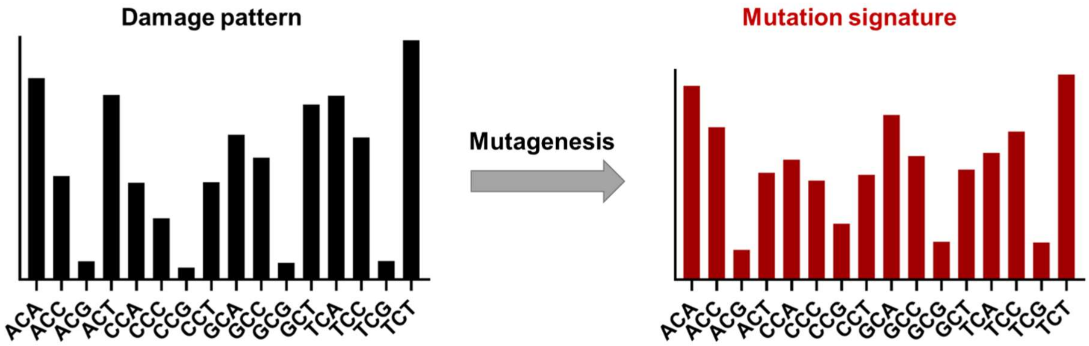
7. Kasinathan, S.; Orsi, G. A.; Zentner, G. E.; Ahmad, K.; Henikoff, S. *Nat. Methods* **2014,** *11*, 203-9.

8. Pan, D.; Zhou, Q.; Rong, S. Z.; Zhang, G. T.; Zhang, Y. N.; Liu, F. H.; Li, M. J.; Chang, D.; Pan, H. Z. *Microchim Acta* **2016,** *183*, 361-368.

9. Fan, J. H.; Liu, Y. J.; Xu, E. S.; Zhang, Y. J.; Wei, W.; Yin, L. H.; Pu, Y. P.; Liu, S. Q. *Anal Chim Acta* **2016,** *946*, 48-55.

10. Ma, B.; Jing, M.; Villalta, P. W.; Kapphahn, R. J.; Montezuma, S. R.; Ferrington, D. A.; Stepanov, I. *Scientific reports* **2016,** *6*.

Chapter 3: Nucleotide-resolution mapping of DNA oxidation in relation to mutagenesis and chromatin state



Chapter 3

Nucleotide-resolution mapping of DNA oxidation in relation to mutagenesis and chromatin state

Abstract

DNA oxidation arising from cellular oxidative stress is a major endogenous damage type with significant toxicological implications in disease development. We recently developed click-code-seq, a nucleotide-resolution genome-wide sequencing method for oxidative damage, however it has only been applied to sequence background oxidation in a yeast genome. To further investigate the distribution of 8-oxoG in human genome, herein, we report the genome-wide distribution of DNA oxidation in human haploid cells. The results demonstrate that the distribution of oxidative DNA damage varies widely across the genome, with distinct patterns related to chromatin architecture, epigenetic modification, DNA damage response and DNA-protein interactions. The patterns of nucleobases flanking damage sites closely match oxidative stress-associated mutation signatures, suggesting a potentially predictive relationship between damage distribution and mutation signatures.

Introduction

The faithful transmission and interpretation of genetic information which rely on the integrity and stability of DNA chemical structure are essential to life. However, DNA is continuously subjected to assault from damaging agents that could alter nucleobase structures. DNA oxidation is one of the most common processes that can arise from endogenous metabolism as well as from exogenous chemical exposure, leading to more than 100 types of lesions.[1-2] Among oxidative DNA lesions, 8-oxo-7,8-dihydro-2'-deoxyguanosine (8-oxoG) is the most abundant and well-characterized base modification with up to 2500 8-oxoG sites per healthy human cell.[3] Efficient search and removal of 8-oxoG to maintain cell integrity are performed by base excision repair (BER) pathway, involving 8-oxoguanine glycosylase.[2] Unrepaired 8-oxoG is highly mutagenic due to 8-oxoG : A mismatch during DNA replication, causing G:C to T:A transversion. The accumulation of 8-oxoG and G>T mutation in the genome have been correlated with oxidative stress associated processes and diseases, including cancer, atherosclerosis, diabetes, accelerated aging, and central nervous system pathologies.[4-5]

Over the last decade, it has become increasingly evident that the location of 8-oxoG, disease progression and gene expression are closely integrated. First, it is known that cancer is driven by natural selection enabled by the evolution of mutations conferring a growth advantage. These mutations often occur within the genome in a characteristic pattern, known as mutational signatures. The excess G:C > T:A transversion mutations in colorectal cancer (CRC) in MUTYH-Associated Polyposis (MAP) syndrome exhibit a novel mutational signature with a strong sequence dependence, termed Signature 36. This mutational signature reflecting persistent 8-oxoG:A mismatches occurs frequently in oncogenes that are associated with CRC. [6] Second, 8-oxoG may regulate gene activation in response to oxidative stress by several potential pathways, including direct interactions of OGG1 with transcription factors (TFs), allosteric transition of G-quadruplex (G4) and signal transduction by OGG1·8-oxoguanine complex.[7] Finally, DNA damage response (DDR) signaling network plays a vital role in cell cycle checkpoints, DNA repair, and DNA-damage tolerance pathways. Changes in histone post-translational modifications have been observed as a part of the DDR network upon DNA damage, leading changes in chromatin environment, damage repair and gene expression.[8] Given the extremely high biological relevance, improved knowledge is needed to understand factors that influence the persistence of DNA oxidation in the genome and their relation to mutagenesis and gene expression. However, strategies for tracking DNA oxidation in the genome are lagging behind because of their inherent chemical complexity.

Recently, several methods were reported map 8-oxoG via next-generation sequencing. Using a selective 8-oxoG chemical labeling method (OG-seq), genome-wide distribution of 8-oxoG in mouse embryonic fibroblasts (MEFs) showed that 8-oxoG occurs at a greater frequency in specific genomic elements.[9] Alternatively, 8-oxoG containing DNA fragment was fished out with an OG-selective antibody for high throughput sequencing with 0.1kb resolution (OxiDIP-Seq). The results revealed the accumulation and co-localization of 8-oxoG sites and γH2AX ChIP-seq signals at transcribed regions in MEFs and MCF10A cells, particularly at long genes and at DNA replication origins.[10]

To achieve single nucleotide resolution sequencing of 8oxoG, we employed a glycosylase excision and click-reaction based sequencing method, named click-code-seq.[11] The principle of click-code-seq mainly involves three steps: First, the 8-oxoG site is recognised and removed by the 8-oxoG glycosylase Fpg, generating a gap with free 3'-hydroxyl at the damage site. Then, a synthetic *O*-3'-propargyl modified nucleotide (prop-dGTP) is incorporated into the resulting gap by Therminator IX DNA polymerase, giving rise to a 3'-alkynyl modified end. After that, the yield 3'-alkynyl DNA is ligated to a 5'-azido-modified code sequence via a copper(I)-catalyzed click reaction, resulting a triazole-linked DNA that could be amplified by DNA polymerases. Via this process, 8-oxoG sites are stably labelled with a code sequence that serves as a tag for affinity enrichment, an adaptor for PCR amplification and a marker of the damage locations. By performing whole genome mapping of DNA oxidation in *S. cerevisiae* genome, initial steps have been taken to gain a better understanding of DNA oxidation in a genome scale.

To further understand the genome-wide distribution of DNA oxidation, we performed single-nucleotide resolution mapping of DNA oxidation in the human genome. Click-code-seq was optimized for the mammalian genome to fit higher throughput sequencing platform. The sequencing results were compared to several existing HAP1 sequencing datasets, providing insights into the genomic distribution of DNA oxidation as a composite of local sequence context, histone modification, DNA-protein interactions and gene size.

Results

DNA oxidation mapping of HAP1 cells by click-code-seq

In principle, click-code-seq is adaptable to genomic DNA from any species, [11] however, a major technical limitation for species with larger genome sizes, i.e. such as human, is that higher cluster density during sequencing is used to achieve higher throughput. However, the final sequencing library of click-code-seq begins with a 6-mer code sequence, thus creating low initial sequence diversity and leading to large scale loss of data.[12] To deal with the low diversity issue, a mixture of a four code sequence was used for the click reaction, thus increasing initial sequence diversity (Table S1). Using this strategy, we successfully obtained high-quality data from click-code-seq library with Illumina Hiseq 2500 platform.

Human haploid cells (HAP1) derived from male chronic myelogenous leukemia cell line KBM-7 was used as biological model in this study. HAP1 contains a single copy of the genome, so gene editing on one allele for loss of function mutations is not masked by additional alleles, and sequencing scale is substantially reduced and more affordable. Both wilde type (WT) HAP1 cells and 8-oxoguanine DNA glycosylase (OGG1) deficient HAP1 cells (Ogg1^{-}) were used, as OGG1 is the BER glycosylase responsible for removing oxidized guanines from the genome.

Genomic DNA was extracted (~5 µg) from HAP1 cells and fragmented by sonication to obtain a population of strands with an average length of 300 bp. The click-code-seq protocol was conducted on the fragmented samples. The sequencing library was submitted for NGS using the Illumina HiSeq 2500 instrument (125 cycle single end). Around 30 million reads were obtained from each sample and aligned to the human reference genome GRCh38.p13.

We analyzed the data by assessing damage distribution relative to published genome annotations and HAP1 sequencing datasets (See supporting information). The 8-oxoG spectrum for individual nuclear chromosomes indicated the widespread and uneven distribution of 8-oxoG in the genomes of HAP1 cell (Figure S2). The quantity of 8-oxoG for each chromosome was plotted and compared (Figure 1a). We observed no major 8-oxoG bias with respect to the strand orientation (5'>3'/ 3'>5') nor with respect to the chromosome size. We then analysed the genomic distribution of 8-oxodG at genome features (Figure 1b). 65.2% of the 8-oxoG sites were mapped within gene loci, including 7.4% in promoter (i.e. upstream and 5'UTR) and 52.4% in gene body. 30.9% of the 8-oxoG sites were mapped within intergenic regions. These results suggest a broad distribution of 8-oxoG throughout the whole genome.

To further examine the resulting damage pattern in more detail, we analyzed two specific regions of chromosome 1, i.e. 152.2 Mb to 152.3 Mb (region I) and 161.4 Mb

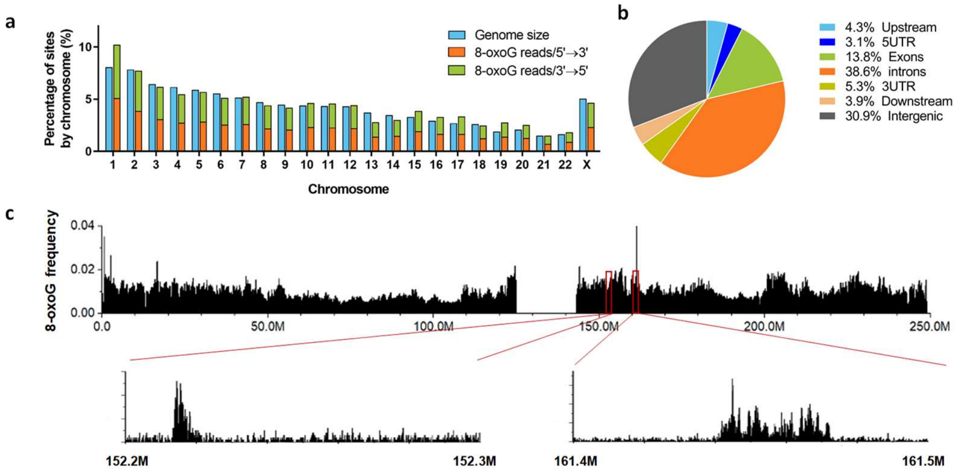


Figure 1: (a). Chromosomal distribution of 8-oxoG compared to the size of each nuclear chromosome. The total amount of genome size or 8-oxoG reads is set to 100. A comparison of 8-oxoG reads for 5'>3' and 3'>5' strands is also displayed. (b). Pie chart showing the annotation of 8-oxodG in several different genomic features. (c). Genome browser view of 8-oxoG distribution on Chromosome 1 at 100-kb resolution. More detail views of two regions at 100 bp resolution. The vertical axes shows the 8-oxoG frequency per bp.

to 161.5Mb (region II) (Figure 1c). Most of 8-oxoG reads at region I are located at Homo sapiens hornerin (HRNR) gene body, where also harbors most of SNPs at this region based on NCBI's dbSNP human build 151 data (SNP151) (Figure S2). Region II showed the highest 8-oxoG peak at 100 kb resolution and was zoomed further into 100 bp resolution. An increasing amount of 8-oxoG was found around 161.45 Mb region, where don't contain SNPs hotspots but is a high transcript density region for transfer ribonucleic acid tRNA. All these results suggest that the distribution of 8-oxoG is not a single factor process and more detailed genomic features and functional elements play an important role in shaping the local damage distributions.

8-oxoG damage signature correlates with cancer mutation signatures

We then try to address the potential relationship between damage pattern and cancer mutation signature. We generated 3-base damage patterns in 5'-XCY-3' triplets with C opposite 8-oxoG site in the middle (Figure 2a). Correlations between damage pattern and mutation signatures were assessed using Spearman's correlation (Table S2). Interestingly, 16 mutation signatures showed a strong correlation (Spearman's r > 0.6) to damage pattern in both WT and $Ogg1^-$ HAP1 genome, including mutation signature 5 (SBS5) , SBS9, SBS10a, SBS18, SBS30, SBS32, SBS36 and so on. Most of these mutation signatures are associated with increasing oxidative stress or DNA repair/replication protein deficiency. For example, SBS18 commonly found in neuroblastoma is potentially due to ROS-induced DNA damage; SBS36 and SBS30 are associated with deficient DNA glycosylases, MUTYH and NTHL1, respectively; SBS9 and 10a are associated with DNA replication process, including translesion

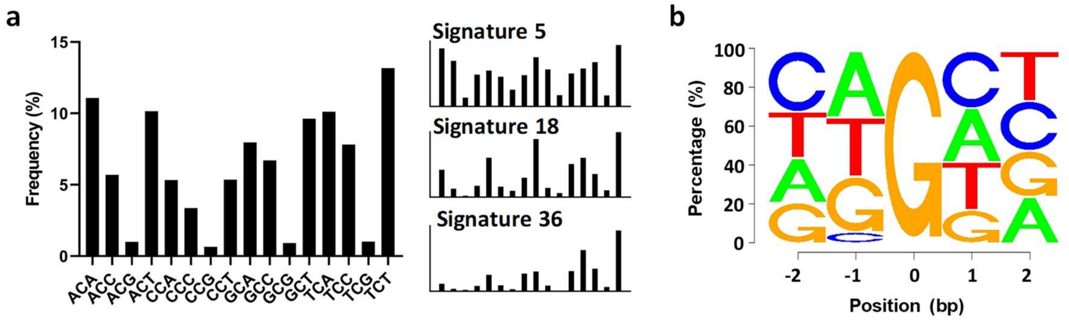


Figure 2. (a) 3-bases damage plot with the complementary sequence of 8-oxoG sites. The sequence contexts are on the horizontal axes, whereas vertical axes show the frequency of the specific sequence context. Right side shows three mutation signature plots regenerated from COSMIC database. (b) Sequence logo plot shows prevalence of bases surrounding 8-oxoG (position 0).

synthesis Pol η and proofreading deficient Pol ε, respectively.[6, 13-14] These results revealed that the mutational process from 8-oxoG may reflect the overloading of 8-oxoG (SBS18), the absence of 8-oxoG repair protein (SBS30, SBS36) and the altered DNA replication process (SBS9, SBS10a). These results provide the first direct evidence that the distribution pattern of 8-oxoG at a local sequence scale may play a vital role to shape the C>T transversions *in vivo*.

Then, we analyzed the sequence context of nucleobases flanking 8-oxoG sites (Figure 2a). The frequency of each base 5' to 8-oxoG was more variable: 35% A, 31% T, 28% G and 5% C than 3' to 8-oxoG (~12% variance) in human genome. This result conflicts with calculated chemical reactivity of middle G in 5'-XGY-3' and our observation in yeast genome.[15-16] However, resistant 8-oxoG in a genome is the balance of damage formation and repair which could be influenced by repair efficiency, chromatin status and DDR. Further studies are needed to reveal the factors that shape the different damage patterns in different species.

Crosstalk between DNA damage and epigenetic histone modification

Our previous 8-oxoG sequencing in yeast genome and several other studies of DNA damage distribution in mammalian genome have showed that histone modifications exhibit significant impact on DNA repair enzyme accessibility.[11, 17-18] DNA damage is repaired faster at sites of histone modifications associated with active chromatin than with repressed and heterochromatic status. Seven histone modifications examined in this study were H3K4me1, H3K4me3 and H3K27Ac, which are associated with active gene transcription; H3K36me3 and H2K119Ub, which are associated with heterochromatic chromatin; H3K27me3, which are associated with repressed chromatin. A ChIP-seq dataset of HAP1 from NCBI (GSE 107599) was used to analyze the 8-oxoG distribution. We used the IgG ChIP-seq data as a control for our analysis. No correlation was found between 8-oxoG reads and IgG enrichment regions. For H3K27me3, H3K36me3 and H2K119Ub, regions with these histone modifications harbor more 8-oxoG reads compared to regions without, showing a positive correlation. Meanwhile, a negative correlation was found between H3K4me1 and 8-oxoG reads. However, for H3K4me3 and H3K27ac, 8-oxoG was accumulated at low-to-middle level of these two modifications and reduced again at high level. On the basic of histone modification distribution, H3K4me1 is more located at active enhancers, H3K4me3 is more located at actively transcribed promoters; H3K27ac is located at both. Interestingly, the relationship between H3K27ac and 8-oxoG is very similar to the average of H3K4me1 and H3K4me3, indicating that the distribution of 8-oxoG at actively transcribed promoters shapes the patterns of H3K4me3 and H3K27ac (Figure S4). In general, these observations are relatively consistent with our previous knowledge that chromatin status has major influence on DNA repair. However, the abnormal observations on H3K4me3 lead us to pay attention to the impact of DNA damage on histone modification.

DDR is a chromatin-associated process that regulates histone modifications to modulate DNA damage repair. SSBs formed during the excision of 8-oxodG can be converted into DSBs during DNA replication. 8-oxoG and DSB biomarkers γ-H2AX are co-localized in the genome.[10] Histone-lysine *N*-methyltransferase (Set1) was reported to accumulate at newly created DSBs in budding yeast cell, increasing the level of H3K4me3 as chromatin remodeler.[19] However, the role of H3K4me3 is disputable in DDR. H3K4me3 could also be down-regulated at DNA damage sites by UV laser microirradiation. The decrease of H3K4me3 at damage sites is required for efficient recruitment of ATP-dependent DNA helicase 2 subunit KU70 (Ku70) and breast cancer type 1 susceptibility protein (BRCA1) for DSB repair.[20] The two-sided role of H3K4me3 may explain the accumulation of H3K4me3 at low-to-middle level of 8-oxoG and decrease of H3K4me3 at high level of 8-oxoG.

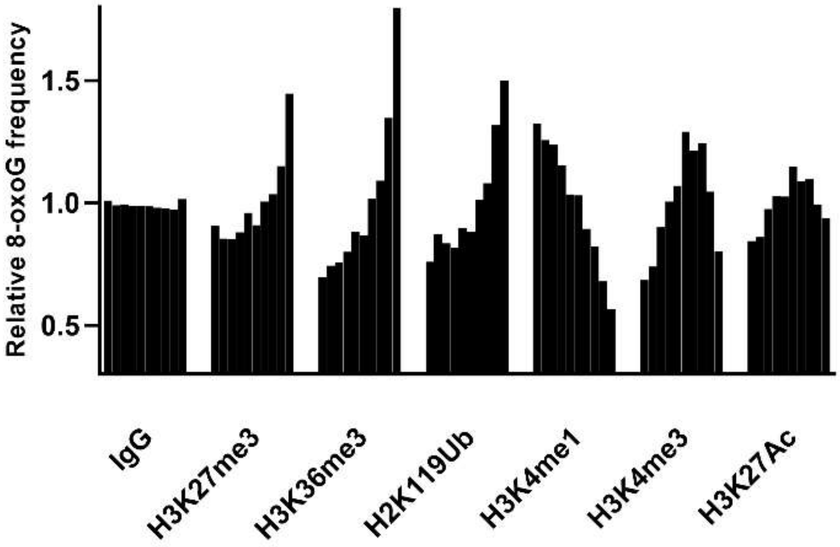


Figure 3: Column plots showing the relative 8-oxoG frequencies within increasing level of histone modifications.

8-oxoG levels are related with gene length but not expression

Increased transcriptions of specific genes when cells response to oxidative stress have been reported. The mechanism behind is not well understood. Several studies suggested that the 8-oxoG at promoter region participate in regulating gene transcription. To further examine the relationship between 8-oxoG and gene expression, 8-oxoG frequencies at each promoter region and gene body were analyzed and compared to HAP1 transcription dataset (GSE107600). No correlation between gene expression and 8-oxoG frequency neither in gene body nor in promoter was found (Figure 4a, 4b). Very recently, two 8-oxoG sequencing studies, one in adipose and lung tissues using OG-seq and another in MEFs and MCF10A cells using OxiDIP-seq also showed gene expression is not associated with 8-oxoG frequency at promoter region.[10, 21] Meanwhile, we found a negative correlation between 8-oxoG frequency and gene length (Figure 4c). In summary, these data show that 8-oxodG doesn't participate in oxidative stress induced gene transcription and is preferentially repaired within long genes.

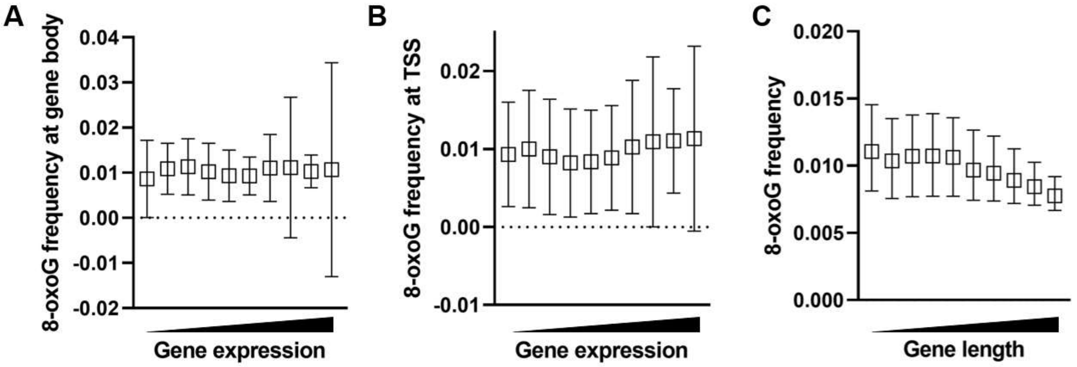


Figure 4. 8-oxoG frequency per bp at gene body (A) and promoter (B) across the increasing gene expression level. (C) 8-oxoG frequency per bp at gene body across the increasing gene length.

Conclusion

Despite extensive chemistry and biology knowledge of 8-oxoG, the role of chemical and biological factors that shape the distribution of 8-oxoG in the human genome is largely unknown. Thus, we applied our previously reported click-code-seq method for single nucleotide resolution mapping of 8oxoG to human HAP1 cell to investigate 8-oxoG distribution in human genome. A key advantage of click-code-seq over other methods is the ability to map 8-oxoG at single nucleotide resolution which is significant to understand how the distribution of 8-oxoG related with mutagenesis, epigenetic modifications, genomic features and DNA-protein interactions.

Mutagenesis is a multi-step process including damage preference, repair efficiency and replication accuracy. The reveal of mutation signature from cancer cell line upon specific chemical exposure or repair protein mutant suggests the important role of DNA damage during cancer development. However, due to the lack of nucleotide-resolution damage sequencing data, the linkage between mutation signature and DNA damage pattern in genome is still ambiguous. Here, for the first time, we observed an 8-oxoG damage pattern in human genome which is similar to several mutation signatures resulted from oxidative stress, DNA repair deficiency or tranlesion synthesis. This finding suggests that the process from 8-oxoG to mutation is mainly dominated by resistant 8-oxoG in genome, thus, leading to a critical question that what factors shaped this specific 8-oxoG pattern in genome. Oxidation of guanine in dsDNA shows a modest preference for the 5'G in a 5'-GG-3' context, due to a lower ionization potential attributed to π- π stacking.[15-16] Meanwhile, the excision of DNA damage by repair proteins is influenced by their neighbor nucleotides.[22-23] The resistant 8-oxoG in genome represents the balance of damage formation, as well as DNA damage repair. Further studies are needed to clarify the contribution of each factor in damage pattern formation.

Besides local sequence context, distribution of 8-oxoG could be influenced more globally upon the accessibility changes of genomic DNA, including chromatin configuration, epigenetic modifications and DNA-protein interactions. We observed 8-oxoG accumulation in heterochromatin regions (H3K27me3, H3K36me3 and H2K119Ub) and decrease amount of 8-oxoG in euchromatin regions (H3K4me1). This observation agrees SNVs distribution in genome, which is positively correlated with heterochromatin marks and negatively correlated with euchromatin marks.[24] However, two euchromatin marks, H3K4me3 and H3K27ac, accumulate at low-to-middle level of 8-oxoG and decrease at high level of 8-oxoG. This observation may related with the two-sided role of epigenetic mark in DNA damage response. Our observations suggest that distribution of 8-oxoG is not only related with DNA accessibility but also related with DNA damage response signaling network.

Furthermore, we observed that the amount of 8-oxoG at gene promoter or gene body have no impact on gene expression regulation. However, either oxidative-stress induced transcription or G-quadruplex oxidation involved gene activation is restricted to a small group of genes.[9, 25] This specific phenomenon may be buried in global genes. Meanwhile, we also observed a decrease of 8-oxoG frequency in long genes. Transcription of very long genes has been shown to cause accumulation of R loops and in turn lead to DNA damage and genomic instability.[26] Indeed, we observed increased 8-oxoG reads in long genes. However, the relative 8-oxoG frequency in long gene is decreased, suggesting a potential protection mechanism to long genes.

References

1. Cadet, J.; Davies, K. J. A., Oxidative DNA damage & repair: An introduction. *Free Radical Bio. Med.* **2017,** *107*, 2-12.
2. Markkanen, E., Not breathing is not an option: How to deal with oxidative DNA damage. *DNA Repair* **2017,** *59*, 82-105.
3. Tubbs, A.; Nussenzweig, A., Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **2017,** *168* (4).
4. Bosshard, M.; Markkanen, E.; van Loon, B., Base excision repair in physiology and pathology of the central nervous system. *Int. J. Mol. Sci.* **2012,** *13* (12), 16172-222.
5. Nakabeppu, Y., Cellular levels of 8-oxoguanine in either DNA or the nucleotide pool play pivotal roles in carcinogenesis and survival of cancer cells. *Int. J. Mol. Sci.* **2014,** *15* (7), 12543-57.
6. Viel, A.; Bruselles, A.; Meccia, E.; Fornasarig, M.; Quaia, M.; Canzonieri, V.; Policicchio, E.; Urso, E. D.; Agostini, M.; Genuardi, M.; Lucci-Cordisco, E.; Venesio, T.; Martayan, A.; Diodoro, M. G.; Sanchez-Mete, L.; Stigliano, V.; Mazzei, F.; Grasso, F.; Giuliani, A.; Baiocchi, M.; Maestro, R.; Giannini, G.; Tartaglia, M.; Alexandrov, L. B.; Bignami, M., A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **2017,** *20*, 39-49.
7. Wu, J. Z.; Sturla, S. J.; Burrows, C. J.; Fleming, A. M., Impact of DNA Oxidation on Toxicology: From Quantification to Genomics. *Chem. Res. Toxicol.* **2019,** *32* (3), 345-347.
8. Chen, Y. C.; Zhu, W. G., Biological function and regulation of histone and non-histone lysine methylation in response to DNA damage. *Acta Bioch. Bioph. Sin.* **2016,** *48* (7), 603-616.
9. Fleming, A. M.; Ding, Y.; Burrows, C. J., Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (10), 2604-2609.
10. Amente, S.; Di Palo, G.; Scala, G.; Castrignano, T.; Gorini, F.; Cocozza, S.; Moresano, A.; Pucci, P.; Ma, B.; Stepanov, I.; Lania, L.; Pelicci, P. G.; Dellino, G. I.; Majello, B., Genome-wide mapping of 8-oxo-7,8-dihydro-2-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res.* **2019,** *47* (1), 221-236.
11. Wu, J. Z.; McKeague, M.; Sturla, S. J., Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J. Am. Chem. Soc.* **2018,** *140* (31), 9783-9787.
12. Mitra, A.; Skrzypczak, M.; Ginalski, K.; Rowicka, M., Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform. *Plos One* **2015,** *10* (4).
13. Alexandrov, L. B.; Jones, P. H.; Wedge, D. C.; Sale, J. E.; Campbell, P. J.; Nik-Zainal, S.; Stratton, M. R., Clock-like mutational processes in human somatic cells. *Nat. Genet.* **2015,** *47* (12), 1402.
14. Petljak, M.; Alexandrov, L. B.; Brammeld, J. S.; Price, S.; Wedge, D. C.; Grossmann, S.; Dawson, K. J.; Ju, Y. S.; Iorio, F.; Tubio, J. M. C.; Koh, C. C.; Georgakopoulos-Soares, I.; Rodriguez-Martin, B.; Otlu, B.; O'Meara, S.; Butler, A. P.; Menzies, A.; Bhosle, S. G.; Raine, K.; Jones, D. R.; Teague, J. W.; Beal, K.; Latimer, C.; O'Neill, L.; Zamora, J.; Anderson, E.; Patel, N.; Maddison, M.; Ng, B. L.; Graham, J.; Garnett, M. J.; McDermott, U.; Nik-Zainal, S.; Campbell, P. J.; Stratton, M. R., Characterizing Mutational

Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **2019,** *176* (6), 1282-+.

15. Margolin, Y.; Shafirovich, V.; Geacintov, N. E.; DeMott, M. S.; Dedon, P. C., DNA sequence context as a determinant of the quantity and chemistry of guanine oxidation produced by hydroxyl radicals and one-electron oxidants. *Journal of Biological Chemistry* **2008,** *283* (51), 35569-78.
16. Senthilkumar, K.; Grozema, F. C.; Guerra, C. F.; Bickelhaupt, F. M.; Siebbeles, L. D., Mapping the sites for selective oxidation of guanines in DNA. *Journal of the American Chemical Society* **2003,** *125* (45), 13658-9.
17. Zhang, T.; Cooper, S.; Brockdorff, N., The interplay of histone modifications - writers that read. *EMBO Rep* **2015,** *16* (11), 1467-81.
18. Weiner, A.; Hsieh, T. H.; Appleboim, A.; Chen, H. V.; Rahat, A.; Amit, I.; Rando, O. J.; Friedman, N., High-resolution chromatin dynamics during a yeast stress response. *Mol Cell* **2015,** *58* (2), 371-86.
19. Faucher, D.; Wellinger, R. J., Methylated H3K4, a Transcription-Associated Histone Modification, Is Involved in the DNA Damage Response Pathway. *Plos Genet.* **2010,** *6* (8).
20. Mosammaparast, N.; Kim, H.; Laurent, B.; Zhao, Y.; Lim, H. J.; Majid, M. C.; Dango, S.; Luo, Y. Y.; Hempel, K.; Sowa, M. E.; Gygi, S. P.; Steen, H.; Harper, J. W.; Yankner, B.; Shi, Y., The histone demethylase LSD1/KDM1A promotes the DNA damage response. *J. Cell Biol.* **2013,** *203* (3), 457-470.
21. Park, J. W.; Han, Y. I.; Kim, T. M.; Yeom, S. C.; Kang, J.; Park, J., 8-OxoG in GC-rich Sp1 binding sites enhances gene transcription during adipose tissue development in juvenile mice. *bioRxiv* **2019**, 538967.
22. Allgayer, J.; Kitsera, N.; von der Lippen, C.; Epe, B.; Khobta, A., Modulation of base excision repair of 8-oxoguanine by the nucleotide sequence. *Nucleic Acids Res.* **2013,** *41* (18), 8559-8571.
23. Xia, B.; Han, D. L.; Lu, X. Y.; Sun, Z. Z.; Zhou, A. K.; Yin, Q. Z.; Zeng, H.; Liu, M. H.; Jiang, X.; Xie, W.; He, C.; Yi, C. Q., Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **2015,** *12* (11), 1047-1050.
24. Schuster-Bockler, B.; Lehner, B., Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **2012,** *488* (7412), 504-+.
25. Pan, L.; Zhu, B.; Hao, W. J.; Zeng, X. L.; Vlahopoulos, S. A.; Hazra, T. K.; Hegde, M. L.; Radak, Z.; Bacsi, A.; Brasier, A. R.; Ba, X. Q.; Boldogh, I., Oxidized Guanine Base Lesions Function in 8-Oxoguanine DNA Glycosylase-1-mediated Epigenetic Regulation of Nuclear Factor B-driven Gene Expression. *J. Bio. Chem.* **2016,** *291* (49), 25553-25566.
26. Helmrich, A.; Ballarino, M.; Tora, L., Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol. Cell.* **2011,** *44* (6), 966-977.

Supporting information

Cell culture and genomic DNA extraction

HAP1 wild type and OGG1-knock out cells were cultured in 10 cm dishes in 10 ml IMDM medium including 10% FCS and 1% Pen/Strep. At around 70% confluence, the cells need to be split as they become diploid at high confluence. For CellTiter Glo cell viability assay, 12’000 cells were seeded to 96-well plates and let recover for 24 h. For treatment, $KBrO_3$ was dissolved in medium, filter-sterilized and serially diluted. The cells were treated with various concentrations up to 16 mM $KBrO_3$ for 24 h. For assessment of cell viability, CellTiter Glo assay was performed according to the manufacturer’s protocol and luminescence was measured on the Tecan plate reader. 9 Mio cells were seeded onto two 15cm dishes per cell line. After 24 h recovery, the cells were incubated in normal IMDM medium or with 1.2 mM $KBrO_3$ dissolved in IMDM medium for 24 h. 15 Mio cells were used to extract the gDNA using QIAamp DNA mini kit according to the manufacturer’s protocol. BHT and deferoxamine were added to the lysis and elution steps to reduce oxidation during the preparation. The eluted gDNA was stored at -20°C and its concentration was measured with Nanodrop.

Library preparation and sequencing

Genomic DNA (5 µg) was sheared in 130 µL Tris-EDTA buffer with a Covaris S220 ultrasonicator (Covaris Inc.) using the following parameters: peak incident power 140 W, cycles/burst 200, duty factor 10% and time 100 s. DNA concentration and distribution were estimated using the Nanodrop 8000 and Agilent 2200 Tapestation with high sensitivity D1000 screen tape.

The fragments were treated with APE 1 (New England Biolabs (NEB), final concentration (fc): 0.2U/µL) and T4 PNK (NEB, fc: 0.2U/µL) in 1 × NEBuffer 2.1 at 37 °C for 0.5 h to remove abasic sites and 3’ phosphoryl group. Dideoxynucleotides (Jena Bioscience GmbH, fc: 200 µM) and therminator IX (NEB, fc: 0.04 U/µL) were added and heated (60 °C, 10 min) to block gaps and terminal nucleotides in DNA fragments. The product was purified using the Monarch nucleic acid purification kit (NEB). The removal of 8-oxoG was carried out with blocked genomic DNA, Fpg (NEB, fc: 0.16 U/ µL) and APE 1 (NEB, fc: 0.2 U/µL) in 1 × NEBuffer 2.1 at 37 °C for 1 h. Then, prop-dGTP (fc: 0.2 mM) and therminator IX (NEB, 0.04 U/ µL) was added and the mixture was heated (60°C, 10 min).The resulting DNA was purified using the Monarch nucleic acid purification kit, completely dried on a vacuum concentrator and re-suspended in 2 µL water. Next, 3 µL Bar-N3-mix (200 µM), 1 µL potassium phosphate buffer (1 M, pH 7.0), 1 µL aminoguanidine hydrochloride solution (50 mM), 1 µL DMSO, 1 µL sodium ascorbate (25 mM) and 1 µL premixed CuSO4:THPTA (1:6, 5 mM in concentration of Cu2+) were add and incubated at room temperature (30 min).

After purification, DNA (50 µL in TE buffer) was subjected to slow rotation with Dynabeads MyOne C1 (Invitrogen Corp.) in bead-binding buffer (50 µL, 5.0 mM Tris-HCl (pH 7.5), 1.0 mM EDTA, 1 M NaCl) at room temperature for 30 min. Afterwards, the beads were washed twice with bead-washing buffer (200 µL, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.2 M NaCl) and resuspended in 49 µL reaction mixture (5 µL T4 PNK reaction buffer (10 ×), 5 µL ATP (10 mM) and 39 µL ddH2O). After adding 1 µL T4 PNK (NEB), the mixture was warmed at 37 °C for 30 min. Next, the beads were washed two times with 200 µL bead-washing buffer and resuspended in 49 µL reaction mixture (5 µL T4 ligase reaction buffer (10 ×), 5 µL PEG-4000 (50 %), 0.5 µL TWEEN-20 (1 %), 3 µL pre-annealed dsDNA adapter (40 µM)). After adding 1 µL T4 ligase (NEB), the mixture was incubated at 16 °C for 2 h. After adaptor ligation, the beads were washed twice with 200 µL bead-washing buffer, resuspended in 20 µL ddH2O and heated at 95 °C for 2 min. The beads were pelleted immediately using a magnetic rack, and the supernatant, which contained the single-stranded library molecules, was transferred to a fresh tube.

The entire library was pre-amplified with Ex-bar-mix and Pr-P7-23nt using Vent exo-polymerase (NEB) in 1 × Thermopol buffer (NEB) for 5 cycles, involving denaturation for 20 s at 95 °C, annealing for 20 s at 64 °C and primer extension for 60 s at 72 °C, final extension for 5 min at 72 °C, and hold at 4 °C. Use of vent exo- polymerase ensured efficient bypass of the triazole produced by click labelling. Finally, the library was indexed and amplified with one of the P7 barcoded primers, and P5-universal using Q5 high fidelity DNA polymerase (NEB) and purified with an SPRIselect kit. The library size was verified using the Agilent 2200 tapestation with high sensitivity D1000 screen tape before being loaded onto the Illumina Hiseq2500 platform (Illumina) with a single-end 125 bp fast run mode.

Sequence alignment and data analysis

Raw reads were aligned to newest human reference genome (GRCh38.p13, National Center for Biotechnology Information (NCBI)) with Bowtie 2. Reads in this study had a five-base code sequence incorporated during click labelling, corresponding to the first five cycles of raw FASTQ sequence. To remove these sequences, --trim5 5 option was used during bowtie 2 alignment with other default settings (bowtie2 --end-to-end --threads 1 -x reference_genome -U input.fastq --trim5 8 -S output.sam). The produced SAM files were sorted and converted to BAM files with Sambamba (sambamba view -S input.sam -f bam -l 0 -o /dev/stdout -t 1|sambamba sort /dev/stdin -o /dev/stdout t 1 -l 0 -m 20GB --tmpdir=TMPDIR > output.bam). Then, PCR duplicates in BAM files were removed with samtools using rmdup command (samtools rmdup input.bam output.bam). Next, bam files were converted to bed files directly (bedtools bamtobed -i input.bam > output.bed). With the bedtools genomecov command, the depth at each genome position with 100kb or 100bp resolution was reported in bed files, which were used to generate the genome wide map of 8-oxoG (genomeCoverageBed -d -i input.bed -g reference_genome > output.bed). Based on the strand tag (+/-), 8-oxoG sites were calculated using the following rules: for read with '+' tag, the 8-oxoG was located at the start site of this read; for reads with '–' tag, the 8-oxoG was located at the end site of this read. The bed files with single resolution data were generated with the above rules, and used for further analysis.

The 8-oxoG frequencies present in each genomic feature were calculated with bedtools intersect command (bedtools intersect -a input.bed -b genomic_feature.bed > output.bed). Genomic features data were downloaded from UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) or Ensembl BioMart (https://www.ensembl.org/biomart). HAP1 specific sequencing datasets were download from NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) and listed in External data. Any genome annotation that was not generated based on GRch38 assembly was converted to correct assembly using UCSC Liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Sequence logo analysis was carried out with bedtools (bedtools getfasta -fi reference_genome -bed inpout.bed -s -fo output.bed) and Weblogo (https://github.com/WebLogo) (seqlogo -f input.bed -k 1 -o output.eps). All of the data described as "relative 8-oxoG frequency" in the manuscript were calculated based on following equation: Relative 8-oxoG frequency = (8-oxoG counts in genome feature/genome feature size(bp))/(8-oxoG counts in whole genome/whole genome size(bp)). Graphs were made using GraphPad Prism 7.03 (GraphPad Software), Origin 9.1 (OriginLab Corp.) or Rstudio 1. All error bars in this study represent standard deviation.

External data

RNA-seq in HAP1 cell:

GSE107600 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107600)

ATF4 binding sites in HAP1 cell:

GSE69304 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69304)

ChIP-seq of histone modifications in HAP1 cell:

GSE107599 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107599)

ChIP-seq of histone modifications and DNA binding proteins in HAP1 cell:

GSE108387 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108387)

Mutation signatures: https://cancer.sanger.ac.uk/cosmic/signatures

NCBI's dbSNP human build 151 data:

https://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi?view+summary=view+summary&build_id=151

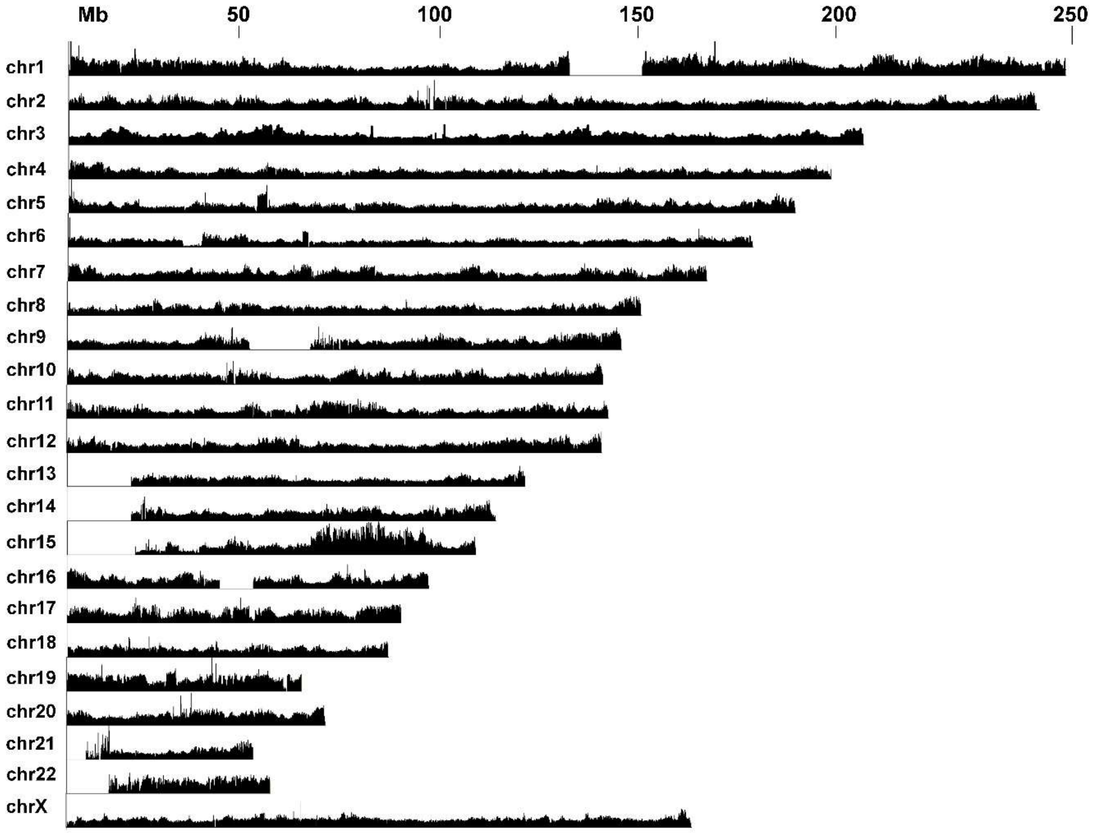


Figure S1: The click-code-seq map of 8-oxoG in genomic DNA from HAP1 cells. The peaks of 8-oxoG reads are shown for the individual nuclear chromosomes (chr 1 – 22 and chrX). The height of each peak corresponds to the number of reads per bp in 100-kb resolution.

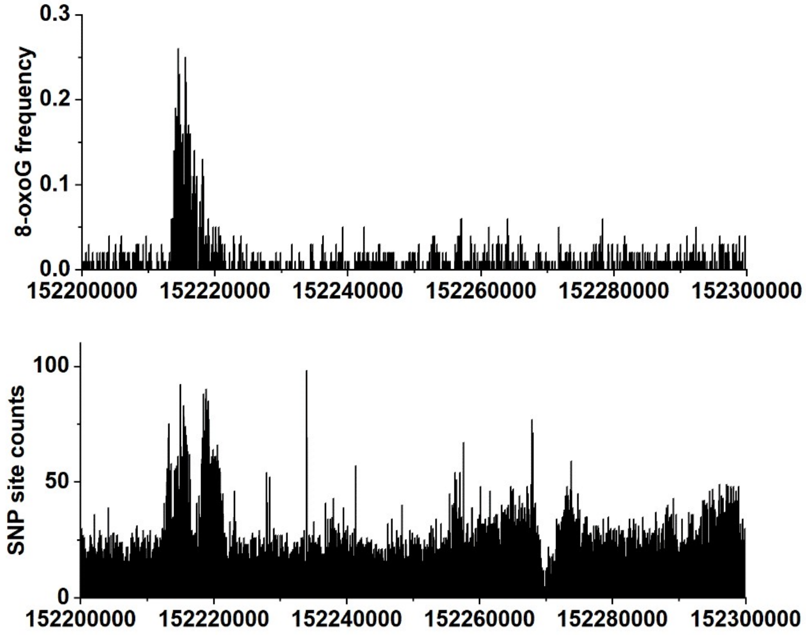


Figure S2: (A) Genome browser views of 8-oxoG levels in 100 bp resolution from 152.2 Mb to 152.3 Mb. (B) Genome browser views of SNP levels in 100 bp resolution from 152.2 Mb to 152.3 Mb.

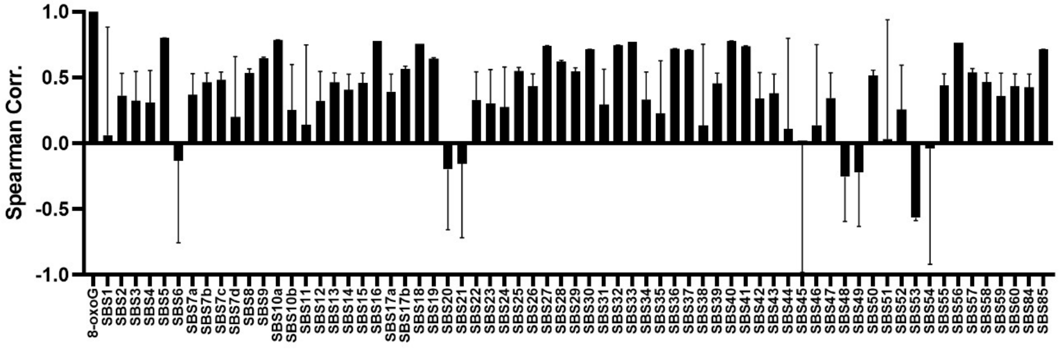


Figure S3: Spearman correlation analysis between 8-oxoG damage pattern and mutation signatures. Bar plots depict the Spearman's r. Error bars depict the p-values from spearman correlation.

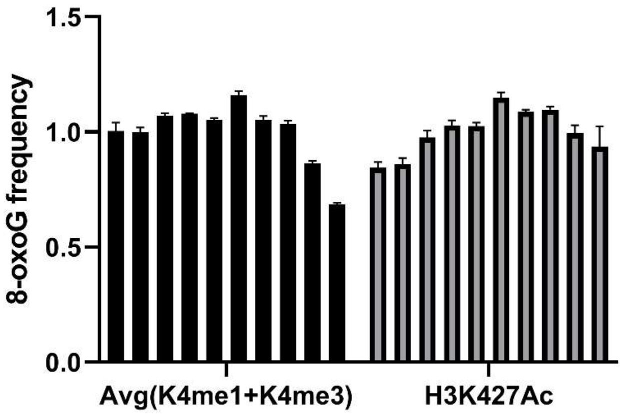
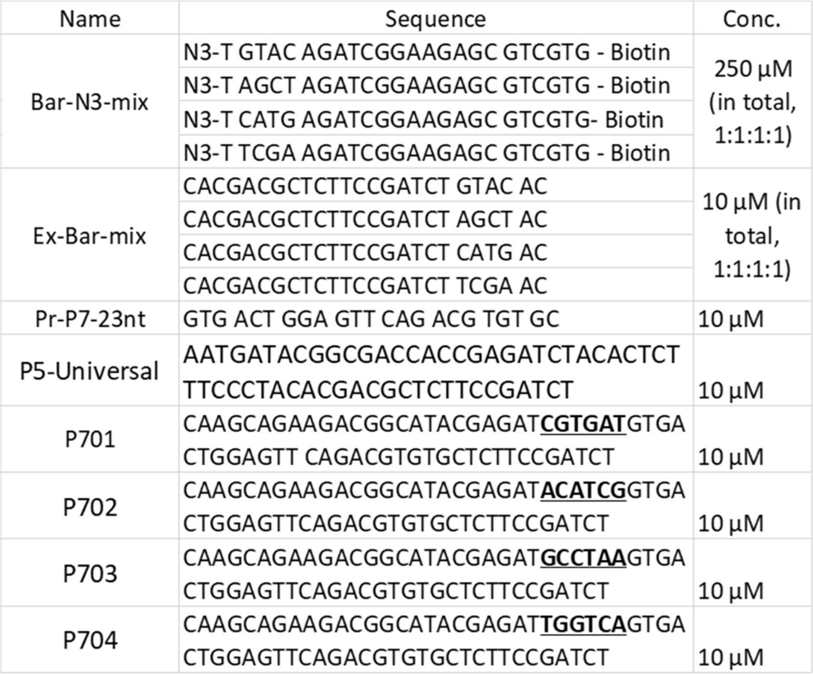


Figure S4: Column plots showing the relative 8-oxoG frequencies within increasing level of H3K27Ac and thhe average of K4me1 and K4me3.



Name	Sequence	Conc.
Bar-N3-mix	N3-T GTAC AGATCGGAAGAGC GTCGTG - Biotin	250 µM (in total, 1:1:1:1)
	N3-T AGCT AGATCGGAAGAGC GTCGTG - Biotin	
	N3-T CATG AGATCGGAAGAGC GTCGTG- Biotin	
	N3-T TCGA AGATCGGAAGAGC GTCGTG - Biotin	
Ex-Bar-mix	CACGACGCTCTTCCGATCT GTAC AC	10 µM (in total, 1:1:1:1)
	CACGACGCTCTTCCGATCT AGCT AC	
	CACGACGCTCTTCCGATCT CATG AC	
	CACGACGCTCTTCCGATCT TCGA AC	
Pr-P7-23nt	GTG ACT GGA GTT CAG ACG TGT GC	10 µM
P5-Universal	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	10 µM
P701	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTT CAGACGTGTGCTCTTCCGATCT	10 µM
P702	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10 µM
P703	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10 µM
P704	CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10 µM

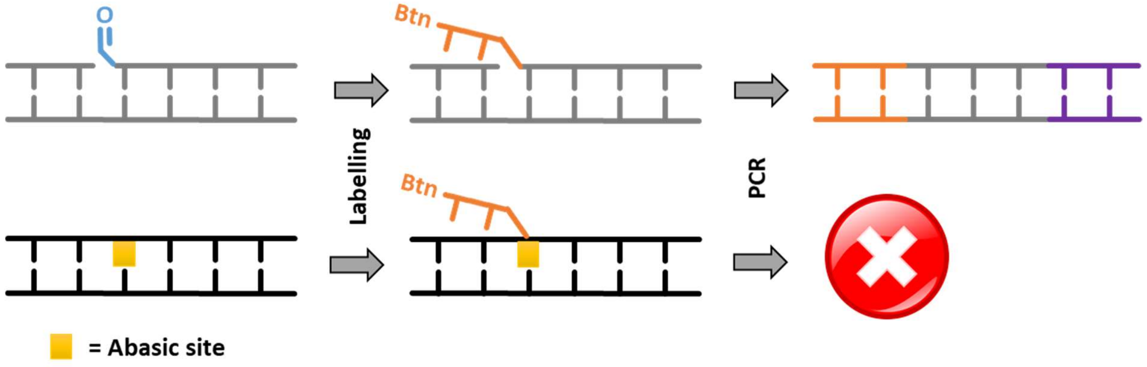
Table S1: Oligonucleotides used in this study

	WT	Ogg1^{-}	Proposed aetiology
SBS5	0.8	0.76765	Unclear, ERCC2 mutations, tobacco smoking
SBS10a	0.7853	0.80882	Polymerase epsilon exonuclease domain mutations
SBS16	0.7794	0.79706	Unknown
SBS40	0.777	0.79912	Unknown
SBS33	0.7706	0.77941	Unknown
SBS56	0.7623	0.75055	Possible sequencing artefact
SBS18	0.7559	0.79412	Possibly damage by reactive oxygen species.
SBS32	0.7441	0.67941	treatment with azathioprine
SBS27	0.7412	0.74118	Possible sequencing artefact.
SBS41	0.7382	0.77059	Unknown
SBS36	0.7177	0.75588	MUTYH mutations
SBS30	0.7118	0.72353	mutations in NTHL1
SBS85	0.7118	0.71176	activation-induced cytidine deaminase (AID)
SBS37	0.7088	0.73529	Unknown
SBS9	0.6471	0.64706	induced during replication by polymerase eta
SBS19	0.6431	0.63429	Unknown
SBS28	0.6206	0.59706	Unknown
SBS17b	0.5637	0.57248	Unknown
SBS25	0.5471	0.58235	chemotherapy treatment
SBS29	0.5445	0.55335	from individuals with a tobacco chewing habit
SBS57	0.5353	0.57647	Possible sequencing artefact
SBS8	0.5324	0.50882	associated with CC>AA mutations
SBS50	0.5147	0.47353	Possible sequencing artefact
SBS7c	0.4827	0.48859	exposure to ultraviolet light
SBS58	0.4651	0.44444	Possible sequencing artefact
SBS7b	0.4647	0.48824	ultraviolet light
SBS13	0.4647	0.45294	activity of cytidine deaminase
SBS15	0.4588	0.52059	Defective DNA mismatch repair
SBS39	0.4559	0.45588	Unknown
SBS55	0.4412	0.42353	chemotherapy treatment with platinum drugs
SBS26	0.4324	0.48529	Defective DNA mismatch repair
SBS60	0.4324	0.40882	Possible sequencing artefact
SBS84	0.4238	0.40177	Activity of activation-induced cytidine deaminase
SBS14	0.4059	0.43235	polymerase epsilon mutation and defective DNA mismatch repair
SBS17a	0.3929	0.37822	Unknown
SBS43	0.3794	0.35588	Possible sequencing artefact
SBS7a	0.3694	0.39294	exposure to ultraviolet light
SBS2	0.362	0.37675	activity of cytidine deaminases
SBS59	0.3559	0.34706	Possible sequencing artefact
SBS47	0.3441	0.31176	Possible sequencing artefact
SBS42	0.3382	0.33529	exposure to haloalkanes
SBS34	0.3324	0.3	Unknown

SBS22	0.3294	0.36471	Aristolochic acid exposure
SBS3	0.3235	0.32941	Defective homologous recombination
SBS12	0.3206	0.34412	Unknown
SBS4	0.3088	0.35	Associated with tobacco smoking
SBS23	0.3	0.29706	Unknown
SBS31	0.2941	0.29118	Prior chemotherapy treatment with platinum drugs.
SBS24	0.2735	0.25882	exposures to aflatoxin
SBS52	0.2559	0.3	Possible sequencing artefact
SBS10b	0.2529	0.23529	Polymerase epsilon exonuclease domain mutations
SBS35	0.2265	0.22647	Prior chemotherapy treatment with platinum drugs
SBS7d	0.1987	0.20162	exposure to ultraviolet light
SBS11	0.1383	0.12656	treatment with the alkylating agent temozolomide
SBS46	0.137	0.16642	Possible sequencing artefact
SBS38	0.1353	0.18529	ultraviolet light associated melanomas
SBS44	0.1089	0.13539	DNA mismatch repair deficiency
SBS1	0.062	0.00295	deamination of 5-methylcytosine
SBS51	0.0324	0.09412	Possible sequencing artefact
SBS45	-0.0088	0.06471	Possible sequencing artefact
SBS54	-0.0412	-0.0324	Possible sequencing artefact
SBS6	-0.1324	-0.0971	DNA mismatch repair deficiency
SBS21	-0.156	-0.1751	DNA mismatch repair deficiency
SBS20	-0.2	-0.1471	DNA mismatch repair deficiency
SBS49	-0.2206	-0.1647	Possible sequencing artefact
SBS48	-0.2559	-0.2441	Possible sequencing artefact
SBS53	-0.5677	-0.5235	Possible sequencing artefact

Table S2: Spearman correlation analysis between 8-oxoG damage pattern and mutation signatures.

Chapter 4: Amplification and sequencing of 5’-aldehyde lesions resulting from DNA oxidation



Chapter 4

Amplification and sequencing of 5’-aldehyde lesions resulting from DNA oxidation

Abstract

DNA single strand breaks are the most abundant DNA oxidative damage resulting from reactive oxygen species. Persistent SSBs could lead to replication folk collapse and double strand break formation upon interaction with the replisome, which may further lead to genome rearrangements, cell death and disease. There are numerous chemical forms of SSBs including 5’ aldehyde terminus, which involves several different repair enzymes. Despite advances in the understanding of SSB repair, the genome-wide landscape of SSBs remains largely unknown because of the lack of methods for sequencing this damage with high specificity and resolution. Therefore, we developed a new strategy to locate the 5’-aldehyde terminus at single nucleotide resolution. The principle of the method involves labelling the 5’-aldehyde terminus with an aminooxy-functionalized oligonucleotide, giving rise to a biocompatible altered DNA linkage and allowing labelled sites to be amplified by the polymerase chain reaction. The specificity of labelling and polymerase bypass of other aldehyde modifications, such as abasic site were characterized, supporting that various sites are labelled but only those derived from 5’-aldehyde precursors could be bypassed and amplified by a polymerase. The results of this work provide a new strategy for studies aiming to provide valuable knowledge on the biological and toxicological impacts of SSB in a genome scale.

Introduction

As an unavoidable consequence of metabolism and stimulation by chemical exposures, DNA oxidation resulting from reactive oxygen species (ROS) constitutes a major threat to genetic integrity and has thus been implicated in a wide variety of oxidative stress-associated diseases.[1] In the past years, oxidative nucleobases such as 8-oxoguanine (8-oxoG) have drawn the most attention due to the mutagenesis and carcinogenesis aspects of altered nucleobase structures.[2] However, there is growing evidence that oxidation of deoxyribose in DNA also induces a serious threat to genetic stability and cell survival by evoking single- and double-strand DNA break (SSB and DSB) and complex protein-DNA cross-links.[3-4] Among the five positions in 2-deoxyribose, C5’ hydrogen atoms in B-DNA are the most accessible to groove binding molecules and diffusible species.[5] Therefore, C5’ hydrogens are believed to be the most frequently abstracted by hydroxyl radicals, resulting in 5'-oxidation.[6-7] A 5'-aldehyde terminus (5'-AT) is then formed concomitantly with strand scission as one of the major and the best-characterized 5'-oxidation product with half-life of around one week.[6, 8] As a subtype of single strand breaks, 5'-AT is repaired through SSB repair pathway (SSBR), involving poly(ADP-ribose) polymerase 1 (PARP1) for damage recognition, DNA polymerase β and flap endonuclease 1 for damage excision.[4, 9] The importance of 5'-AT together with other types of SSBs is highlighted by the observation that involving repair proteins are mutated in cancer and promising targets for anticancer therapy such as PARP inhibitor.[10-13] Unfortunately, the biological consequences of unrepaired 5'-AT and other SSBs are not well understood due to the lack of high sensitive detection method.

SSB with 3'-hydroxyl group has been mapped based on nick translation with digoxigenin labelled dUTP, and affinity enrichment with anti-digoxigenin antibody.[14-15] However, SSBs with 3'-hydroxyl group could be artificially formed during genomic DNA extraction, yielding false positive results. Moreover, there is no detectable marker for SSB sites during sequencing, making it difficult to map the exact position of a SSB. Meanwhile, single-nucleotide-resolution sequencing of several DNA adducts have been achieved by adduct-specific antibodies and repair proteins, such as UV photoproducts[16-18], cisplatin adducts[19], benzo[α]pyrene adducts[20] and 8-oxoG[21-22]. However, these methods are not congruent for 5'-AT sequencing because of the lacking of specific antibody and repair protein. The method used for 5'-AT quantification relies on the derivatization by the biorthogonal reaction between aldehyde group and reactive amine nucleophile.[6] This biorthogonal chemistry could also reacts with other aldehyde-containing nucleotides, such as 5-formyl-2'-deoxycytidine (5fC) and abasic sites.[23-25] Thus, the main challenge of 5'-AT sequencing is how to enrich 5'-AT containing fragments specifically without the interference of other aldehyde-containing nucleotides.

Here we present a novel method that detects 5'-AT specifically. The core idea embedded in this approach is to label 5'-AT with 3'-aminooxy 5'-biotin modified oligodeoxynucleotide (ODN) as a code sequence instead of widely using small molecules. The resulting biocompatible oxime-linked DNA could be read through by DNA polymerases and amplified using the code sequence as down-stream primer binding site during polymerase chain reaction (PCR) amplification. Through this way, only the ligated product yielded from 5'-AT could be amplified and sequenced, and the code sequence could be used as a readable marker to achieve single nucleotide resolution (Figure 1).

Results and Discussion

Synthesis of modified oligonucleotides

In this study, three modified ODNs containing 5'-AT, 3'-aminooxy terminus and oxime internal linkage were synthesized to test our hypothesis for 5'-AT labelling. The synthesis of 5'-aldehyde modified ODN was achieved via a previously reported method with a vicinal diol modified phosphoramidite and standard solid phase synthesis (supporting information).[26] The aldehyde was produced via $NaIO_4$ treatment of freshly deprotected and purified ODN (Figure 2a). For 3'-aminooxy modified ODN, phthalimide was first used as the aminooxy protection group as previously reported.[27-28] All attempts to purify ODN with deprotected aminooxy group failed, in all likelihood leading to a formaldehyde adduct supported by ESI-MS (SI). For this reason, a 5'-phosphoramidite with dimethoxytrityl (DMT) protected 3'-aminooxy was synthesized and used for 5' > 3' solid phase synthesis. ODN was purified with the DMT group retained and deprotected freshly every time before use under standard ODN detritylation condition with 80% acetic acid / 20% water (Figure 2a). The synthesis of oxime modified ODN was achieved with a novel on-support method instead of dimer phosphoramidite. In detail, a fully protected 5'-OH ODN on a solid support is mildly

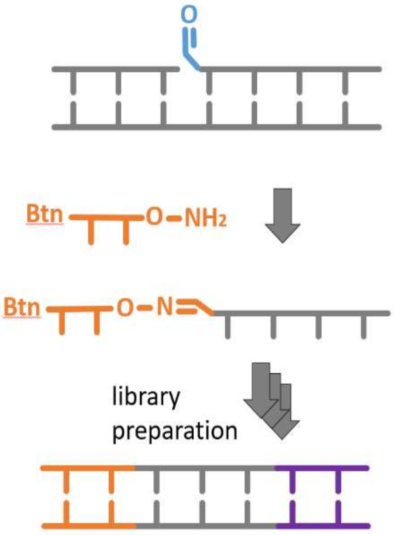


Figure 1: Schematic of a new strategy to detect 5'-AT. DNA containing these sites are biotin-tagged using an 3' aminooxy-functionalized oligonucleotide, pulled down with streptavidin, and amplified by PCR. The damage site is marked with code sequence which coud be read out during sequencing.

oxidized through Moffat reaction.[29] The resulting 5'-aldehyde ODN is then reacted with 3'-aminooxy modified nucleoside analogues to form oxime linkage. Full-length ODN is further synthesized and purified with standard protocol (Figure S1).

Labelling of aldehyde modified oligonucleotide with the code sequence

3'-Aminooxy, 5'-aldehyde and abasic site modified ODNs were used to test the labelling reaction. Abasic site modified ODN was produced by treating a uracil containing ODN with uracil-DNA glycosylase (UDG) and confirmed by human apurinic/apyrimidinic endonuclease (APE1) treatment. (Figure S2) To catalyze the reaction for rapid oxime conjugations, mildly acidic condition (pH 5.4) and large excessive amount of 3'-aminooxy code sequence were used. In 2 hours, the 5'-aldehyde labelling reaction was quantitative completely compared to 7% labelling yield from abasic site modified ODN which simutaneouslyformed 74% β-elimination product under acidic condition (Figure 2b, Figure S2). The results demonstrated that 5'-aldehyde terminus in DNA could be efficiently labelled by a 3'-aminooxy code sequence under a mild acidic condition.

Bypass and amplification of ligated products raised from aldehyde labelling

The specificity of this method lies on the successful amplification of labelled product raised from 5'-aldehyde terminus but not from other aldehyde-modified nucleotides by

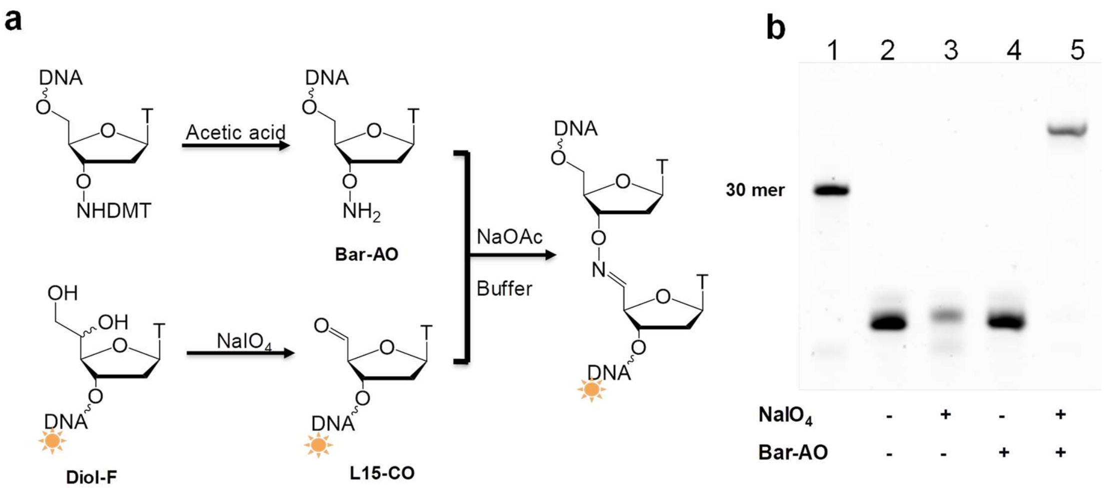


Figure 2: Labelling of aldehyde modified oligonucleotide with code sequence. (a) Structures of 3' aminooxy-functionalized ODN, 5'-AT containing ODN and their conjugated product. (b) Denaturing PAGE analysis of labelling reaction. Only fluorophore labelled 5'-AT containing ODN is visible. Lane 1 shows a 30 mer ODN marker. Slower migrating product in Lane 5 is attributed to conjugated oxime linkage ODN.

DNA polymerases. Two ODNs containing the ligation products raised from 5’-aldehyde terminus and abasic site were used to investigate DNA polymerases bypass of oxime linkages. The abasic site labelled product was synthesized by the reaction between abasic site modified ODN and 3’-aminooxy code sequence, later purified by denaturing polyacrylamide gel electrophoresis (PAGE) and characterized by PAGE and ESI-MS. Deep vent, Vent (exo-) and Therminator IX DNA polymerases could bypass the oxime linkage raised from 5’-aldehyde labelling with more than 50% efficiency compared to Taq and Q5 polymerases with less than 10% efficiency (Figure 3). As for abasic site labelled product, only Vent (exo-) DNA polymerase could bypass with less than 10% efficiency (Figure S3). These results indicate that 5’-AT labelled product is easier to be bypassed than abasic site labelled product as expected.

It is possible however, that PCR amplification of the oxime linked DNA could be efficient even if bypass of the oxime linkage is invisible on a gel. The ability of DNA polymerases to replicate through the oxime linkage was therefore evaluated more rigorously by PCR amplification of ligated plasmid using the code sequence as reverse primer binding site (Figure S4). Taq, Vent (exo-), Deep vent and Q5 DNA polymerase could amplify 5’-AT labelled product successfully (Figure 4). The amplicon was purified and sequenced using Sanger sequencing. Insertion of the code sequence immediately after the known 5’-aldehyde terminus was confirmed (Figure 4b). As for abasic site labelled product, Taq, Deep vent and Q5 DNA polymerase did not give target product during PCR amplification. Vent (exo-) DNA polymerase gave the expected amplicon from abasic site labelled product, agreeing the ODN bypass study on PAGE gel (Figure S3). Considering both the bypass and PCR amplification studies, Taq, Deep vent and

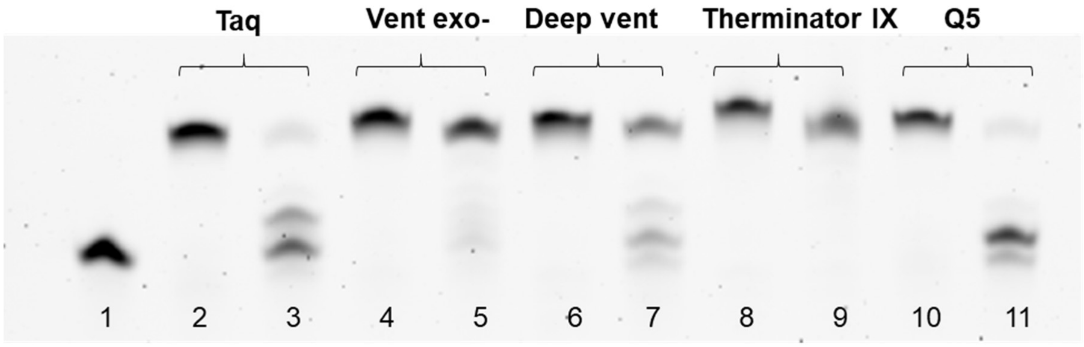


Figure 3: Bypass of DNA template containing a site-specific oxime linkage (lanes 3, 5, 7, 9, 11) or identical template without modification (lanes 2, 4, 6, 8, 10) by different DNA polymerases. Lane 1 indicates a non-template control.

Q5 DNA polymerases showed good specificity to amplify 5’-aldehyde terminus but not abasic site.

qPCR quantification of 5’-aldehyde terminus in dsDNA

To determine whether this method could be used to quantify low amounts of 5’-aldehyde terminus in a DNA sample, various amounts of 5’-aldehyde modified dsDNA were mixed with native dsDNA. The final 5’-aldehyde amount (5’-aldehyde termini/total nucleotides) is ranged from 0 to 0.05% (Figure S5). After labelling reaction with aminooxy modified code sequence, excess code sequence was removed and labelled dsDNA was enriched by streptavidin beads. qPCR quantification studies were performed with either Deep vent or Q5 polymerase using EvaGreen dye (Figure 5a, Figure S6). The specificity of amplification during qPCR was further checked by agarose gel and melting curve analysis (Figure 5b). For both DNA polymerases, Ct values increased linearly as a function of the relative amount of 5’-aldehyde termini, indicating a good enrichment and amplification specificity. Our labelling and pulldown method demonstrated efficient enrichment for 5’-aldehyde containing dsDNA; a single 5’-aldehyde modification enriched the sequence by ~30,000-fold based on qPCR results by Q5 polymerase. As for the lowest amount of input 5’-aldehyde termini (5 x

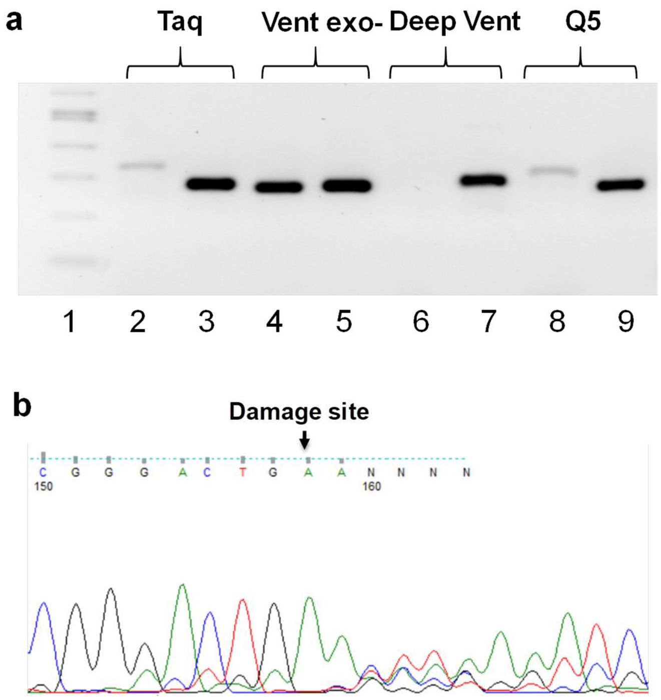


Figure 4: (a) Amplification of labelling products from abasic site (lanes 2, 4, 6, 8) or 5’-AT (Lanes 3, 5, 7, 9) by different DNA polymerases. (b) Sanger sequencing data from 5’-aldyhyde amplicon showing code sequence marked 5’-AT lesion.

10^{-7} damage/total nucleotides), the signal with Q5 DNA polymerase showed 5.3 Ct lower than native dsDNA, indicating 30 times higher in concentration. Meanwhile, Deep vent polymerase was not able to distinguish between the lowest input and native dsDNA based on the Ct values. From a perspective of DNA polymerase fidelity, Q5 as an ultra-high fidelity enzyme shows around 280 times higher fidelity than Taq polymerase and 60 times than Deep vent polymerase. Taking all the ODN bypass, dsDNA amplification, qPCR quantification and polymerase fidelity assays into consideration, Q5 DNA polymerase showed the best specificity to amplify 5'-aldehyde terminus.

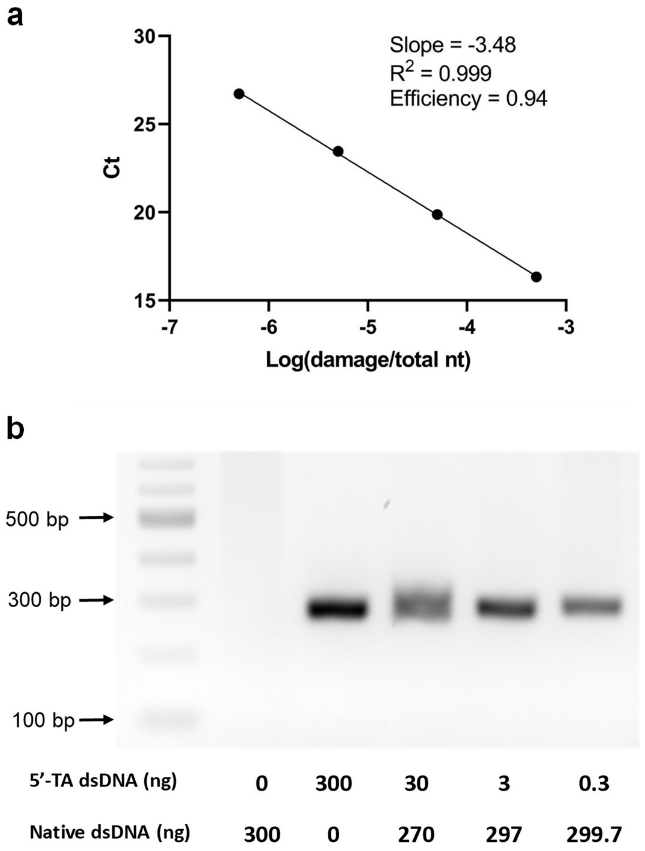


Figure 5: (a) qPCR Ct values using Q5 polymerase as a function of relative 5'-AT concentrations ([5'-AT lesion]/[total DNA]). (b) PCR amplification of DNA samples with different concentrations of initial 5'-AT using Q5 polymerase.

Conclusion

In this study, we designed a new method to label and detect 5'-AT lesion specifically in DNA. We synthesized a aminooxy functioned ODN as a code sequence and found that both 5'-AT and abasic site containing ODNs can react with this code sequence to form oxime conjugation. Several polymerases showed high selectivity to amplify 5'-AT labelled product but not abasic site labelled product. qPCR results showed the method is sufficiently sensitive to label 5'-AT at a frequency as low as 10^{-7} lesions/unmodified bases or several thousand lesions per genome. To the best of our knowledge, this is the first strategy for sensitive and specific detection of 5'-AT sites. We think that this strategy can be further developed as a useful method for quantification and sequencing of 5'-AT lesion in genomic DNA.

References

1. Cooke, M. S.; Evans, M. D.; Dizdaroglu, M.; Lunec, J., Oxidative DNA damage: mechanisms, mutation, and disease. *Faseb J.* **2003,** *17* (10), 1195-1214.
2. Markkanen, E., Not breathing is not an option: How to deal with oxidative DNA damage. *DNA Repair* **2017,** *59*, 82-105.
3. Dedon, P. C., The chemical toxicology of 2-deoxyribose oxidation in DNA. *Chem. Res. Toxicol.* **2008,** *21* (1), 206-219.
4. Caldecott, K. W., Single-strand break repair and genetic disease. *Nat. Rev. Genet.* **2008,** *9* (8), 619-631.
5. Balasubramanian, B.; Pogozelski, W. K.; Tullius, T. D., DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95* (17), 9738-9743.
6. Chan, W.; Chen, B.; Wang, L.; Taghizadeh, K.; Demott, M. S.; Dedon, P. C., Quantification of the 2-Deoxyribonolactone and Nucleoside 5′-Aldehyde Products of 2-Deoxyribose Oxidation in DNA and Cells by Isotope-Dilution Gas Chromatography Mass Spectrometry: Differential Effects of γ-Radiation and Fe2+−EDTA. *J. Am. Chem. Soc.* **2010,** *132* (17), 6145-6153.
7. Boussicault, F.; Kaloudis, P.; Caminal, C.; Mulazzani, Q. G.; Chatgilialoglu, C., The fate of C5 ' radicals of purine nucleosides under oxidative conditions. *J. Am. Chem. Soc.* **2008,** *130* (26), 8377-8385.
8. Rana, A.; Yang, K.; Greenberg, M. M., Reactivity of the Major Product of C5'-Oxidative DNA Damage in Nucleosome Core Particles. *Chembiochem* **2019,** *20* (5), 672-676.
9. Abbotts, R.; Wilson, D. M., 3rd, Coordination of DNA single strand break repair. *Free Radic Biol. Med.* **2017,** *107*, 228-244.
10. Zheng, L.; Jia, J.; Finger, L. D.; Guo, Z. G.; Zer, C.; Shen, B. H., Functional regulation of FEN1 nuclease and its link to cancer. *Nucleic Acids Res.* **2011,** *39* (3), 781-794.
11. Alshammari, A. H.; Shalaby, M. A.; Alanazi, M. S.; Saeed, H. M.; Azzam, N. A., Novel Mutations of the PARP-1 Gene Associated with Colorectal Cancer in the Saudi Population. *Asian Pac. J. Cancer P.* **2014,** *15* (8), 3667-3673.
12. Jayasinghe, R. G.; Cao, S.; Gao, Q. S.; Wendl, M. C.; Vo, N. S.; Reynolds, S. M.; Zhao, Y. Y.; Climente-Gonzalez, H.; Chai, S. J.; Wang, F.; Varghese, R.; Huang, M.; Liang, W. W.; Wyczalkowski, M. A.; Sengupta, S.; Li, Z.; Payne, S. H.; Fenyo, D.; Miner, J. H.; Walter, M. J.; Vincent, B.; Eyras, E.; Chen,

K.; Shmulevich, I.; Chen, F.; Ding, L.; Network, C. G. A. R., Systematic Analysis of Splice-Site-Creating Mutations in Cancer. *Cell Rep.* **2018,** *23* (1), 270.

13. Tangutoori, S.; Baldwin, P.; Sridhar, S., PARP inhibitors: A new era of targeted therapy. *Maturitas* **2015,** *81* (1), 5-9.

14. Baranello, L.; Kouzine, F.; Wojtowicz, D.; Cui, K. R.; Przytycka, T. M.; Zhao, K. J.; Levens, D., DNA Break Mapping Reveals Topoisomerase II Activity Genome-Wide. *Int. J. Mol. Sci.* **2014,** *15* (7), 13111-13122.

15. Baranello, L.; Kouzine, F.; Wojtowicz, D.; Cui, K.; Zhao, K.; Przytycka, T. M.; Capranico, G.; Levens, D., Mapping DNA Breaks by Next-Generation Sequencing. *Methods Mol. Biol.* **2018,** *1672*, 155-166.

16. Hu, J.; Adar, S.; Selby, C. P.; Lieb, J. D.; Sancar, A., Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. *Genes Dev.* **2015,** *29* (9), 948-60.

17. Adar, S.; Hu, J.; Lieb, J. D.; Sancar, A., Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (15), E2124-33.

18. Hu, J.; Adebali, O.; Adar, S.; Sancar, A., Dynamic maps of UV damage formation and repair for the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (26), 6758-6763.

19. Hu, J.; Lieb, J. D.; Sancar, A.; Adar, S., Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (41), 11507-11512.

20. Li, W.; Hu, J.; Adebali, O.; Adar, S.; Yang, Y.; Chiou, Y. Y.; Sancar, A., Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc. Natl. Acad. Sci. U.S.A.* **2017,** *114* (26), 6752-6757.

21. Gorini, F.; Di Palo, G.; Scala, G.; Lania, L.; Cocozza, S.; Amente, S.; Castrignanò, T.; Moresano, A.; Pucci, P.; Ma, B.; Stepanov, I.; Pelicci, P. G.; Dellino, G. I.; Majello, B., Genome-wide mapping of 8-oxo-7,8-dihydro-2′-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res.* **2018,** *47* (1), 221-236.

22. Wu, J. Z.; McKeague, M.; Sturla, S. J., Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J. Am. Chem. Soc.* **2018,** *140* (31), 9783-9787.

23. Poetsch, A. R.; Boulton, S. J.; Luscombe, N. M., Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Bio.* **2018,** *19*.

24. Xia, B.; Han, D.; Lu, X.; Sun, Z.; Zhou, A.; Yin, Q.; Zeng, H.; Liu, M.; Jiang, X.; Xie, W.; He, C.; Yi, C., Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **2015,** *12*, 1047.

25. Condie, A. G.; Yan, Y.; Gerson, S. L.; Wang, Y. M., A Fluorescent Probe to Measure DNA Damage and Repair. *Plos One* **2015,** *10* (8).

26. Kodama, T.; Greenberg, M. M., Preparation and analysis of oligonucleotides containing lesions resulting from C5 '-oxidation. *J. Org. Chem.* **2005,** *70* (24), 9916-9924.

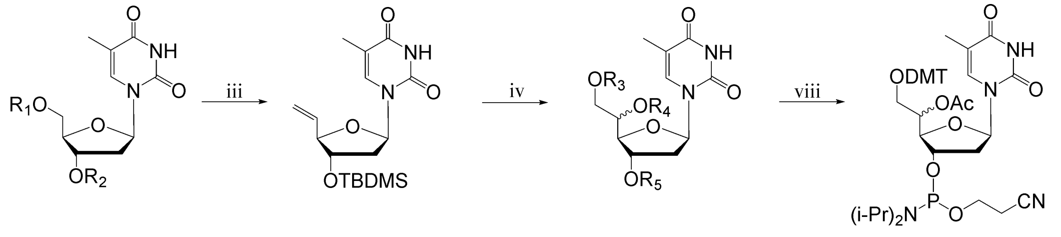
27. Meyer, A.; Vasseur, J. J.; Dumy, P.; Morvan, F., Phthalimide-Oxy Derivatives for 3-or 5-Conjugation of Oligonucleotides by Oxime Ligation and Circularization of DNA by "Bis- or Tris-Click" Oxime Ligation. *Eur. J. Org. Chem.* **2017,** (46), 6931-6941.

28. Salo, H.; Virta, P.; Hakala, H.; Prakash, T. P.; Kawasaki, A. M.; Manoharan, M.; Lonnberg, H., Aminooxy functionalized oligonucleotides: preparation, on-support derivatization, and postsynthetic attachment to polymer support. *Bioconjug. Chem.* **1999,** *10* (5), 815-23.

29. Lartia, R.; Constant, J. F., Synthetic Access to the Chemical Diversity of DNA and RNA 5 '-Aldehyde Lesions. *J. Org. Chem.* **2015,** *80* (2), 705-710.

Supporting information

Chemical Synthesis



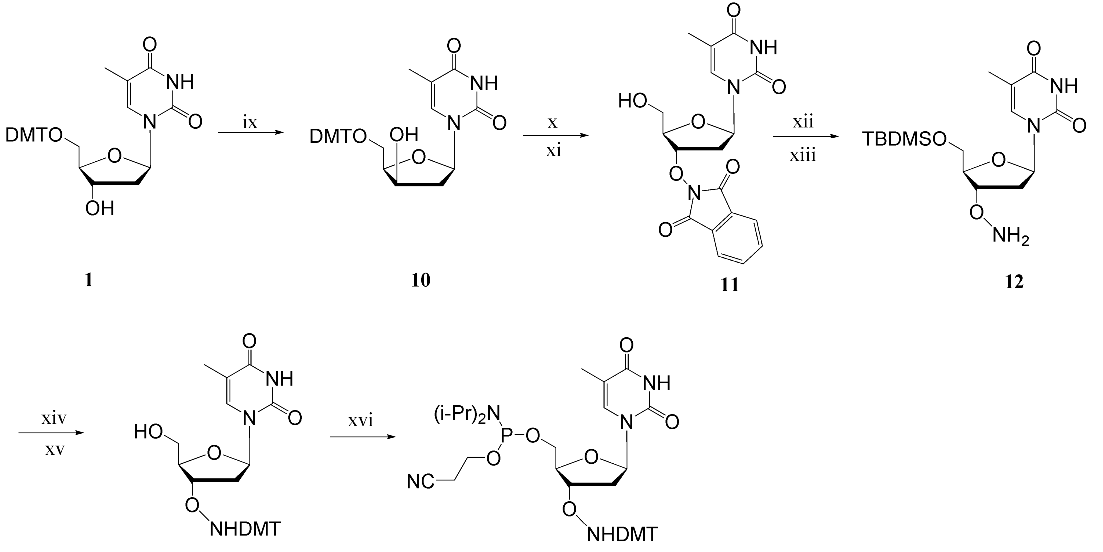
1: R_1 = DMT; R_2 = H (i)
2: R_1 = DMT; R_2 = TBDMS (ii)
3: R_1 = H; R_2 = TBDMS

4

5: R_3 = H; R_4 = H; R_5 = TBDMS (v)
6: R_3 = DMT; R_4 = H; R_5 = TBDMS (vi)
7: R_3 = DMT; R_4 = Ac; R_5 = TBDMS (vii)
8: R_3 = DMT; R_4 = Ac; R_5 = H

9

i) *tert*-Butyldimethylsilyl chloride, Imidazole, Acetonitrile; ii) Trifluoroacetic acid, Dichloromethane; iii) *N'*-ethylcarbodiimide hydrochloride, Dichloroacetic acid, DMSO, Methyltriphenylphosphonium bromide, Potassium tert-butoxide, Tetrahydrofuran; iv) AD-mix-beta, tert-Butyl alcohol, H_2O; v) 4,4'-Dimethoxytrityl chloride, Pyridine; vi) Acetic anhydride, Triethylamine, 4-Dimethylaminopyridine, Acetonitrile; vii) Tetrabutylammonium fluoride trihydrate, Tetrahydrofuran; viii) 2-Cyanoethyl *N,N*-diisopropylchlorophosphoramidite, *N,N*-Diisopropylethylamine, Dichloromethane



1 **10** **11** **12**

13 **14**

ix) Methanesulfonyl chloride, Pyridine; NaOH, H_2O, Ethanol; x) *N*-hydroxyphthalimide, Triphenylphosphine, Diisopropyl azodicarboxylate, THF; xi) Trifluoroacetic acid, Dichloromethane; xii) *tert*-Butyldimethylsilyl chloride, Imidazole, Acetonitrile; xiii) Hydrazine monohydrate, Ethanol; xiv) Tetrabutylammonium fluoride trihydrate, Tetrahydrofuran; xv) 4,4'-Dimethoxytrityl chloride, Pyridine; xvi) 2-Cyanoethyl *N,N*-diisopropylchlorophosphoramidite, *N,N*-Diisopropylethylamine, Dichloromethane

General synthesis information

All chemical reagents were purchased from Sigma Aldrich and used without further purification. All reactions were monitored by TLC using commercial Merck Plates coated with silica gel GF254 (0.24 mm thick). Flash column chromatography was performed on a Biotage SP4 system with pre-packed cartridges. ^{1}H, ^{13}C, and ^{31}P NMR spectra were recorded on a Bruker Biospin 400 MHz NMR instrument at 25 °C. Chemical shifts (δ, ppm) are reported relative to the residual solvent peaks, together with coupling constants (J). The mass spectrometry analysis was measured on Velos Ion Trap Mass spectrometer (Thermo Scientific).

5'-*O*-[bis(4-methoxyphenyl)phenylmethyl]-3'-*O*-[(1,1-dimethylethyl)dimethylsilyl]-thymidine (2)

To a solution of compound **1** (1.63 g, 3 mmol) in acetonitrile (10 mL), *tert*-butyldimethylsilyl chloride (*t*-BDMSCl) (750 mg, 5 mmol) and imidazole (410 mg, 6mmol) were added. The reaction mixture was stirred at ambient temperature for 5 h and then concentrated under vacuum. The resulting mixture was poured into ethyl acetate (100 mL), and washed with brine (3 × 50 mL), and dried over Na_2SO_4. The organic layer was concentrated under reduced pressure, and the resulting residue was used for next step without further purification.

3'-*O*-[(1,1-dimethylethyl)dimethylsilyl]-thymidine (3)

Crude compound **2** was dissolved in 20 mL of dichloromethane and trifluoroacetic acid (0.68 g in 5 mL dichloromethane) added dropwise. The reaction mixture was stirred at ambient temperature for 1 h. Triethylamine was added dropwise to the mixture to quench the excess acid. The crude reaction mixture was evaporated to dryness, diluted with ethylacetate (100 mL) and washed with saturated $NaHCO_3$ (50 mL) and brine (50 mL). The ethylacetate layer was dried over Na_2SO_4. The organic layer was concentrated under vacuum, and the resulting residue was purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (3% - 5 % MeOH) yielding product **3** (0.98 g, 92%) as a white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 9.00 (s, 1H), 7.37 (d, J = 1.5 Hz, 1H), 6.14 (t, J = 6.8 Hz, 1H), 4.48 (dt, J = 6.8, 3.6 Hz, 1H), 3.99 – 3.83 (m, 2H), 3.82 – 3.62 (m, 1H), 2.44 (s, 1H), 2.34 (dt, J = 13.5, 6.8 Hz, 1H), 2.21 (ddd, J = 13.4, 6.5, 3.8 Hz, 1H), 1.90 (d, J = 1.2 Hz, 3H), 0.88 (s, 9H), 0.08 (s, 6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.95, 150.48, 137.19, 111.12, 87.72, 87.05, 71.72, 62.12, 40.60, 25.85, 18.09, 12.64, -4.56, -4.71.

1-[(2R,4S,5R)-4-[[(1,1-dimethylethyl)dimethylsilyl]oxy]-5-ethenyltetrahydro-2-furanyl]-5-methyl-2,4(1H,3H)-pyrimidinedione (4)

To a suspension of methyltriphenylphosphonium bromide (1.43 g, 4 mmol) in dry THF (10 mL) at 0 °C, potassium *tert*-butoxide (0.45 g, 4 mmol) was added. After stirring at 0 °C to ambient temperature for 2 h gave methyltriphenylphosphorane ylide, which was used for the next step synthesis without isolation. In another round bottom flask, to a solution of compound **3** (0.71 g, 2 mmol) and 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride (1.15 g, 6 mmol) in DMSO (5 mL), dichloroacetic acid (0.13 g, 1 mmol) was added and the reaction mixture was stirred for 4 h at ambient temperature. The resulting mixture was poured into ethyl acetate (100 mL), and washed with brine (3 × 50 mL), and dried over Na_2SO_4. The organic layer was concentrated under reduced pressure to get crude aldehyde compound **3** to which above prepared methyltriphenylphosphorane ylide was added at 0 °C slowly. The reaction was stirred at ambient temperature overnight, then reaction was concentrated and purified flash chromatography using a hexane/ethyl acetate gradient (20% - 30% ethyl acetate) yielding product **4** (0.51 g, 73%) as a light yellow solid. ^1H NMR (400 MHz, Chloroform-*d*) δ 8.99 (s, 1H), 7.19 (d, *J* = 1.5 Hz, 1H), 6.23 (t, *J* = 6.4 Hz, 1H), 5.89 (ddd, *J* = 17.1, 10.5, 6.5 Hz, 1H), 5.39 (dt, *J* = 17.2, 1.3 Hz, 1H), 5.29 (dt, *J* = 10.5, 1.3 Hz, 1H), 4.30 – 4.22 (m, 1H), 4.17 (dt, *J* = 6.4, 4.5 Hz, 1H), 2.32 (ddd, *J* = 13.5, 6.4, 4.4 Hz, 1H), 2.10 (dt, *J* = 13.3, 6.5 Hz, 1H), 1.93 (d, *J* = 1.2 Hz, 3H), 0.89 (s, 9H), 0.07 (s, 6H). ^{13}C NMR (101 MHz, Chloroform-*d*) δ 163.83, 150.36, 135.42, 135.27, 118.22, 111.13, 87.64, 85.36, 75.43, 40.71, 25.83, 18.11, 12.79, -4.55, -4.60.

1-[(5ξ)-2-deoxy-3-O-[(1,1-dimethylethyl)dimethylsilyl]-β-D-erythro-hexofuranosyl]-5-methyl-2,4(1H,3H)-pyrimidinedione (5)

To a suspension of AD-mix-β (1 g) in ^tBuOH-H_2O (1:1, 10 mL) at 0 °C, compound **4** (0.49 g, 1.4 mmol) was added. After stirring vigorously at ambient temperature for 24 h. Solid sodium bisulfite (1 g) was added to the reaction mixture. The crude reaction mixture was poured into ethyl acetate (100 mL) and washed with brine (2 x 50 mL). The ethylacetate layer was dried over Na_2SO_4. The organic layer was concentrated under vacuum, and the resulting residue was purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (2% - 4% MeOH) yielding a diastereomeric mixture (ca. 3:1) of **5** (0.51 g, 95%) as a white solid. ^1H NMR (400 MHz, Chloroform-*d*) δ 9.51 (d, *J* = 20.4 Hz, 1H), 7.45 (dd, *J* = 38.2, 1.5 Hz, 1H), 6.20 – 6.03 (m, 1H), 4.57 (ddt, *J* = 9.8, 6.3, 3.0 Hz, 1H), 4.02 – 3.83 (m, 2H), 3.78 – 3.60 (m, 2H), 2.39 (ddd, *J* = 10.8, 7.9, 6.1 Hz, 1H), 2.15 (ddd, *J* = 13.3, 6.3, 3.1 Hz, 1H), 1.89 (d, *J* = 1.1 Hz, 3H), 0.88 (d, *J* = 1.3 Hz, 9H), 0.13 – 0.04 (m, 6H). ^{13}C NMR (101 MHz, Chloroform-*d*) δ 164.23, 150.78, 137.62, 111.24, 88.19, 88.01, 87.52, 87.46, 72.95, 72.42, 71.58, 71.28, 63.76, 40.31, 40.11, 25.87, 25.85, 17.98, 12.59, 12.57, -4.37, -4.51, -4.66, -4.69.

Compound 6

4, 4'-Dimethoxytriphenylmethyl chloride (DMT-Cl) (0.88 g, 2.6 mmol) was added to a solution of compound **5** (0.51 g, 1.3 mmol) in pyridine (5 mL). The mixture was stirred at ambient temperature under an argon atmosphere. After complete consumption of the starting material (6 hours), MeOH (2 mL) was added and the resulting mixture was concentrated under vacuum. The resulting residue was purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (1% - 3 % MeOH) yielding product **6** (0.82 g, 90%) as a light yellow foam. ^{1}H NMR (400 MHz, Chloroform-*d*) δ 8.51 (d, *J* = 8.9 Hz, 1H), 7.66 – 7.51 (m, 1H), 7.51 – 7.40 (m, 2H), 7.39 – 7.18 (m, 8H), 6.86 (dd, *J* = 8.6, 6.1 Hz, 4H), 6.24 (dt, *J* = 25.7, 6.9 Hz, 1H), 4.58 – 4.36 (m, 1H), 4.12 (dq, *J* = 9.1, 3.2 Hz, 1H), 3.82 (d, *J* = 4.4 Hz, 7H), 3.46 – 3.12 (m, 2H), 2.87 (dd, *J* = 14.7, 3.1 Hz, 1H), 2.32 – 2.11 (m, 2H), 1.95 (d, *J* = 6.0 Hz, 3H), 1.70 (s, 1H), 0.88 (d, *J* = 30.6 Hz, 9H), 0.13 – -0.12 (m, 6H). ^{13}C NMR (101 MHz, Chloroform-*d*) δ 163.72, 158.83, 150.41, 144.62, 136.93, 136.47, 135.82, 135.74, 130.20, 130.17, 130.14, 129.29, 128.28, 128.21, 128.10, 128.01, 127.17, 113.43, 113.33, 111.06, 88.11, 87.15, 86.83, 86.34, 85.71, 71.70, 71.61, 70.62, 64.68, 55.37, 55.35, 41.13, 40.65, 25.89, 25.83, 17.92, 12.78, 12.76, -4.47, -4.51, -4.76..

Compound 7

To a solution of compound **6** (0.82 g, 1.2 mmol), triethylamine (420 µL, 303 mg, 3 mmol,) and 4-dimethylaminopyridine (20 mg, 0.16 mmol) in acetonitrile (5.0 mL), acetic anhydride (190 µL, 204 mg, 2 mmol) was added. The reaction mixture was stirred at ambient temperature for 2 h. The resulting mixture was diluted with ethyl acetate (100 mL), washed with saturated $NaHCO_3$ (50 mL), and brine (2 x 50 mL). The ethyl acetate layer was dried over Na_2SO_4. The organic layer was concentrated under vacuum, and the resulting residue was purified by flash chromatography on a Biotage SP4 system using a hexane/ethyl acetate gradient (0% - 30% ethyl acetate) yielding compound **7** (0.84 g, 96%) as a light yellow foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.80 (d, *J* = 5.4 Hz, 1H), 7.70 (td, *J* = 5.7, 4.5, 1.4 Hz, 2H), 7.65 – 7.47 (m, 8H), 7.12 (dd, *J* = 9.0, 2.4 Hz, 4H), 6.51 (ddd, *J* = 31.2, 8.2, 5.6 Hz, 1H), 5.67 – 5.49 (m, 1H), 4.69 – 4.48 (m, 1H), 4.40 – 4.32 (m, 1H), 4.08 (s, 6H), 3.75 – 3.52 (m, 2H), 2.54 (ddd, *J* = 13.3, 5.5, 2.0 Hz, 1H), 2.44 (d, *J* = 7.0 Hz, 3H), 2.18 (dd, *J* = 23.7, 1.2 Hz, 4H), 1.17 (d, *J* = 3.5 Hz, 9H), 0.32 (d, *J* = 16.9 Hz, 6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.23, 163.49, 158.74, 150.10, 144.66, 135.86, 134.87, 130.17, 130.11, 128.22, 127.99, 127.05, 113.32, 111.12, 86.53, 86.05, 85.63, 85.02, 72.69, 72.56, 72.44, 72.21, 62.54, 60.51, 55.35, 40.56, 25.82, 21.28, 21.22, 18.00, 14.33, 12.79, 12.63, -4.48, -4.74.

Compound 8

To a solution of compound **7** (0.81 g, 1.1 mmol) in tetrahydrofuran (5 mL), tetrabutylammonium fluoride (0.5 g, 1.6 mmol) was added. The reaction mixture was stirred at ambient temperature for 4 h. The resulting mixture was diluted with ethyl acetate (50 mL), washed with brine (2 x 50 mL). The ethyl acetate layer was dried over Na_2SO_4. The organic layer was concentrated under vacuum, and the resulting residue was purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (0% - 2% MeOH) yielding compound **8** (0.60 g, 89%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 9.01 (s, 1H), 7.61 (dd, *J* = 7.1, 1.9 Hz, 2H), 7.57 – 7.37 (m, 8H), 7.09 – 6.99 (m, 4H), 6.41 (t, *J* = 6.5 Hz, 1H), 5.49 (d, *J* = 5.9 Hz, 2H), 4.87 (dq, *J* = 7.6, 3.9 Hz, 1H), 4.21 (t, *J* = 3.9 Hz, 1H), 4.04 – 3.94 (m, 6H), 3.72 – 3.39 (m, 2H), 3.07 (dd, *J* = 37.7, 3.4 Hz, 1H), 2.61 (tdd, *J* = 19.9, 8.9, 5.0 Hz, 1H), 2.40 – 2.25 (m, 4H), 2.11 (t, *J* = 2.0 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 170.39, 170.05, 163.70, 163.58, 158.88, 158.86, 158.83, 150.26, 150.23, 144.39, 144.18, 135.57, 135.51, 135.48, 135.35, 135.15, 134.97, 130.08, 130.05, 129.95, 129.28, 128.19, 128.10, 128.03, 127.25, 127.17, 113.51, 113.44, 113.32, 111.27, 111.04, 87.17, 86.95, 86.28, 85.59, 84.89, 84.50, 72.55, 71.99, 71.30, 70.21, 63.06, 62.95, 55.37, 53.53, 40.32, 40.16, 21.22, 21.20, 12.74.

Compound 9

To a solution of compound **8** (62 mg, 0.1 mmol) and *N, N*-diisopropylethylamine (48 µL, 36 mg, 0.3 mmol) in dry DCM (3 mL) at 0 °C under argon atmosphere, 2-cyanoethoxy *N, N*-diisopropylaminochlorophosphine (33 µL, 35 mg, 0.15 mmol) was added. The reaction mixture was stirred at ambient temperature for 2 h. The resulting mixture was diluted with DCM (20 mL), washed with cold $NaHCO_3$ (20 mL). The DCM layer was dried over Na_2SO_4. The organic layer was concentrated under vacuum. The oily residues was kept under high vacuum for 2 h, dissolved in acetonitrile (1 mL) and used for oligonucleotide synthesis without further purification.

1-[5’-*O*-(4,4'-dimethoxytrityl)-2’-deoxy-*β*-D-furanosyl] thymidine (10)

To a solution of 5'-*O*-[bis(4-methoxyphenyl)phenylmethyl]-thymidine (3.8 g, 7 mmol) and *N,N*-diisopropylethylamine (3.7 mL, 2.7 g, 21 mmol) in dry THF (20 mL) at 0 °C under argon atmosphere, mesyl chloride (1.1 mL, 1.6 g, 14 mmol) was added dropwise. The resulting reaction mixture was stirred for 2 h at ambient temperature. Ethanol (50 ml) and NaOH aq (1.0 M, 30 mL) were added and the reaction stirred at 85 ^{o}C overnight. After concentration under vacuum, the resulting mixture was diluted with DCM (200 mL), washed with brine (100 mL), 1.0 M HCl (100 mL), $NaHCO_3$ (100 mL) and brine (100 mL). The DCM layer was dried over Na_2SO_4 and concentrated under vacuum. The resulting residue was purified by flash chromatography using a dichloromethane/methanol (DCM/MeOH) gradient (0% - 4% MeOH) yielding compound **10** (3.2 g, 84%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.70 (d, J = 1.6 Hz, 1H), 7.46 (d, J = 7.6 Hz, 2H), 7.39 – 7.23 (m, 8H), 6.87 (d, J = 8.6 Hz, 4H), 6.22 (dd, J = 8.3, 2.5 Hz, 1H), 4.50 (t, J = 4.1 Hz, 1H), 4.03 (q, J = 4.8 Hz, 1H), 3.82 (s, 6H), 3.59 (ddd, J = 42.5, 10.2, 5.2 Hz, 2H), 3.13 (s, 1H), 2.61 (ddd, J = 14.3, 8.3, 5.4 Hz, 1H), 2.15 (dd, J = 14.9, 2.6 Hz, 1H), 1.82 (d, J = 1.3 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.87, 158.90, 158.88, 150.66, 144.35, 137.21, 135.43, 135.37, 130.06, 130.01, 128.22, 128.03, 127.27, 113.53, 110.55, 87.27, 85.03, 82.46, 71.30, 61.97, 55.39, 40.82, 12.64.

3’-*O*-phthalimido-2’-deoxythymidine (11)

To a solution of compound **10** (1.1 g, 2 mmol), *N*-hydroxyphthalimide (0.49 g, 3 mmol) and triphenylphosphine (0.79 g, 3 mmol) in THF (10 mL) at 0 ^{o}C, diisopropyl azodicarboxylate (0.59 mL, 0.61 g, 3 mmol) was added dropwise. The reaction mixture was allowed to warm slowly to ambient temperature and stirred for 2 h. After the reaction was complete, MeOH (2 mL) was added to quench the reaction. The mixture was concentrated under vacuum, and the residue was dissolved in AcOH (80 %; 30 mL). This solution was stirred for 3 h at ambient temperature, then the mixture was concentrated under vacuum. The resulting residue was purified by flash chromatography using a dichloromethane/methanol (DCM/MeOH) gradient (0% - 6% MeOH) yielding compound **11** (0.53 g, 68%) as a white solid. ^{1}H NMR (400 MHz, DMSO-d6) δ 11.34 (s, 1H), 7.89 (s, 4H), 7.72 (s, 1H), 6.37 (dd, J = 8.8, 5.7 Hz, 1H), 5.18 (t, J = 5.2 Hz, 1H), 4.96 (d, J = 5.5 Hz, 1H), 4.24 (d, J = 4.1 Hz, 1H), 3.62 (t, J = 4.7 Hz, 2H), 2.31 (ddd, J = 14.6, 8.9, 5.9 Hz, 1H), 1.78 (s, 3H). ^{13}C NMR (101 MHz, DMSO-d6) δ 164.17, 164.11, 150.94, 136.36, 135.33, 129.13, 123.85, 101.18, 88.57, 84.14, 83.41, 61.87, 35.88, 22.33, 12.74.

3'-*O*-Amino-5'-*O*-[(1,1-dimethylethyl)dimethylsilyl]-thymidine (12)

To a solution of compound **11** (0.39 g, 1 mmol) in acetonitrile (10 mL), *t*-BDMSCl (0.45 g, 3 mmol) and imidazole (0.41 g, 6 mmol) were added. The reaction mixture was stirred at ambient temperature for 3 h and then concentrated under vacuum. Ethanol (5 mL) and hydrazine hydrate (1 mL) were added and the reaction stirred at ambient temperature for 3 h. The resulting mixture was poured into ethyl acetate (100 mL), and washed with brine (3 × 50 mL), and dried over Na_2SO_4. The mixture was concentrated under vacuum and purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (0% - 3% MeOH) yielding compound **12** (0.31 g, 84%) as a white form. ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.71 (s, 1H), 7.56 (d, J = 1.5 Hz, 1H), 6.29 (dd, J = 9.1, 5.4 Hz, 1H), 4.30 (d, J = 5.8 Hz, 1H), 4.20 (q, J = 1.9 Hz, 1H), 3.87 (ddd, J = 53.4, 11.4, 2.2 Hz, 2H), 2.51 (dd, J = 13.7, 5.4 Hz, 1H), 2.02 – 1.86 (m, 4H), 0.92 (s, 9H), 0.12 (d, J = 2.6 Hz, 6H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.88, 150.44, 135.61, 110.97, 85.27, 84.75, 84.13, 64.44, 37.61, 26.08, 18.50, 12.64, -5.21, -5.31.

3'-*N*-[(4,4'-dimethoxytrityl)-aminooxy]-thymidine (13)

4,4'-Dimethoxytriphenylmethyl chloride (DMT-Cl) (0.60 g, 1.8 mmol) was added to a solution of compound **12** (0.30 g, 0.81 mmol) in dry pyridine (5 mL). The mixture was stirred at ambient temperature under an argon atmosphere overnight. Pyridine was removed under vacuum. Then, THF (5 mL) and *tetra*-n-butylammonium fluoride trihydrate (0.70 mg, 2.2 mmol) were added and the mixture was stirred at ambient temperature for 5 h. The resulting mixture was poured into ethyl acetate (50 mL), and washed with brine (3 × 50 mL), and dried over Na_2SO_4. The mixture was concentrated under vacuum and purified by flash chromatography on a Biotage SP4 system using a dichloromethane/methanol (DCM/MeOH) gradient (0% - 2% MeOH) yielding compound **13** (0.40 g, 88%) as a light yellow form.

3'-*N*-[(4,4'-dimethoxytrityl)-aminooxy]-5'-[(2-cyanoethyl)-(N,N-diisopropyl)]-phosphoramidite-thymidine (14)

To a solution of compound **13** (56 mg, 0.1 mmol) and *N, N*-diisopropylethylamine (48 μL, 36 mg, 0.3 mmol) in dry DCM (3 mL) at 0 ^{o}C under argon atmosphere, 2-

cyanoethoxy *N, N*-diisopropylaminochlorophosphine (33 µL, 35 mg, 0.15 mmol) was added. The reaction mixture was stirred at ambient temperature for 2 h. The resulting mixture was diluted with DCM (20 mL), washed with cold $NaHCO_3$ (20 mL). The DCM layer was dried over Na_2SO_4. The organic layer was concentrated under vacuum. The oily residues was kept under high vacuum for 2 h, dissolved in acetonitrile (1 mL) and used for oligonucleotide synthesis without further purification.

Experimental section

Oligonucleotide synthesis

ODNs with commercial avabliable modifications were purchased from Eurogentec. 5'-Diol, 3'-aminooxy and intern oxime modified ODNs were synthesized on a 1 µmol scale on a MerMade 4 Oligonucleotide synthesizer (BioAutomation Corporation, USA).For the coupling step, 5-Ethylthio-1H-tetrazole (ETT) was used as activator (0.5 M in anhydrous CH_3CN), with a coupling time of 2 x 15 s for standard nucleoside phosphoramidites (0.1 M in anhydrous CH_3CN) and 4 x 30 s for phosphoramidites **9** and **14** (0.1 M in anhydrous CH_3CN). The capping step was performed with acetic anhydride by using commercial solutions (Cap A: Ac_2O/pyridine/THF, 10:10:80, v/v/v; Cap B: 10 % N-methylimidazole in THF) for 30 s. Oxidation was performed for 15 s by using 0.1 M Iodine in THF/pyridine/water (78:20:2). Detritylation was performed with 3 % trichloroacetic acid (TCA) in DCM for 2 x 30 s. Final trityl group were remained (DMT-on) for all ODNs. For diol modified ODNs, standard DNA phosphoramidites and phosphoramidite 9 were used (Glen Research, USA). For aminooxy modified ODN, 5' -> 3' synthesis phosphoramidites and phosphoramidite 14 were used (Glen Research, USA). For intern oxime modified ODN, the sequence was first synthesized until the modification site using standard DNA phosphoramidites. After detritylation, CPG-linked ODN were stirred in a 0.5 M *N, N'*-dicyclohexylcarbodiimide in DMSO (contain 1.7% dichloroacetic acid) mixture for 30 min at ambient temperature. The solvent was filtered off and the CPG washed with DMSO and acetonitrile. Then compound **12** (20 mg) was added to the CPG and suspended in methanol (250 µL) and acetic acid (10% in methanol, 2.5µL). The mixture was incubated overnight at ambient temperature. The solvent was then filtered off and the CPG was washed with methanol and DCM. Then, 5'-TBDMS on the modified nucleotide was removed by treating the CPG with 1 M tetrabutylammonium fluoride hydrate in THF for 4 hours. The solvent was filtered off and the CPG was washed with dry THF and acetonitrile. The second half of the sequence after modification was synthesized automatically using standard DNA phosphoramidites.

The deprotection of modified ODNs were carried out with 30% aqueous NH_4OH at ambient temperature for 16 hours. Decant the supernatant liquid from the support and evaporate to dryness with the addition of TEA. The ODNs purified by HPLC (Agilent 1200 Series) using a Phenomenex Luna C18 250 x 4.6 mm column with a gradient of acetonitrile (10% to 40% buffer B over 20 min, flow rate 1 mL/min), buffer A: 0.05 M triethylammonium bicarbonate buffer, pH 7.5, buffer B: acetonitrile. Elution was monitored by UV absorption at 260 nm. After HPLC purification, ODNs were concentrated to dryness in a MiVac centrifugal evaporator (GeneVac). Aminooxy modified ODNs with DMT-on were resuspended in deionized water for further use. Diol and oxime modified ODNs were re-dissolved in 200 µL 80% acetic acid and stood for 30 minutes at room temperature. Acetic acid was removed by a MiVac centrifugal evaporator. The ODNs were desalted by sep-pak columns, concentrated and re-dissolved in deionized water. The composition was confirmed by direct-injection mass spectrometric analysis with an Thermo LTQ Orbitrap Velos.

Labelling of an 5’-aldehyde modified oligonucleotide with the code sequence

The 5’-aldehyde modified ODN was prepared by treating diol-modified ODN (Diol-F, 2 µM) with 25 mM $NaIO_4$ in 100 mM NaOAc (20 µL, pH 6.0) at room temperature for 60 min. The ODN was desalted by passing through a micro bio-spin 6 column (Bio-Rad). The aminooxy modified ODNs without DMT protection were prepared freshly every time before use by treating Bar-AO (1 µL) with acetic acid (4 µL) at room temperature for 30 minutes. Acetic acid was removed by a MiVac centrifugal evaporator. Then, 18 µL 5’-aldehyde modified ODN and 2 µL NaOAc (1 M, pH 6.0) were added into dried aminooxy modified ODN. The reaction mixture was incubated at room temperature for 2 hours. The reaction was quenched with formamide loading buffer, and analyzed by denaturing urea polyacrylamide gel electrophoresis (urea-PAGE). Gels were imaged with a ChemiDoc XRS+ System (Bio-Rad). Band intensities were quantified with Image Lab (Bio-Rad).

Labelling of an oligonucleotide containing abasic site with the code sequence

The ODN containing abasic site was prepared with 3 µM T21F-U, 2 µL UDG (New England Biolabs (NEB), 5 U/ µL) in 100 µL 1 × UDG buffer at 37 °C for 1 h. The resulting

ODN was purified with Monarch PCR & DNA Cleanup Kit (NEB) using a protocol for ODNs. The ODN containing abasic site was characterized by ESI-MS and human apurinic/apyrimidinic endonuclease (APE1) digestion. The labelling reaction was carried out between the ODN containing abasic site and the aminooxy modified ODNs as described above for the 5'-aldehyde modified ODN. The reaction was quenched with formamide loading buffer, and analyzed by urea-PAGE. Gels were imaged with a ChemiDoc XRS+ System (Bio-Rad). Band intensities were quantified with Image Lab (Bio-Rad).

To produce the labelled product for polymerases bypass study, 2 nmol purified ODNs containing an abasic site and 10 nmol aminooxy modified ODNs were reacted in 50 µL PBS buffer (100 mM, pH 7.0) at ambient temperature overnight. The ODNs were desalted by passing through a micro bio-spin 6 column (Bio-Rad). The resulting mixture was concentrated to dryness and re-dissolved in 5 µL water. Target ODN was purified by urea-PAGE. The labelled product was characterized by urea-PAGE and ESI-MS.

Preparation of site-specific modified dsDNA

dsDNA fragment containing a site-specific modification was prepared by ligation of a short modified ODNs and a 0.9-kb gapped DNA duplex. ODNs and primers are presented in Table S1. 5'-aldehyde or abasic site modified ODN was prepared as described above.

In brief, a pEGFP-W1 plasmid was constructed to contain two Nb.BbvCI, a nicking endonuclease, cleavage sites. A 0.9 kb DNA duplex was amplified from the plasmid with primers GFP-Pr409 and GFP-Pr1296 and Taq DNA polymerase (NEB), and then subjected to Nb.BbvCI digestion to generate a gapped dsDNA. The 20-mer cleaved single-stranded DNA was removed by annealing with a 20-mer complementary ODN in large excess. The gapped dsDNA was purified with Monarch Nucleic Acid Purification Kit (NEB). Then, 5'-aldehyde or abasic site modified ODN was annealed together with gapped dsDNA. Two nick sites were ligated by T4 DNA ligase at 16 °C for 4 h and purified with Monarch Nucleic Acid Purification Kit (NEB).

	Sequence (5' > 3')	Resource	MW (theor.)	MW (exp.)	Stock (μM)
Diol-F	[Diol]TGA AGA TGT GGC TTT[FAM]	Synthesis	5267.5	5267.1	10
GFP-diol	[Diol]TCAGTCCCGGGCGGGC	Synthesis	4929.2	4929.4	10
T21-ON	TGCACGT[ON]ATTGAAGATGTGGC	Synthesis	6434.3	6435.1	10
Bar-AO	[Biotin]ACGCTCTTCCGATCTGTAC T[ONHDMT]	Synthesis	6905.1	6906.0	1000 (1:1:1:1)
	[Biotin]ACGCTCTTCCGATCTAGCT T[ONHDMT]	Synthesis	6905.1	6905.8	
	[Biotin]ACGCTCTTCCGATCTCATG T[ONHDMT]	Synthesis	6905.1	6906.3	
	[Biotin]ACGCTCTTCCGATCTTCGA T[ONHDMT]	Synthesis	6905.1	6905.9	
P14F	[FAM]GCCACATCTTCAAT	Eurogentec	--	--	10
T21F-U	TGCACGUATTGAAGATGTGGC[FAM]	Eurogentec	--	--	10
Ex-Bar	TACACGACGCTCTTCCGATCT GTACT	Eurogentec	--	--	10 (1:1:1:1)
	TACACGACGCTCTTCCGATCT AGCTT	Eurogentec	--	--	
	TACACGACGCTCTTCCGATCT CATGT	Eurogentec	--	--	
	TACACGACGCTCTTCCGATCT TCGAT	Eurogentec	--	--	
GFP-Pr409	TGACTCACGGGGATTTCCAAGTCT	Eurogentec	--	--	10
GFP-Pr1296	TTCTCGTTGGGGTCTTTGCTCA	Eurogentec	--	--	10
GFP-U648	phos-TGAGUCAGTCCCTGGCGGGC	Eurogentec	--	--	10

Table S1: Oligonucleotides used in this study.

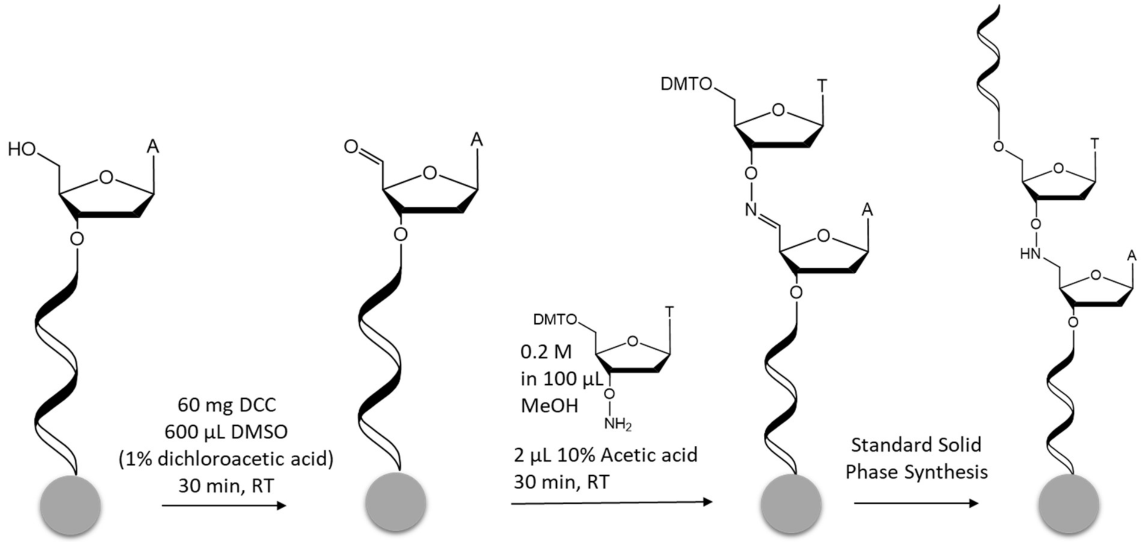


Figure S1: On-support synthesis of oxime modified ODN. A fully protected 5’-OH ODN on a solid support is mildly oxidized by Moffat reaction. The resulting 5’-aldehyde ODN is then reacted with 3’-aminooxy modified nucleoside analogues to form the oxime linkage. Full-length ODN is further synthesized and purified with standard solid phase synthesis protocol.

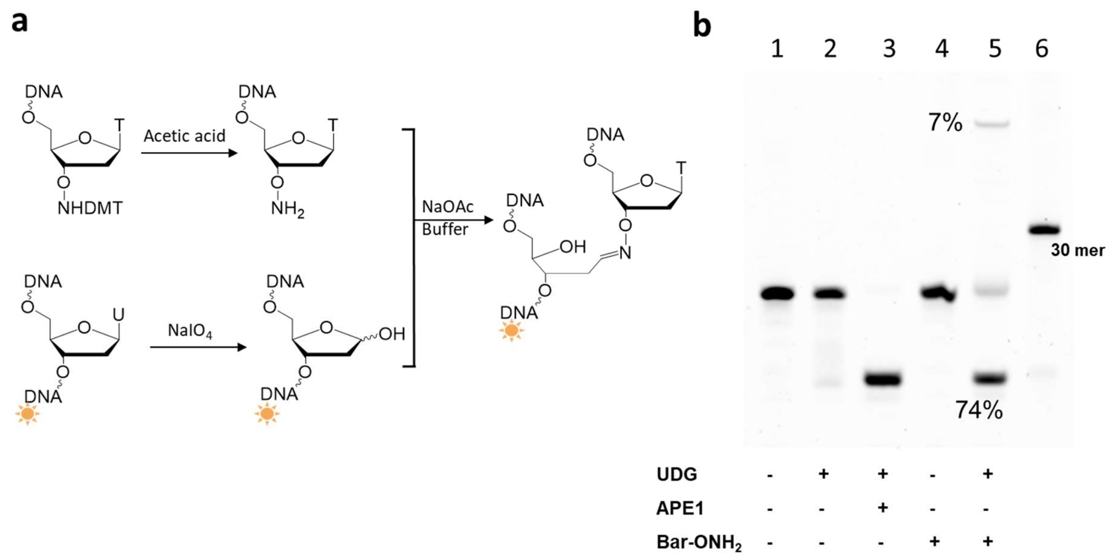


Figure S2: Labelling of abasic site containing ODN with code sequence. (A) Structures of 3' aminooxy-functionalized ODN, abasic site containing ODN and their conjugated product. (B) Denaturing PAGE analysis of labelling reaction. Only fluorescent labelled abasic site containing ODN is visible. Lane 6 shows a 30 mer ODN marker. Faster migrating product in lane 3 is attributed to abasic site excision product. Slower migrating product in lane 5 is attributed to conjugated oxime linkage ODN.

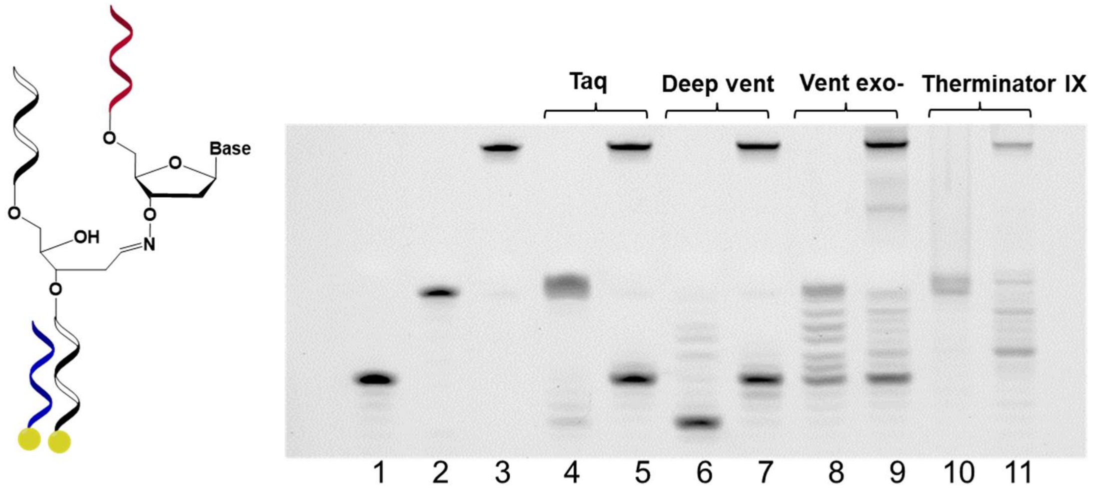


Figure S3: Bypass of DNA template containing a abasic site labelled linkage (lanes 5, 7, 9, 11) or identical template without modification (lanes 4, 6, 8, 10) by different DNA polymerases. Lane 1: primer only, lane 2: 21 mer marker, lane 3: conjugated oxime linkage ODN.

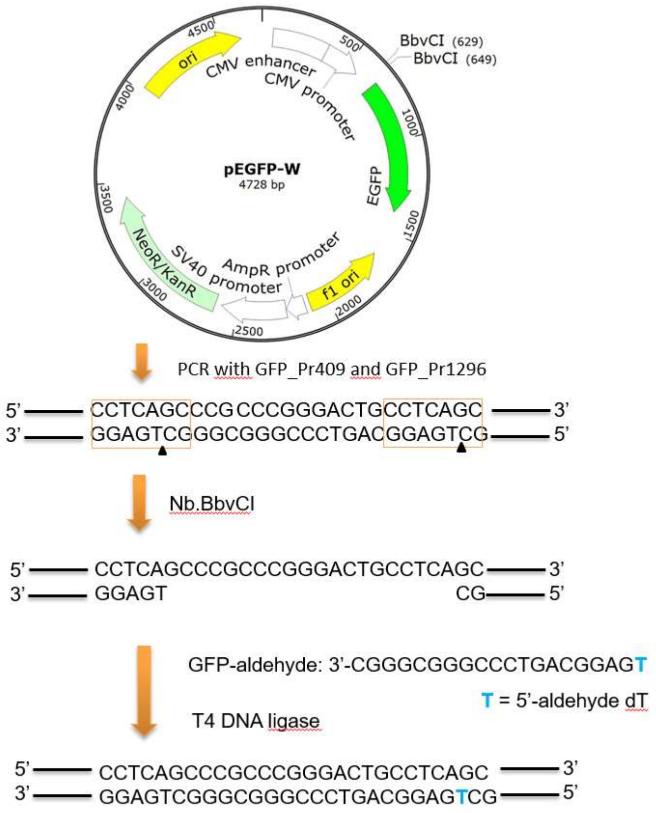

4500
500
BbvCI (629)
BbvCI (649)
ori
CMV enhancer
CMV promoter
1000
4000
pEGFP-W
4728 bp
EGFP
3500
1500
NeoR/KanR
SV40 promoter
AmpR promoter
f1 ori
3000
2500
2000
PCR with GFP_Pr409 and GFP_Pr1296
5' CCTCAGCCCG CCCGGGACTGCCTCAGC 3'
3' GGAGTCGGGCGGGCCCTGACGGAGTCG 5'
Nb.BbvCI
5' CCTCAGCCCGCCCGGGACTGCCTCAGC 3'
3' GGAGT CG 5'
GFP-aldehyde: 3'-CGGGCGGGCCCTGACGGAGT
T = 5'-aldehyde dT
T4 DNA ligase
5' CCTCAGCCCGCCCGGGACTGCCTCAGC 3'
3' GGAGTCGGGCGGGCCCTGACGGAGTCG 5'

Figure S4: The plasmid map of pEGFP-W used in this study to generate site-specific modified dsDNA.

	Native dsDNA	5'-AT dsDNA	Damage/ Total nt
A	300 ng	0	0
B	0	300 ng	5×10^{-4}
C	270 ng	30 ng	5×10^{-5}
D	297 ng	3 ng	5×10^{-6}
E	299.7 ng	0.3 ng	5×10^{-7}

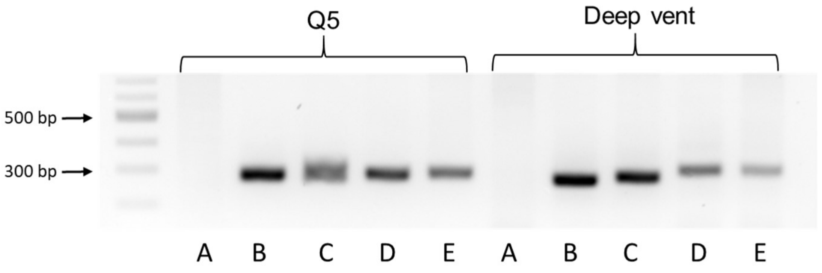
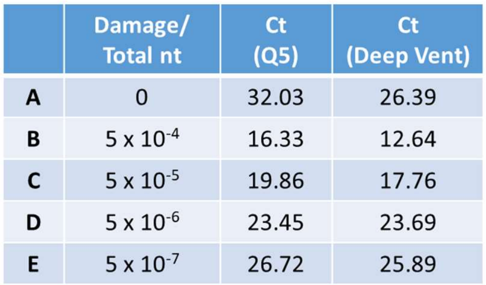


Figure S6: Detection limit study by Q5 or Deep vent polymerase. Various amounts of 5'-AT modified dsDNA were mixed with native dsDNA as listed on the table. After labelling reaction, the resulting DNA samples were amplified by PCR.



	Damage/ Total nt	Ct (Q5)	Ct (Deep Vent)
A	0	32.03	26.39
B	5 x 10^{-4}	16.33	12.64
C	5 x 10^{-5}	19.86	17.76
D	5 x 10^{-6}	23.45	23.69
E	5 x 10^{-7}	26.72	25.89

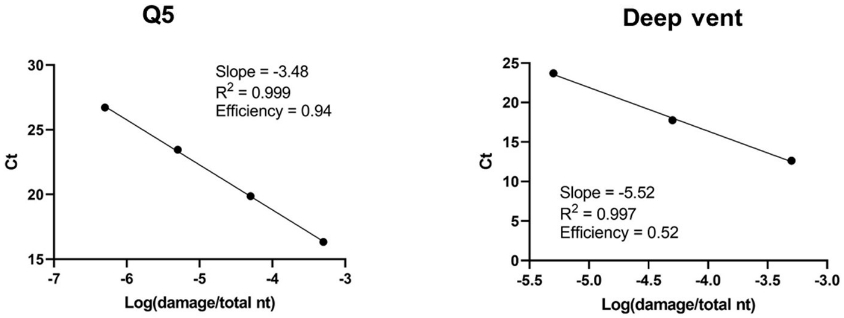
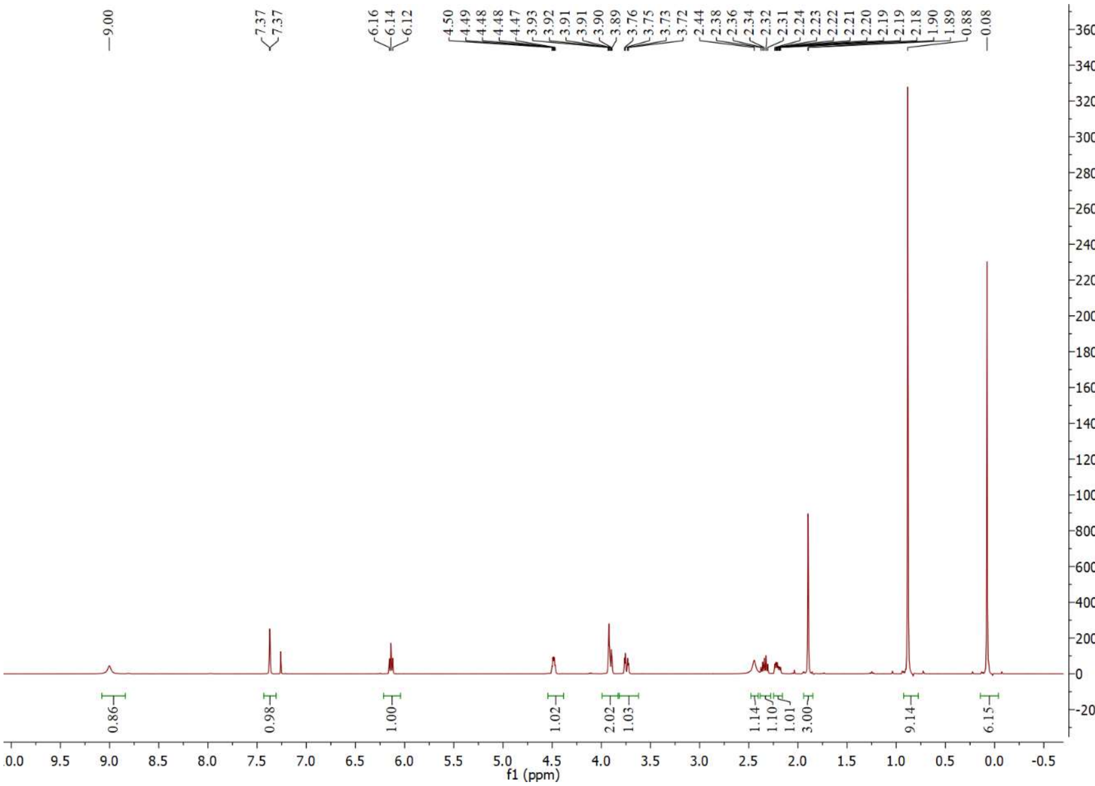
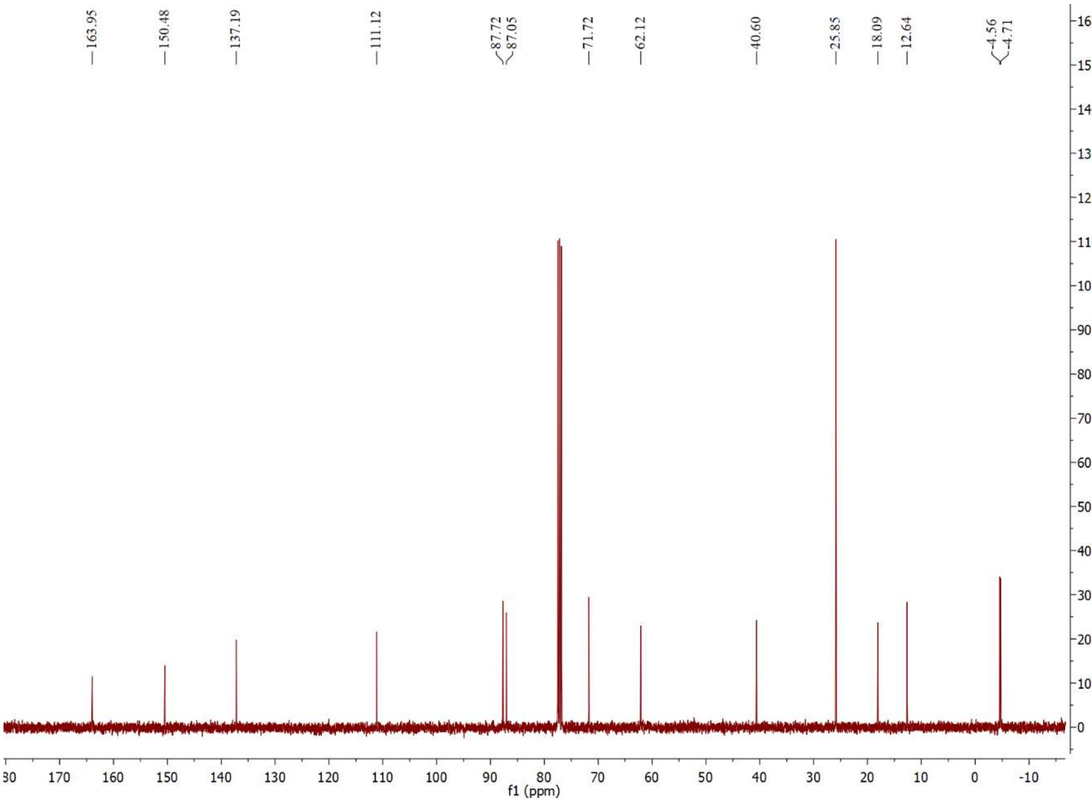


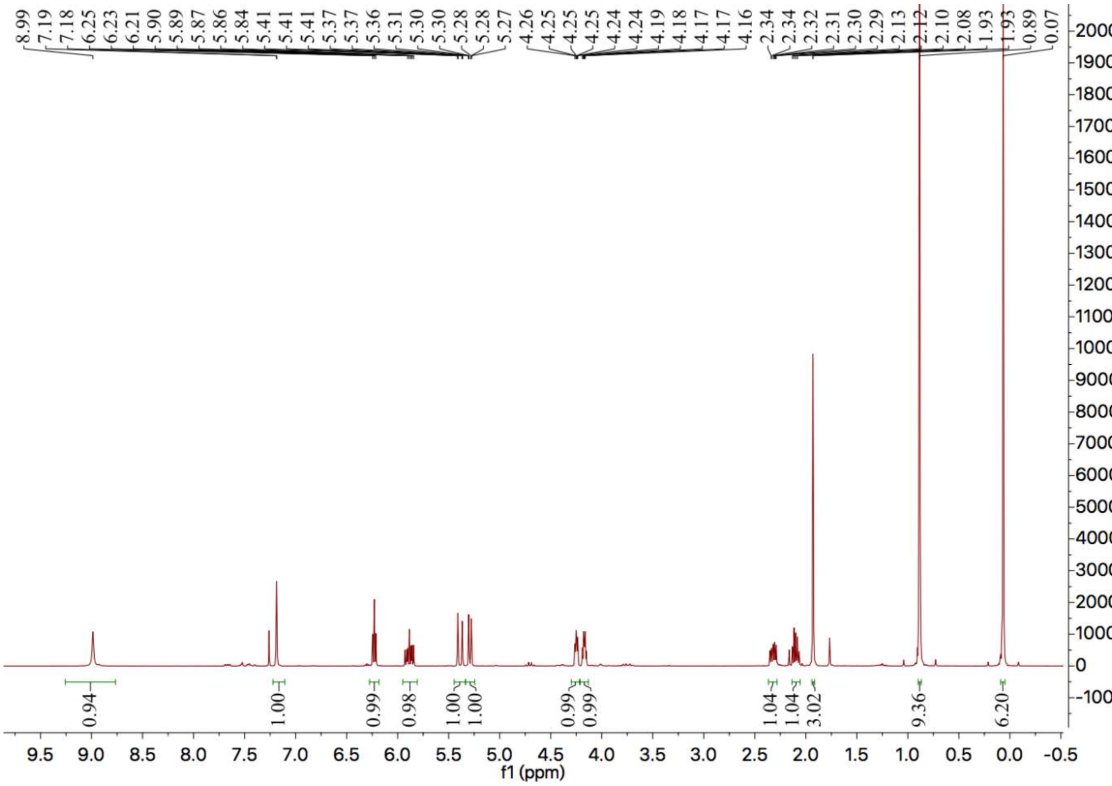
Figure S7: (A) qPCR Ct values using Q5 or Deep vent polymerase as a function of relative 5’-AT concentrations ([5’-AT lesion]/[total DNA]).



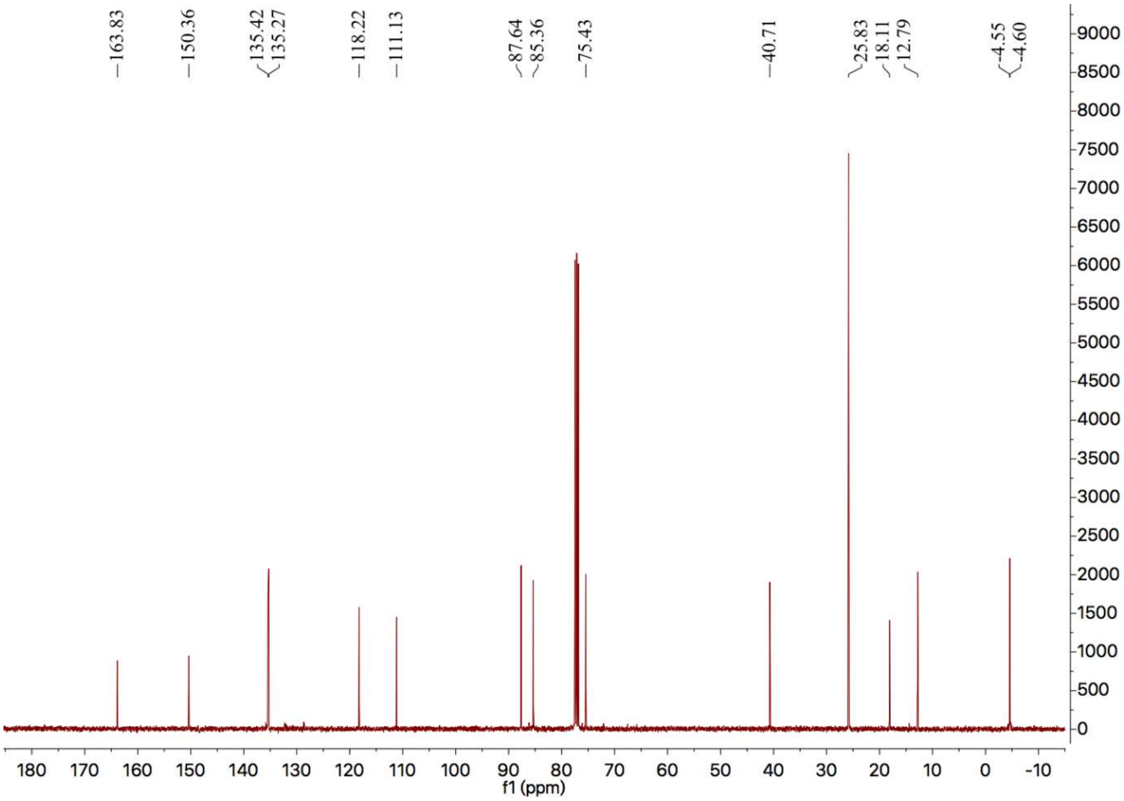
¹HNMR of compound 3



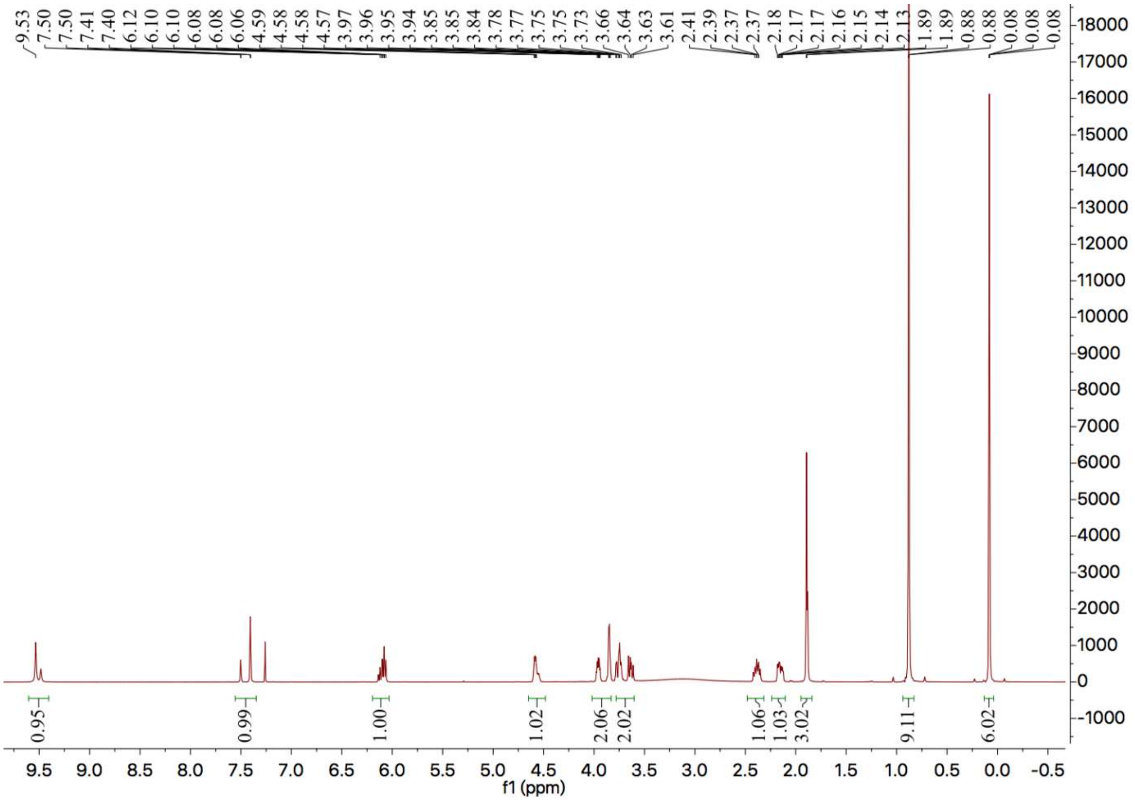
¹³CNMR of compound 3



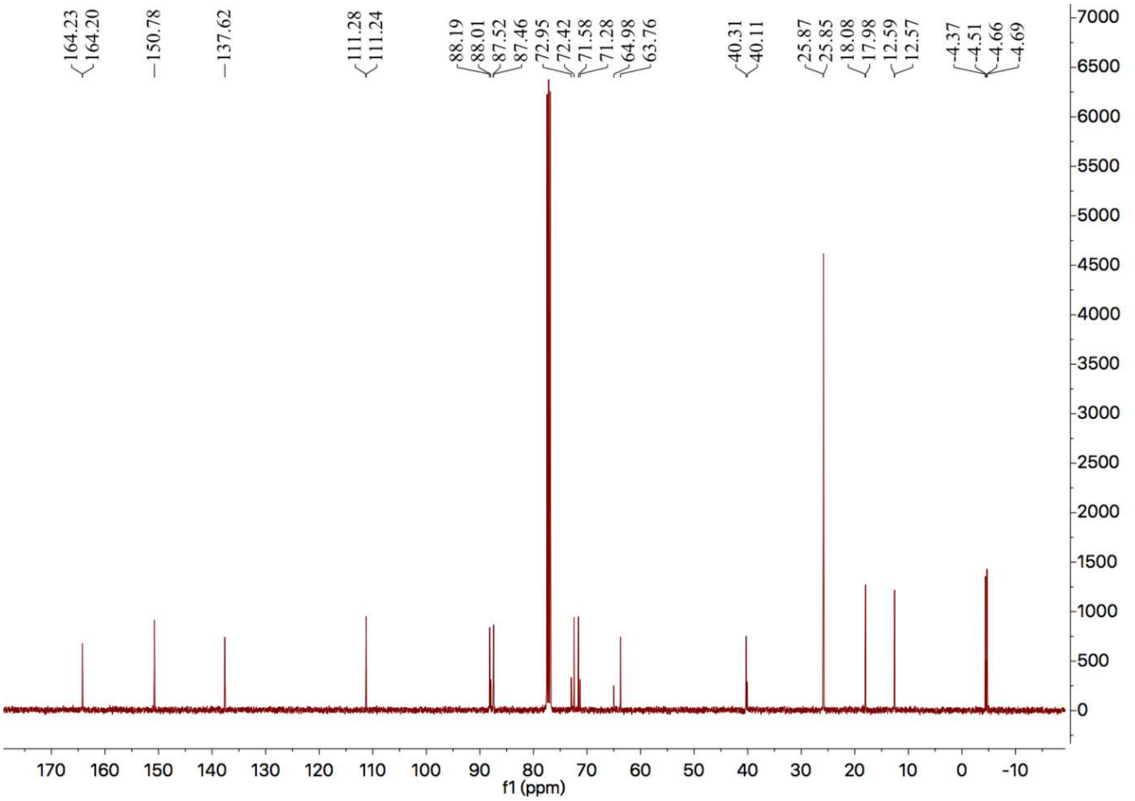
^{1}HNMR of compound 4



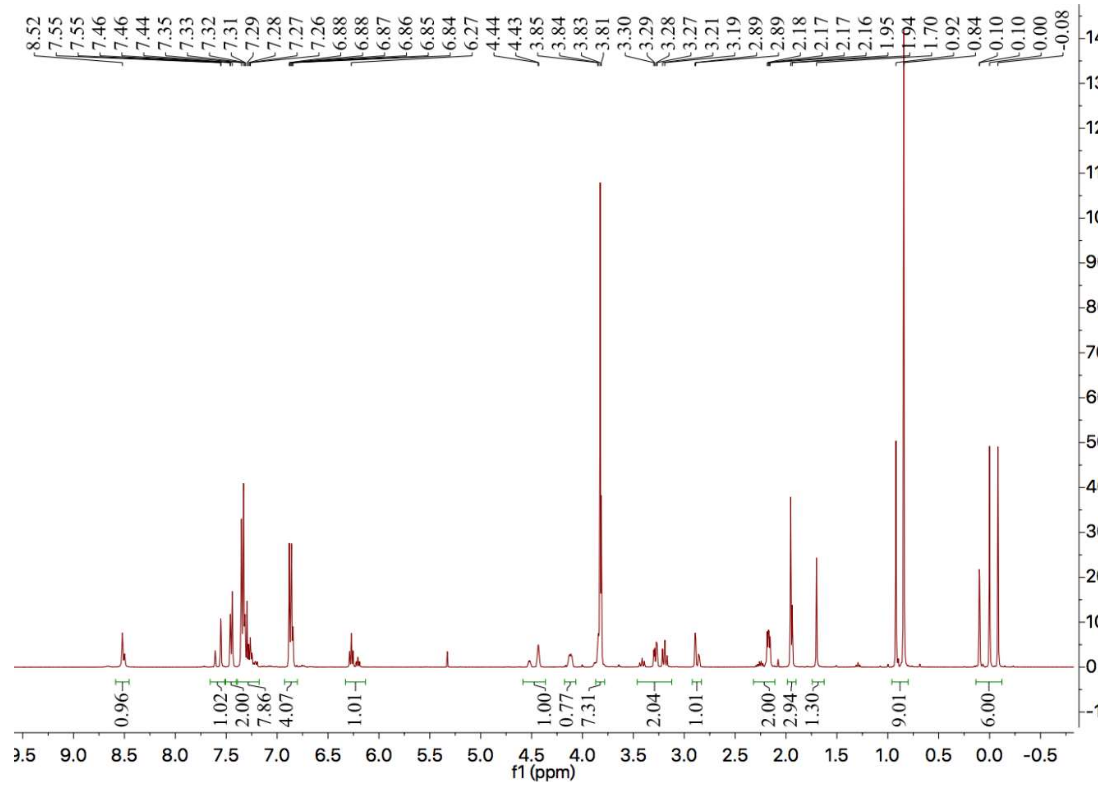
^{13}CNMR of compound 4



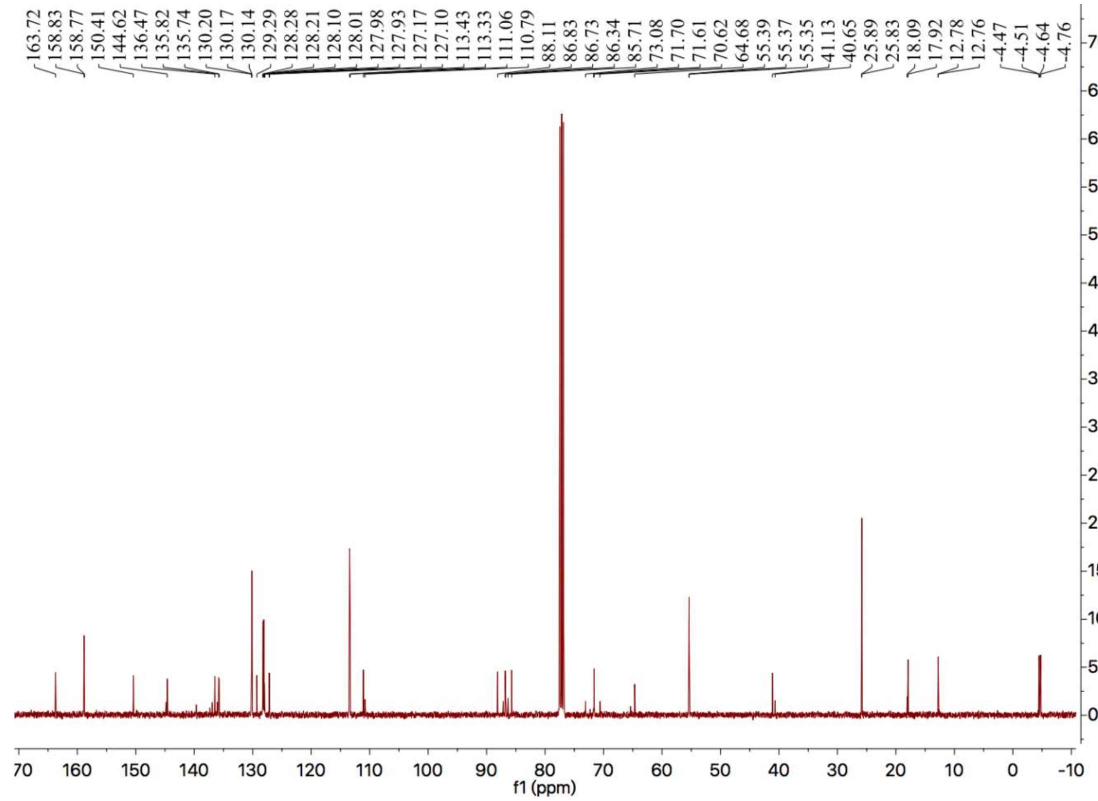
^{1}HNMR of compound 5



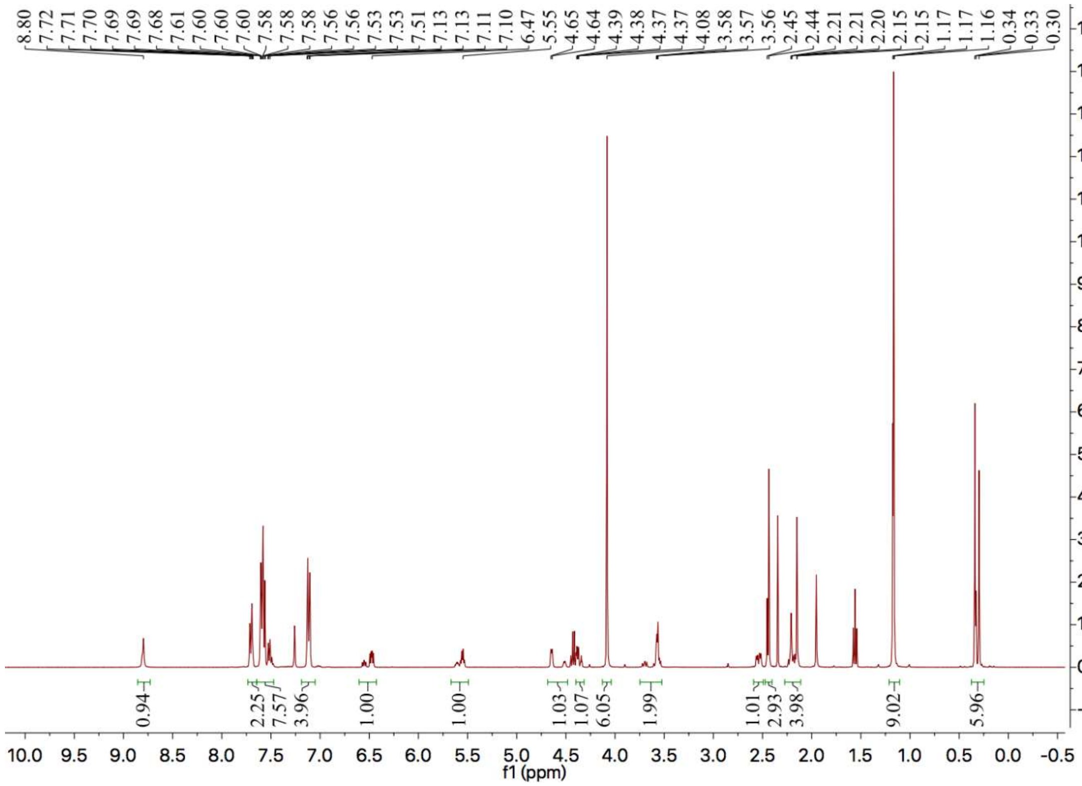
^{13}CNMR of compound 5



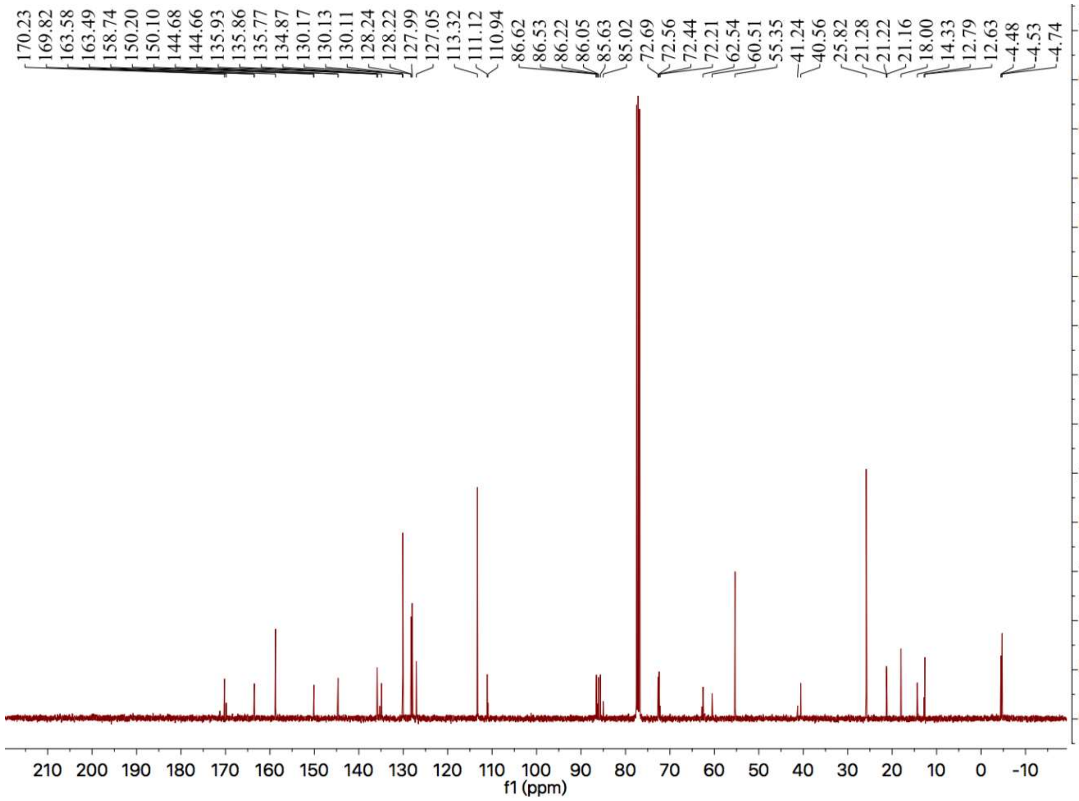
^{1}HNMR of compound 6



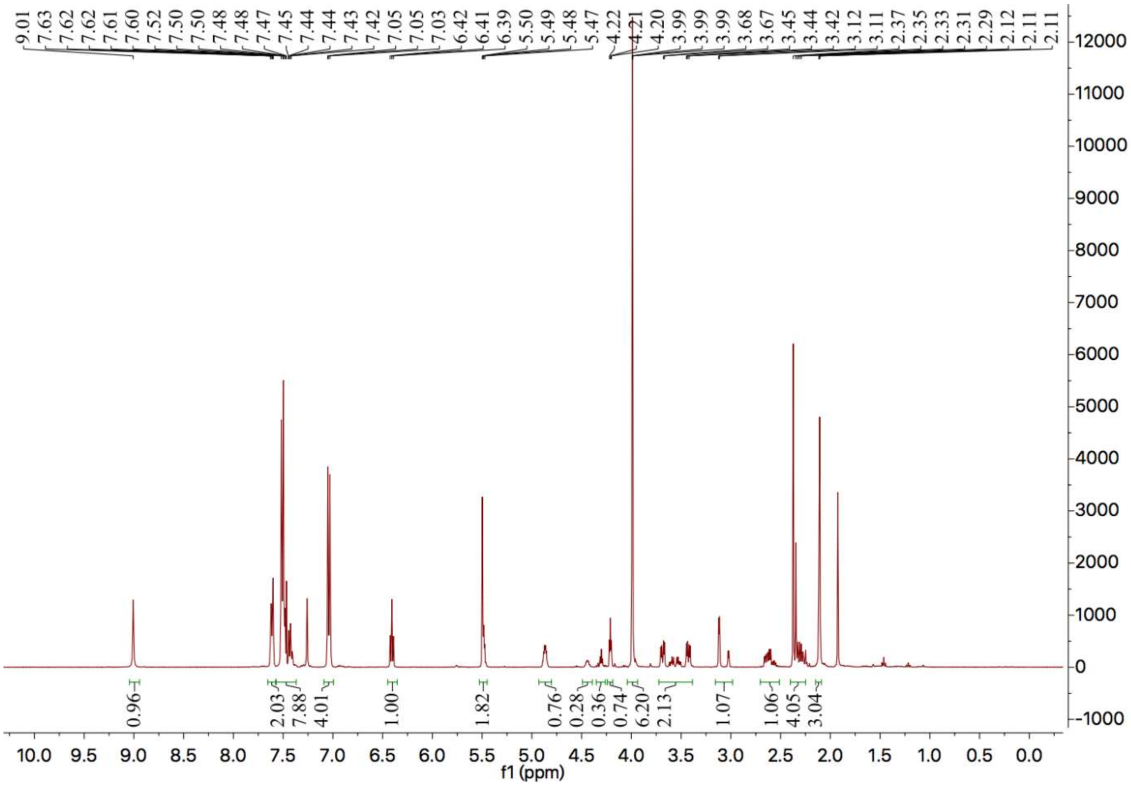
^{13}CNMR of compound 6



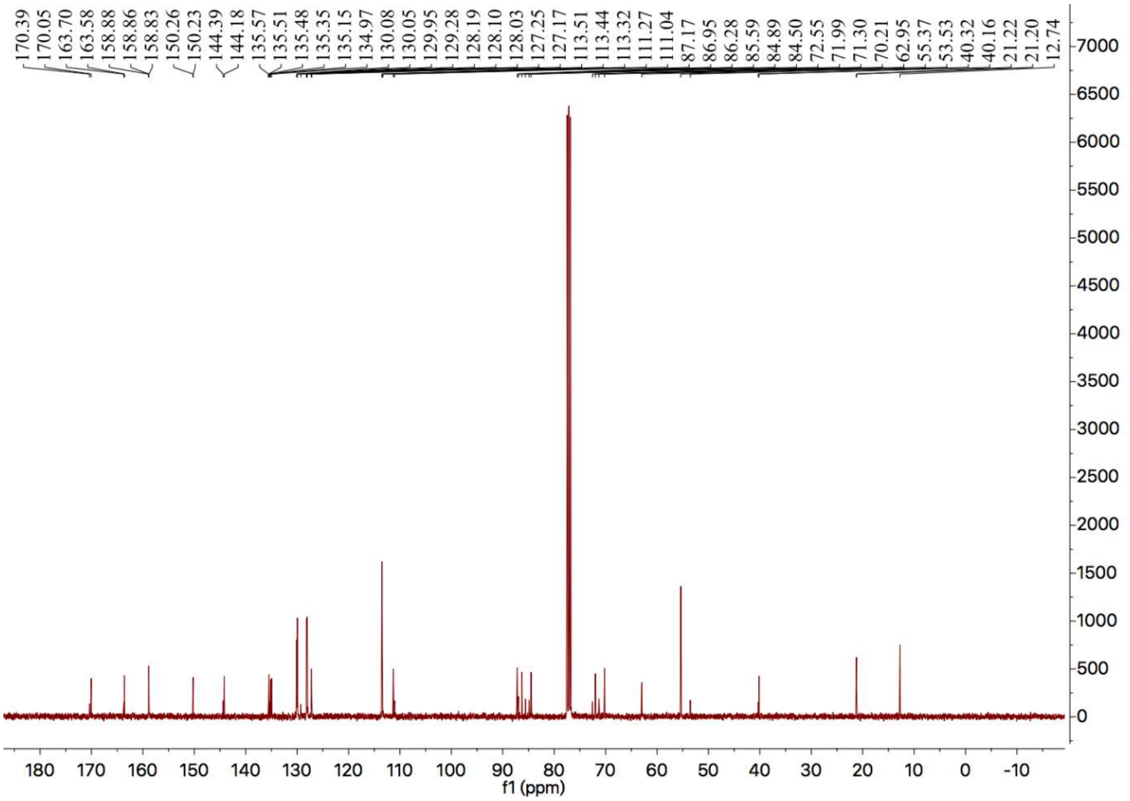
¹HNMR of compound 7



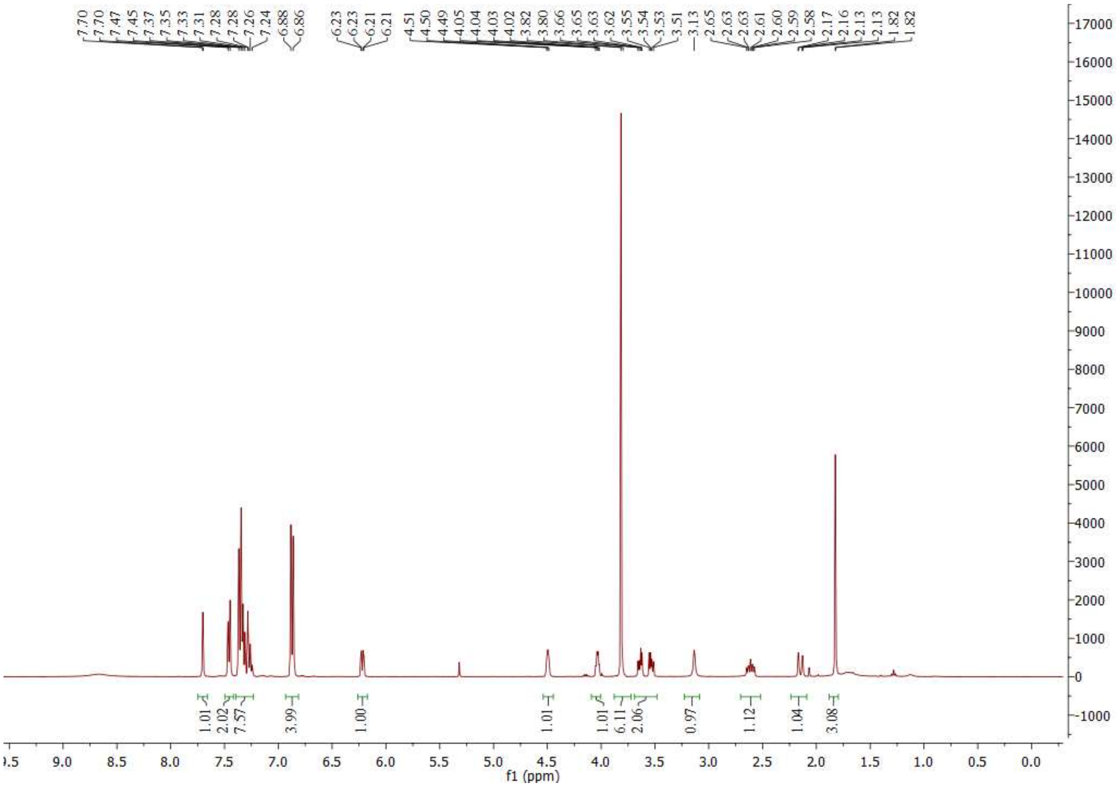
¹³CNMR of compound 7



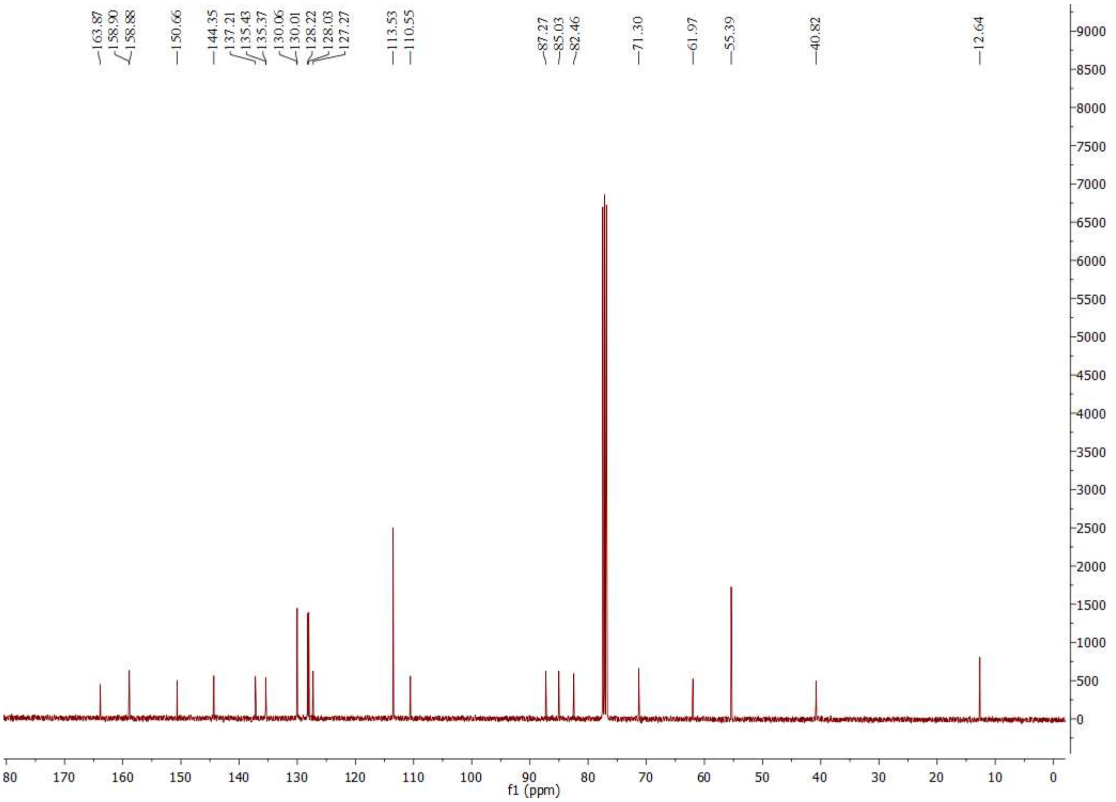
^{1}HNMR of compound 8



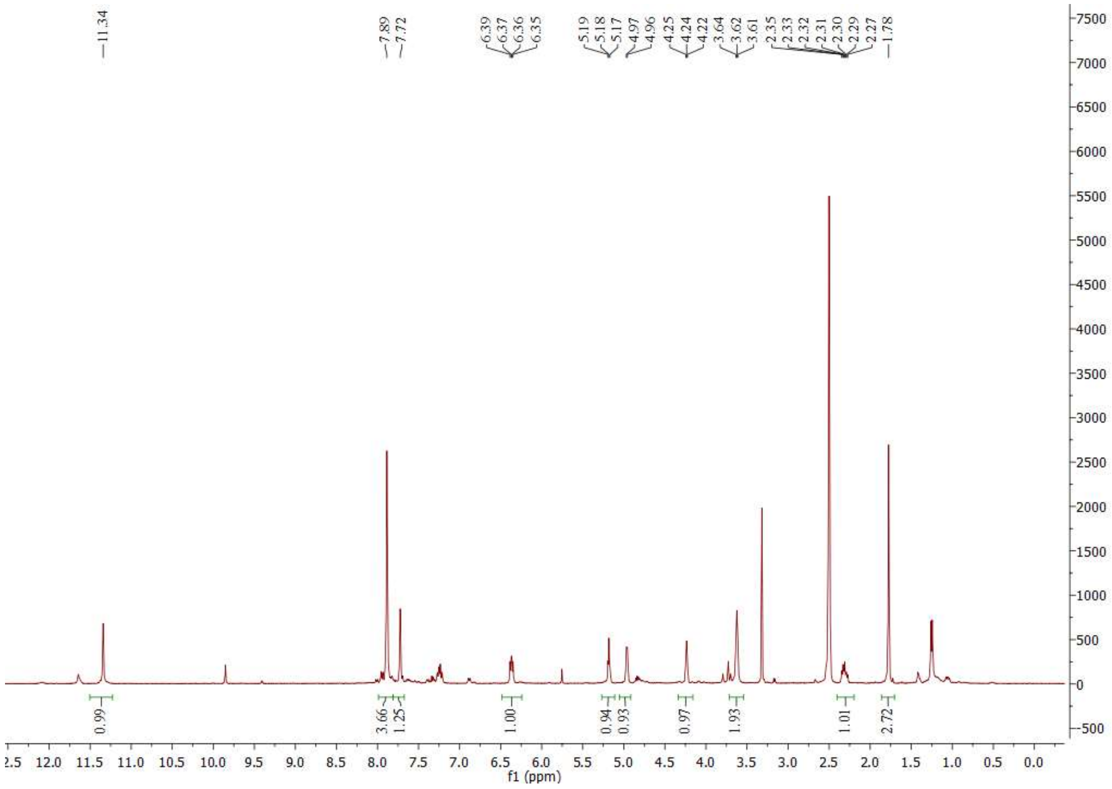
^{13}CNMR of compound 8



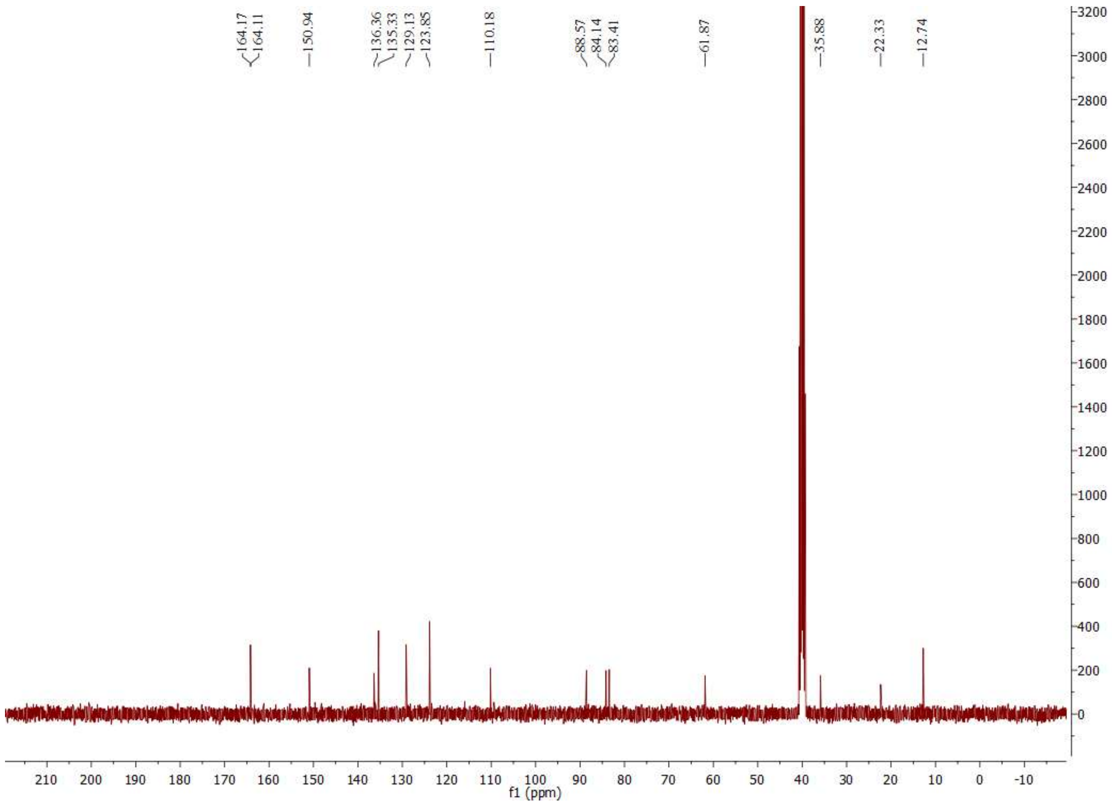
^{1}HNMR of compound 10



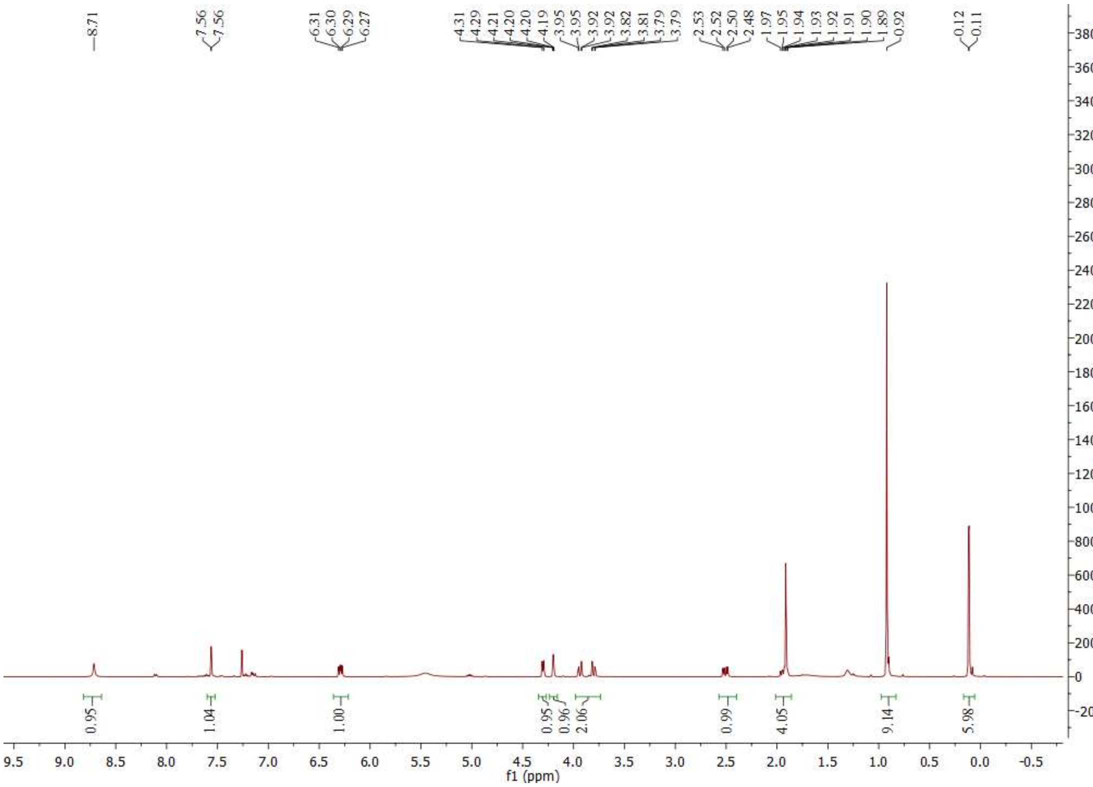
^{13}CNMR of compound 10



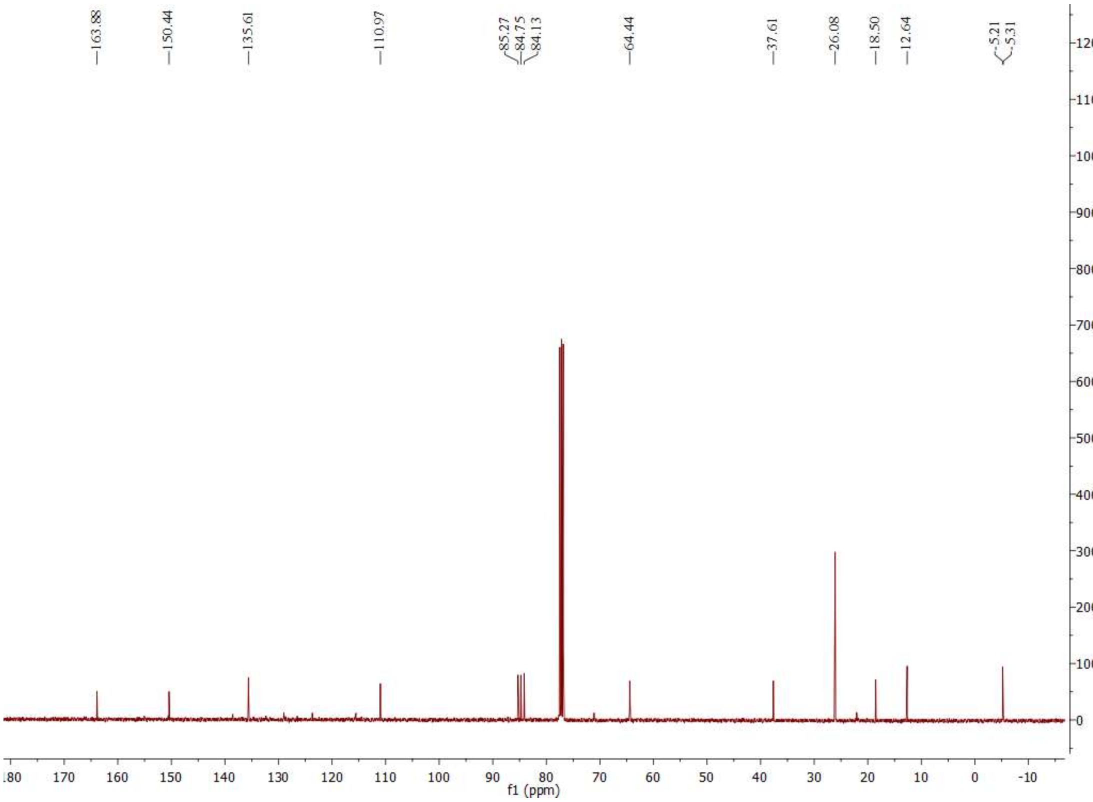
^{1}HNMR of compound 11



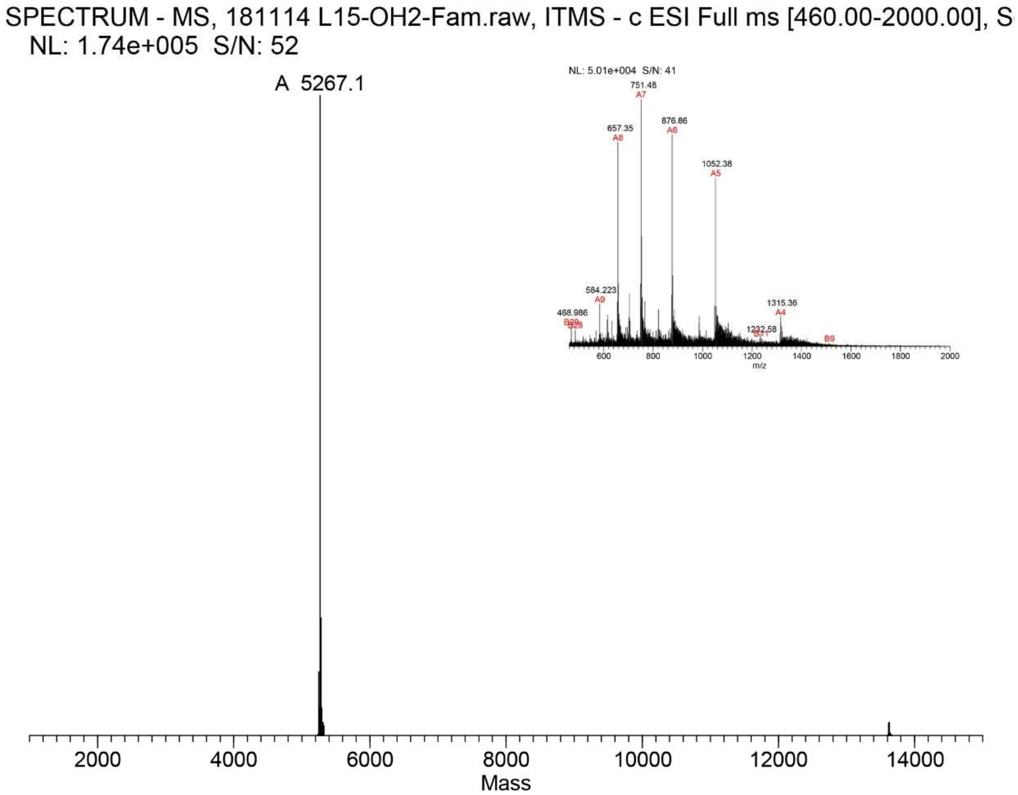
^{13}CNMR of compound 11



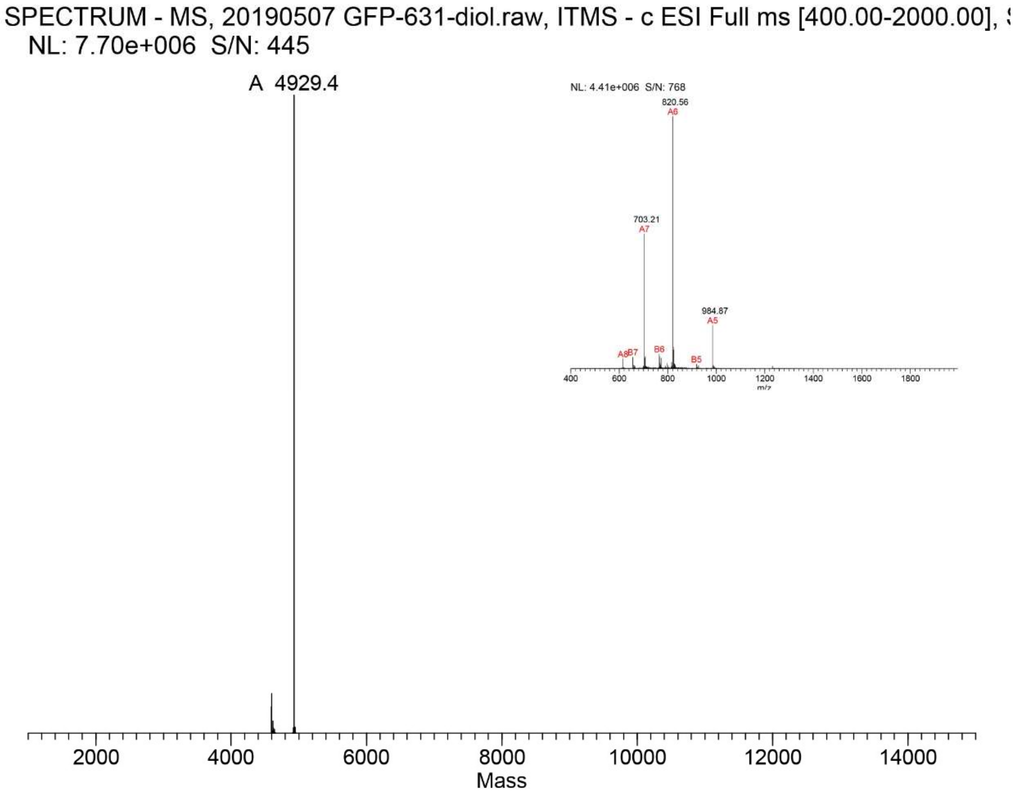
¹HNMR of compound 12



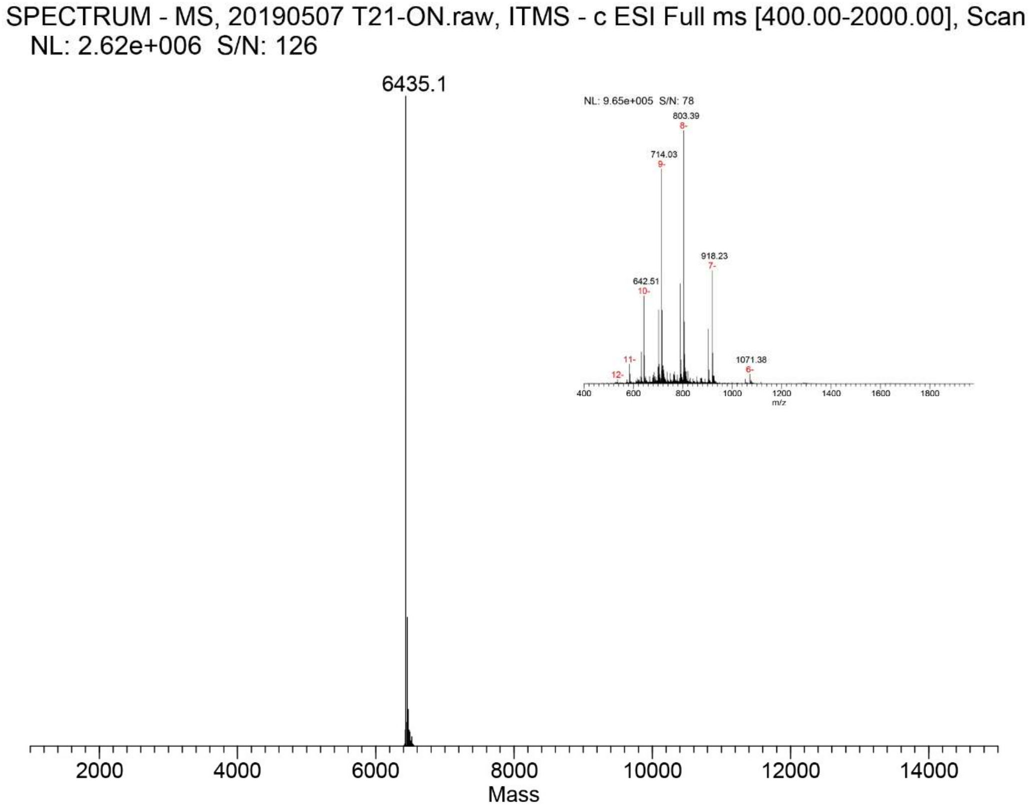
¹³CNMR of compound 12



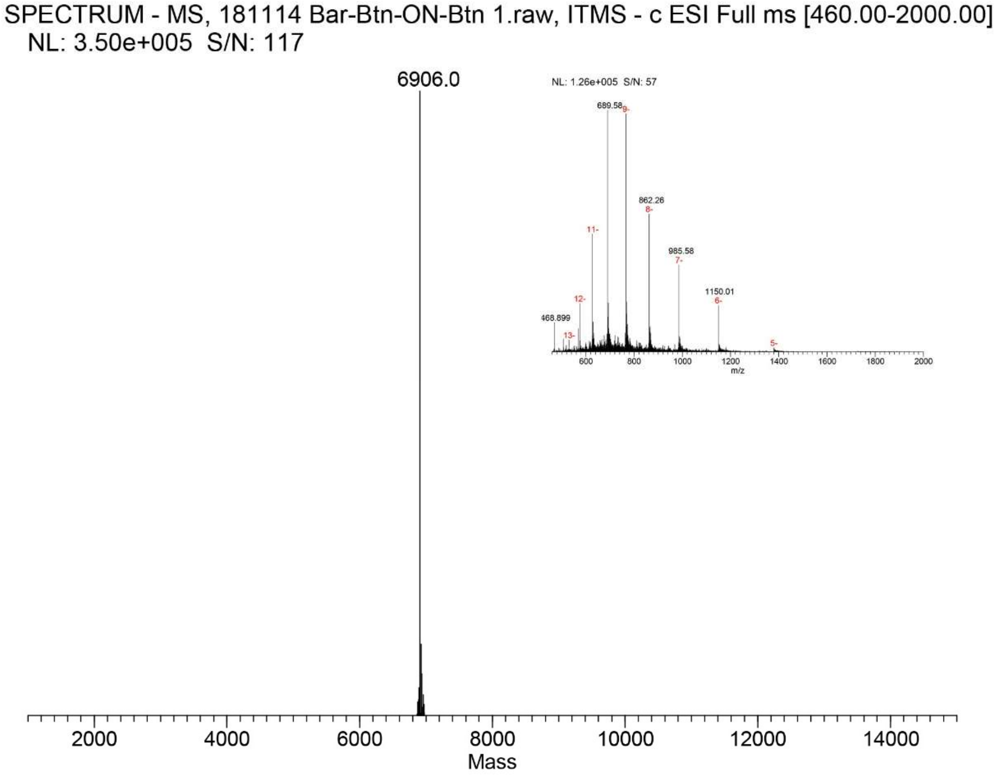
Mass spectrometry characterization of Diol-F



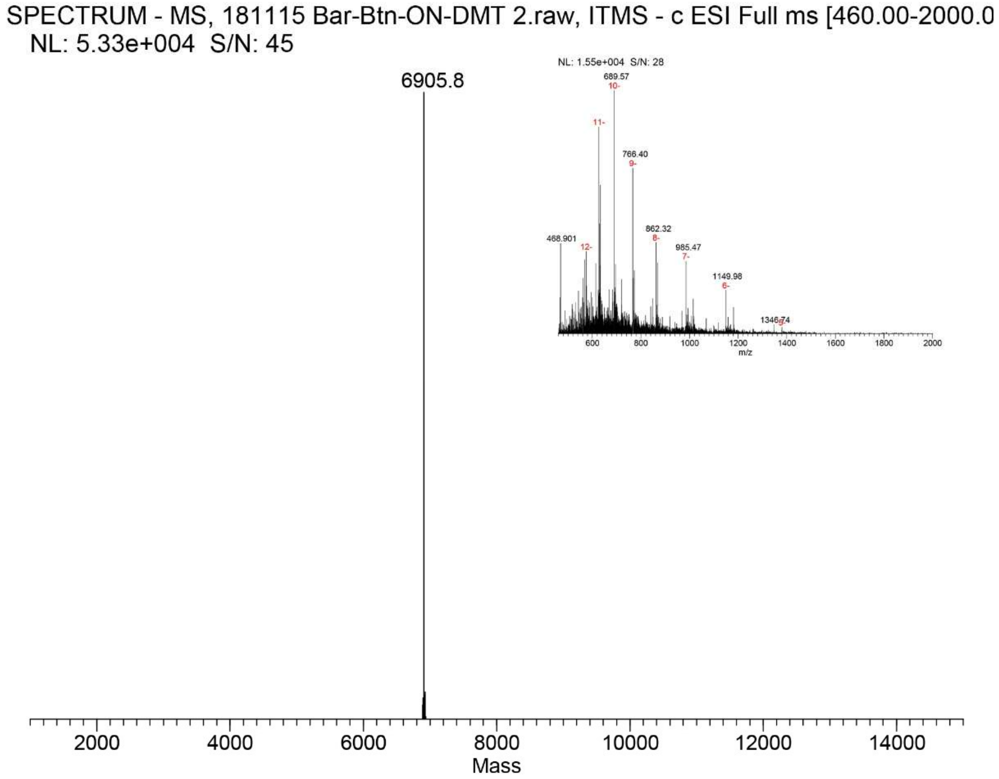
Mass spectrometry characterization of GFP-Diol


SPECTRUM - MS, 20190507 T21-ON.raw, ITMS - c ESI Full ms [400.00-2000.00], Scan
NL: 2.62e+006 S/N: 126

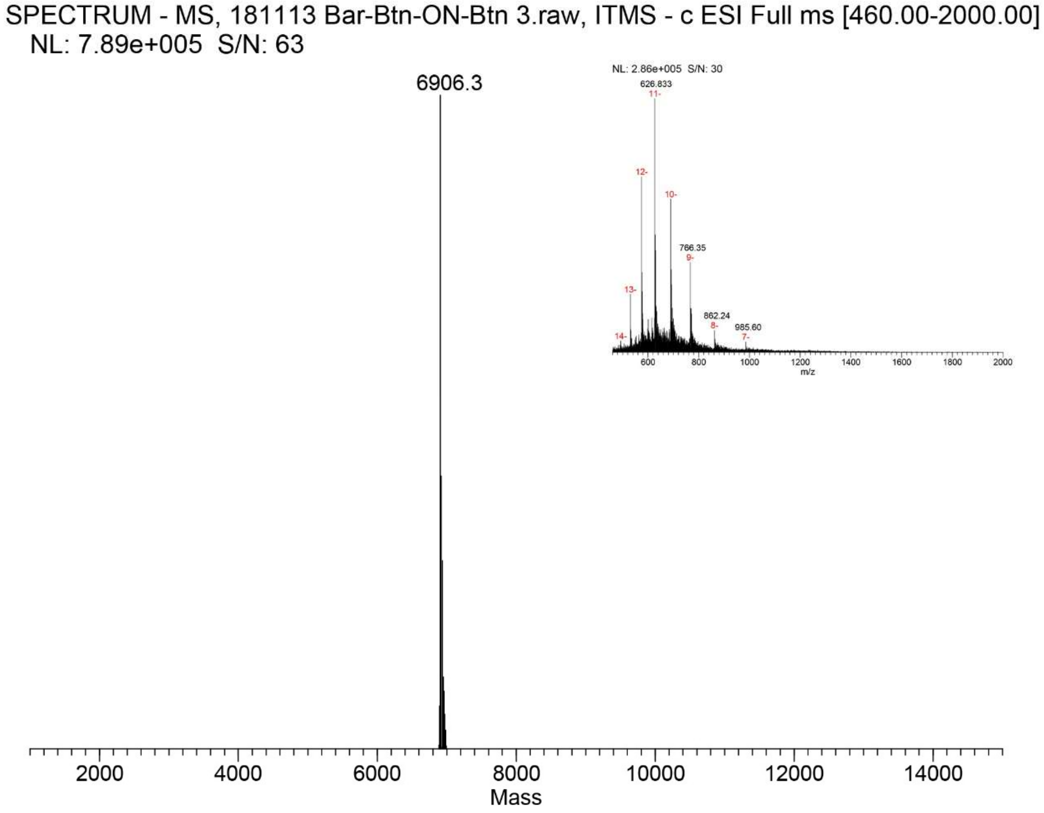
Mass spectrometry characterization of T21-ON


SPECTRUM - MS, 181114 Bar-Btn-ON-Btn 1.raw, ITMS - c ESI Full ms [460.00-2000.00]
NL: 3.50e+005 S/N: 117

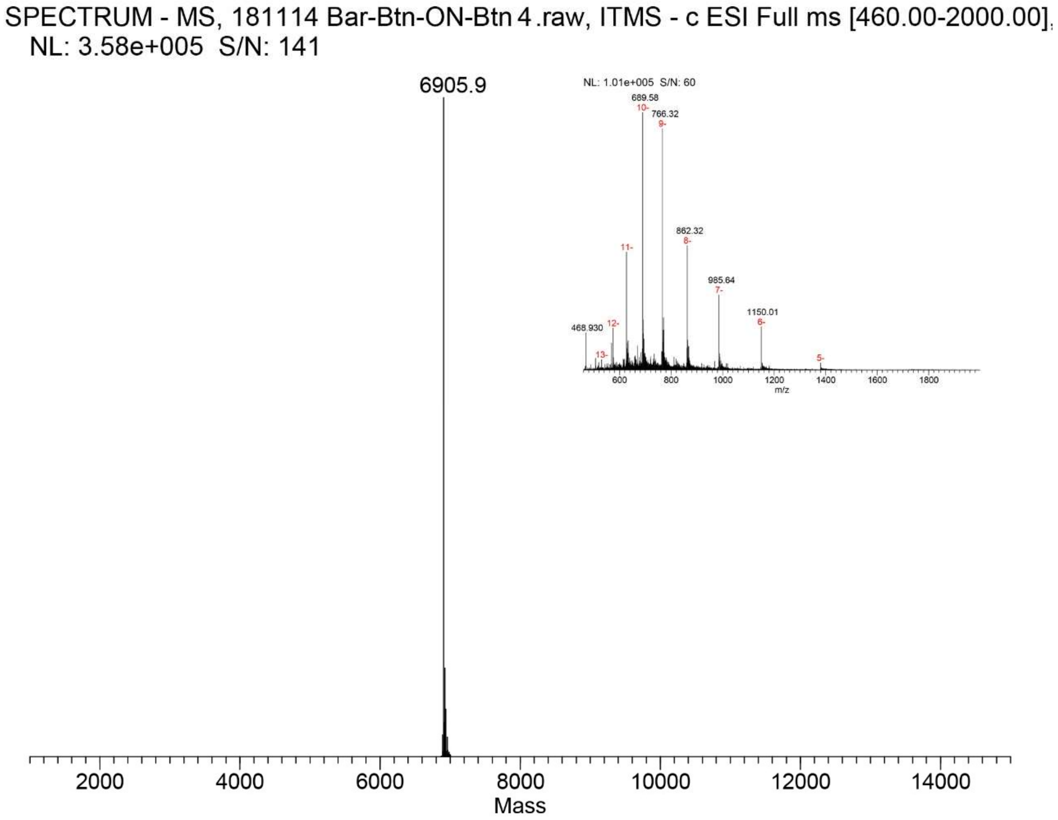
Mass spectrometry characterization of Bar-AO-1



Mass spectrometry characterization of Bar-AO-2



Mass spectrometry characterization of Bar-AO-3



Mass spectrometry characterization of Bar-AO-4

Chapter 5: Summary and Outlook

Exposure to endogenous and exogenous chemicals could lead to the formation of various DNA lesions. Decoding of DNA lesions in genome is of critical interest, enabling valuable insights into DNA damage repair and tolerance, mutagenesis process and their toxicological implications. However, due to the low abundance of DNA lesions, it is challenging to address the location and chemical information of DNA lesions in a genome.

The work presented in this thesis concerned developing novel sequencing methods for DNA lesions. The overarching strategy was the use of DNA probe to label damage site as a readable barcode. Based on this, two specific methods were developed for major nucleobase oxidation, 8-oxoG and major 2-deoxyribose oxidation, 5’-aldehyde terminus, respectively.

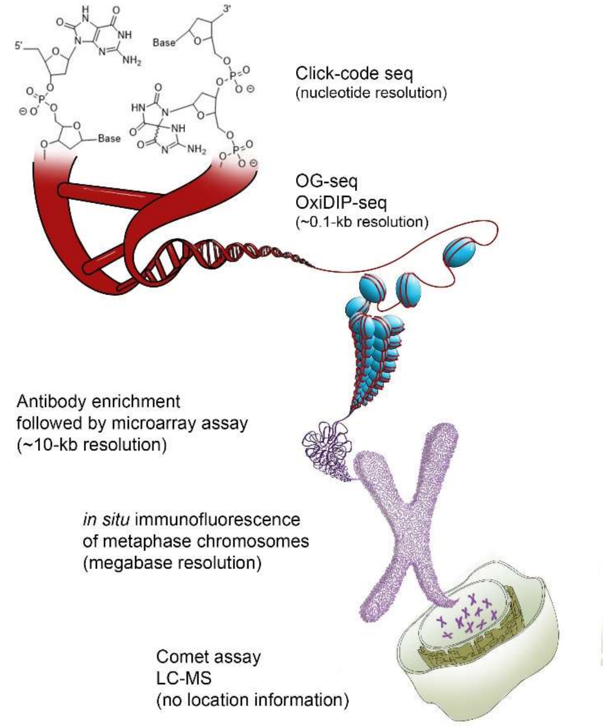
In **Chapter 2**, a glycosylase excision and click reaction based method was developed for 8-oxoG sequencing, named click-code-seq. The method was validated with oligonucleotide and dsDNA models and then applied to yeast genome as a proof of concept. In **Chapter 3**, click-code-seq was further expanded to human genome. Nucleotide-resolution genome-wide mapping of 8-oxoG in yeast and human genome were achieved, respectively. Both studies uncovered distinct patterns of oxidation sites, relating to chromatin architecture, histone modification, DNA-protein interactions and DNA damage response network. In particular, nucleotide-resolution of this method enabled the analysis of flanking sequence around 8-oxoG sites for the first time. In yeast genome, we observed that the first G in a 5'-GG-3' dimer is more easily oxidized due to its lower ionization potential. More interesting, in human genome, the 3-bases damage pattern showed strong correlation with several mutation signatures that were related with increasing oxidative stress or repair protein deficiency, including SBS 5, SBS 10a, SBS 18, SBS30, SBS 32 and SBS 36. This exciting result provided the first direct observation of the mutagenesis process from DNA damage.

Further research will be carried out in several major directions. First, embedded biases in this method is not well understood, including artifactual DNA oxidation during sample preparation, excision preference of glycosylase and triazole linkage bypass efficiency. Further improvements are required to address these biases and improve reliability and sensitivity of this method. Second, focusing on 8-oxoG, we expect applications to understand how the damage distribution is governed by dynamic processes of repair and genomic architecture maintenance, and relate these factors with mutation signatures. For example, by altering and measuring chromatin architecture, click-code-seq could be used to understand how DNA repair status and genome architecture impacts human DNA oxidation maps. By exposing to different chemical sources of oxidative stress *in vitro* and *in vivo*, click-code-seq could be used to investigate the unique impacts of chemical sources on genome-wide profiles of DNA damage as a key initiating event in mutagenesis. Moreover, by studying both damage sequencing and mutation sequencing of repair protein deficient cell lines, click-code-seq could be used to investigate the linkages between early DNA damage pattern and mutation signature raised afterward. Finally, this method could be easily adapted to analyze other DNA damage/modifications including abasic sites, deoxyuridine, ribonucleotides, 3-methyladenine, single strand breaks, and cyclobutane pyrimidine dimers.

In **Chapter 4**, an aminooxy functioned DNA probe was synthesized to label and detect 5'-aldehyde terminus. Preliminary data suggested that both 5'-aldehyde terminus and abasic site could be labelled successfully. Taq, Deep vent and Q5 DNA polymerases could selectively amplify labelled product from 5'-aldehyde terminus but not from abasic site. Among these, Q5 polymerase is able to detect 5'-aldehyde terminus at a frequency as low as 10^{-7} lesions/unmodified bases. However, there are more aldehyde modified nucleotides in genomic DNA may disturb the detection of 5'-aldehyde terminus, such as 5-formylcytosine (5-fc). Further validation with 5-fc and other common aldehyde resources are needed. Upon full validation, this method could be further applied to map 5'-aldehyde terminus at nucleotide-resolution in genomic DNA.

In conclusion, the work presented here represents a significant advance from quantification of total amount of DNA damage to locate DNA damage sites at nucleotide-resolution in a genome. With the nucleotide-resolution maps of yeast and human genome, we are able to take initial steps to gain a better understanding of DNA oxidation damage and repair. However, the opportunities offered by these two methods are more valuable than the questions we could answer now. We believe these two sequencing methods will offer exciting prospects for addressing the biological and toxicological impacts of DNA damage in a genome scale.

Appendix A: Impact of DNA oxidation on toxicology: from quantification to genomics



Reprinted with permission from

Junzhou Wu. Shana J. Sturla, Cynthia J. Burrows and Aaron M. Fleming. Impact of DNA Oxidation on Toxicology: From Quantification to Genomics, *Chem. Res. Toxicol*. 2019, 32, 3, 345-347

https://doi.org/10.1021/acs.chemrestox.9b00046

Copyright © 2019 American Chemical Society

Appendix A

Impact of DNA oxidation on toxicology: from quantification to genomics

Junzhou Wu,[†] Shana J. Sturla,[†] Cynthia J. Burrows,[‡] and Aaron M. Fleming[‡]*

[†]Department of Health Sciences and Technology, ETH Zürich, Schmelzbergstrasse 9, 8092 Zürich, Switzerland

[‡]Department of Chemistry, University of Utah, Salt Lake City, Utah 84112-0850 United States

Abstract: Understanding the toxicological implications of DNA oxidation arising from cellular oxidative stress depends on identifying DNA oxidation products, their location in the genome and their interaction with repair, replication and gene expression.

DNA is a target of cellular oxidation processes that lead to the formation of mutagenic DNA oxidation including 8-oxo-7,8-dihydro-2'-deoxyguanosine (OG). OG and other oxidized bases are efficiently removed by the base excision repair (BER) pathway wherein base removal is catalyzed by DNA glycosylases, e.g. 8-oxoguanine DNA glycosylase (OGG1) targets oxidized purines and Nth like DNA glycosylase 1 (NTHL1) targets oxidized pyrimidines. Unrepaired OG is prone to G>T transversion mutations. Three decades of research using the comet assay has given rapid access to determining levels of DNA damage leading to strand breaks, including glycosylase-induced strand breaks. However, the comet assay provides no information on the location of the lesion in the genome, which restrict the conclusions drawn from the data. Advancements in quantitative mass spectrometry have permitted more selective identification and quantification of specific lesions; however, like the comet assay, sequence information is lost during analysis. Furthermore, errors can be introduced by artifactual oxidation during sample workup, especially for OG. There are many genomic questions we cannot address with these sequence-agnostic methods. To what degree is DNA damage randomly dispersed in the genome? How do genomic features impact repair function? What is the interplay between DNA damage and repair efficiency in a genome leading to cancer-causing mutations? How does DNA damage drive gene expression and cellular proliferation?

Precisely where DNA oxidation persists in a genome is hypothesized to influence cellular fitness and disease development by impacting mutagenesis and gene transcription. A mutation signature has been extracted from the cells of colorectal cancer patients with a deficiency in mutY DNA glycosylase (MUTYH), an enzyme responsible for removal of A opposite OG. This signature is similar to COSMIC signature 18, which is dominated by G>T transversions, particularly in the GCA trinucleotide context.[1] In addition to mutagenesis, evidence supports that DNA oxidation facilitates activation of protective or beneficial genes in response to oxidative stress, against the conventional model of DNA damage.[2] The molecular mechanisms invoked for DNA oxidation-induced gene expression involve several potential pathways, including direct interactions of OGG1 with transcription factors (TFs) or chromatin remodelers, allosteric transition of G-quadruplex-forming sequences (G4) and signal transduction by the post-repair OGG1·OG complex. Given the extremely high biological and health relevance of oxidative stress and genome integrity, improved knowledge is needed to understand factors that influence the persistence of DNA oxidation in the genome and its relation to mutagenesis and gene expression. However, these events are exceedingly rare, and chemical damage cannot be read by standard polymerase-based sequencing.

Recent oxidation-targeted library preparation protocols combined with next generation sequencing techniques afford new opportunities for genome-wide mapping of DNA oxidation. One strategy developed by the Burrows laboratory harnesses the sensitivity of OG to hyperoxidation leading to covalent biotinylation of OG (OG-Seq).[3] When applied to a fragmented mammalian genome, the OG-containing oligomers could be selectively enriched and submitted to next-generation sequencing. Alternatively, Amente et al. enriched a fragmented mammalian genome with an OG-selective antibody to fish out the OG-containing strands for sequencing (OxiDIP-Seq).[4] Both strategies lead to sequencing data that identify the locations of OG at a resolution of the length of fragmentation (~0.1 kb), and both give similar results, regarding genomic regions enriched in OG. Indeed, OG occurs at a greater frequency in specific genomic elements. By analyzing overlapped regions of sequence reads, enriched regions of OG clearly emerge from the one-in-a-million frequency of the oxidized base. Mammalian gene promoters with potential G4 were found to harbor OG at a greater frequency. This single finding is important because oxidative modification of a promoter G4 at any location in the sequence can regulate transcription.[2] Thus, sequencing for OG at ~0.1 kb resolution can reveal answers to profound questions regarding how genes are regulated during oxidative stress, but insight on the relationship of this distribution with particular DNA sequences or signatures requires complementary single base resolution approaches.

A strategy to locate oxidized bases at single base nucleotide resolution is to use nature's own recognition system to identify OG, namely a base excision repair glycosylase such as OGG1 or formamidopyrimidine DNA glycosylase (Fpg) to create a single-nucleotide gap followed by insertion of a chemically modified base used for amplification or tagging of the position of oxidation.[5-6] By inserting a alkynylated nucleotide after oxidation damage excision, for example, a code oligonucleotide could be then incorporated *via* click chemistry at oxidation sites (Click-code-seq).[6] With this approach, the code-sequence serves as a tag for affinity enrichment, an adaptor for PCR amplification, and a marker of the damage locations that is identified by high throughput sequencing. By mapping of DNA oxidation of the *S. cerevisiae* genome with click-code-seq, millions of DNA oxidation sites were uncovered with distinct patterns related to transcription, chromatin architecture, and chemical oxidation potential. These nucleotide resolution data revealed that the first G in a 5'-GG-3' dimer is more easily oxidized on a genome level. Results further suggest that DNA oxidation formation appears to be influenced by local chemistry of DNA sequence context and repair is more governed by genomic features and protein interactions.

With the power of newly developed genome-wide approaches, we are able to take first steps to gain a better understanding of DNA oxidation damage and repair in a genome scale. Nevertheless, further improvements are required to avoid embedded biases, reduce artifactual DNA oxidation during sample preparation, and improve reliability and sensitivity of these methods. These data regarding DNA oxidation mapping is limited, and provide no insight on the expected dynamic and variable patterns of DNA oxidation in cells. Emerging evidence suggests that DNA repair activities vary greatly between species, considering different repair pathways, life styles and life spans. Moreover, DNA oxidation formation and repair greatly depend on the heterogeneous structure of a chromosome, consisting of protein-bound regions, open regulatory regions and actively transcribed genes. In addition, abnormal endogenous metabolism status during disease processing and exogenous chemical/stressor exposures will also alter DNA oxidation patterns. Besides DNA oxidation patterns, we expect applications to further expand to broaden biology and toxicology studies, for example, the genomic connections of DNA oxidation during cancer development, the relationships between DNA oxidation and mutation signatures, and the mechanisms of oxidative stress adaptation when pathogens and plants respond to environmental shifts in levels of oxidative stress. Considering all the knowledge gaps, the methods we have described here offer exciting prospects for addressing the toxicological impacts of DNA oxidation in a genome scale.

References

1. Viel, A.; Bruselles, A.; Meccia, E.; Fornasarig, M.; Quaia, M.; Canzonieri, V.; Policicchio, E.; Urso, E. D.; Agostini, M.; Genuardi, M.; Lucci-Cordisco, E.; Venesio, T.; Martayan, A.; Diodoro, M. G.; Sanchez-Mete, L.; Stigliano, V.; Mazzei, F.; Grasso, F.; Giuliani, A.; Baiocchi, M.; Maestro, R.; Giannini, G.; Tartaglia, M.; Alexandrov, L. B.; Bignami, M., A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **2017,** *20*, 39-49.

2. Fleming, A. M.; Zhu, J.; Ding, Y.; Burrows, C. J., 8-Oxo-7,8-dihydroguanine in the context of a promoter G-quadruplex is an on-off switch for transcription. *ACS Chem. Biol.* **2017,** *12*, 2417-2426.

3. Ding, Y.; Fleming, A. M.; Burrows, C. J., Sequencing the mouse genome for the oxidatively modified base 8-oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **2017,** *139*, 2569-2572.

4. Amente, S.; Di Palo, G.; Scala, G.; Castrignanò, T.; Gorini, F.; Cocozza, S.; Moresano, A.; Pucci, P.; Ma, B.; Stepanov, I.; Lania, L.; Pelicci, P. G.; Dellino, G. I.; Majello, B., Genome-wide mapping of 8-oxo-7,8-dihydro-2′-deoxyguanosine reveals accumulation of oxidatively-generated damage at DNA replication origins within transcribed long genes of mammalian cells. *Nucleic Acids Res.* **2019,** *47* (1), 221-236.

5. Riedl, J.; Ding, Y.; Fleming, A. M.; Burrows, C. J., Identification of DNA lesions using a third base pair for amplification and nanopore sequencing. *Nat. Commun.* **2015,** *6*, 8807.

6. Wu, J.; McKeague, M.; Sturla, S. J., Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq. *J. Am. Chem. Soc.* **2018,** *140* (31), 9783-9787.

Appendix A

Appendix B: Click-code-seq Protocol

Reagents

Aminoguanidine hydrochloride (Sigma-Aldrich, cat. no. 396494)

APE 1 (New England Biolabs, cat. no. M0282S)

Copper(II) sulfate pentahydrate (Sigma-Aldrich, cat. no. C8027)

Deoxynucleotides (dNTPs) solution mix (New England Biolabs, cat. no. N0447S)

Dideoxynucleotides (ddNTPs) (Jena bioscience, cat. no. NU-1019)

Dynabeads MyOne streptavidin C1 (Life Technologies, cat. no. 65001)

Ethanol (Merckmillipore, cat. no. 1.100983.1011)

Fpg (New England Biolabs, cat. no. M0240S)

Methyl sulfoxide (DMSO) (Sigma-Aldrich, cat. no. W387520)

Poly(ethylene glycol) (PEG-4000) (Sigma-Aldrich, cat. no. 81240)

Q5 high-fidelity DNA polymerase (New England Biolabs, cat. no. M0491S)

Sigmacote (Sigma, cat. no. SL2)

(+)-Sodium L-ascorbate (Sigma-Aldrich, cat. no. A4034)

T4 DNA ligase (New England Biolabs, cat. no. M0202S)

Therminator IX DNA polymerase (New England Biolabs, cat. no. M0557B)

Tris(3-hydroxypropyltriazolylmethyl)amine (THPTA) (Sigma-Aldrich, cat. no. 762342)

Tween 20 (Sigma-Aldrich, cat. no. P9416)

Vent (exo-) DNA polymerase (New England Biolabs, cat. no. M0257S)

Equipment

2200 TapeStation (Agilent Genomics)

8-strip PCR tubes (Bioconcept, cat. no. 3131-00)

LoBind tubes, 1.5 ml (Eppendorf, cat. no. 0030 108.116)

LoBind tubes, 0.5 ml (Eppendorf, cat. no. 0030 108.094)

MagRack 6 (Jena bioscience, cat. no. PP-229)

Microcentrifuge (Labnet)

Monarch PCR & DNA Cleanup Kit (NEB, cat.no. T1030L)

Nanodrop ND-1000 (Thermofisher Scientific)

ProNex Size-Selective Purification System (Promega, cat. no. NG2001)

Quantus Fluorometer (Promega, cat. no. E6150)

Rotor-Gene 6000 (Corbett Life Science)

S220 Focused-ultrasonicators (Covaris)

T3000 Thermocycler (Biometra)

Reagent setup

Bead-binding buffer (2x): 10.0 mM Tris-HCl (pH 7.5), 1.0 mM EDTA, 2.0 M NaCl

Bead-wash buffer (1x): 10.0 mM Tris-HCl (pH 7.5), 1.0 mM EDTA, 0.2 M NaCl

Double-stranded adapter (40 µM): Set up the following hybridization reaction mixture in a 1.5 mL tube. Combine 19 µl of TE buffer, 1 µl of 5 M NaCl, 40 µl of 100 µM oligonucleotide L-P7-3 and 40 µl of 100 µM oligonucleotide L-P7-3c. Incubate the reaction mixture in a thermal cycler for 10 s at 95 °C and slowly decrease the temperature at the rate of 0.1 °C per second until reaching 14 °C. Store the resulting solution - 20 °C.

Oligonucleotides

Name	Sequence	Conc.
Bar-N3-mix	N3-T GTAC AGATCGGAAGAGC GTCGTG - Biotin	250 µM (in total, 1:1:1:1)
	N3-T AGCT AGATCGGAAGAGC GTCGTG - Biotin	
	N3-T CATG AGATCGGAAGAGC GTCGTG- Biotin	
	N3-T TCGA AGATCGGAAGAGC GTCGTG - Biotin	
Ex-Bar-mix	CACGACGCTCTTCCGATCT GTAC AC	10 µM (in total, 1:1:1:1)
	CACGACGCTCTTCCGATCT AGCT AC	
	CACGACGCTCTTCCGATCT CATG AC	
	CACGACGCTCTTCCGATCT TCGA AC	
Pr-P7-23nt	GTG ACT GGA GTT CAG ACG TGT GC	10 µM
P5-Universal	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	10 µM
P701	CAAGCAGAAGACGGCATACGAGAT**CGTGAT**GTGACTGGAGTT CAGACGTGTGCTCTTCCGATCT	10 µM
P702	CAAGCAGAAGACGGCATACGAGAT**ACATCG**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10 µM
P703	CAAGCAGAAGACGGCATACGAGAT**GCCTAA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10 µM
P704	CAAGCAGAAGACGGCATACGAGAT**TGGTCA**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	10 µM

1. DNA shearing

Fragmented genomic DNA was obtained by shearing DNA in 130 µl of TE buffer with a Covaris S220 ultrasonicator using the following parameters: peak incident power 140 W, cycles/burst 200, duty factor 10% and time 100 s. DNA concentration and distribution were estimated using the Nanodrop 8000 and Agilent 2200 Tapestation with high sensitivity D1000 screen tape.

2. Abasic sites remove and free 3'-OH block

2.1 For each sample, prepare the following reaction mixture with a total volume of 50 µl in tubes. Mix by flicking the tubes with a finger and spin the tubes briefly in a microcentrifuge.

Reagent	Volume (µl) per sample	Final conc.
Water (to 50 µl)	43-x	
NEBuffer 2.1	5	1 x
Fragmented DNA (up to 5 µg)	x	
T4 PNK (10 U/µl)	1	0.2 U/µl
APE 1 (10 U/µl)	1	0.2 U/µl

2.2 Incubate the reactions in a thermal shaker for 0.5 h at 37 °C.

2.3 Add the following components to the reaction mixtures to obtain a final reaction volume of 60 µl. Mix by flicking the tubes with a finger and spin the tubes briefly in a microcentrifuge.

Reagent	Volume (µl) per sample	Final conc.
Step 2.2	50	
Water (to 60 µl)	2.8	
ddNTPs (2 mM)	6	200 µM
NEBuffer 2.1	1	1 x
Therminator IX (10 U/µl)	0.2	0.033 U/µl

2.4 Incubate the reactions in a thermal cycler with a heated lid for 10 min at 60 °C.

2.5 The products were purified using Monarch PCR & DNA Cleanup Kit (NEB) according to the manufacturer's instructions.

3. 8-oxoG excision and prop-dGTP incorporation

3.1 For each sample, prepare the following reaction mixture with a total volume of 50 µl in 1.5 ml tubes. Mix by flicking the tubes with a finger and spin the tubes briefly in a microcentrifuge.

Reagent	Volume (µl) per sample	Final conc.
Water (to 50 µl)	43 - X	
NEBuffer 2.1	5	1 x
DNA (Step 2.5)	X	
Fpg (8 U/µl)	1	0.16 U/µl
T4 PNK (10 U/µl)	1	0.2 U/µl

3.2 Incubate the reactions in a thermal cycler for 1 h at 37 °C.

3.3 Add the following components to the reaction mixtures to obtain a final reaction volume of 60 µl. Mix by flicking the tubes with a finger and spin the tubes briefly in a microcentrifuge.

Reagent	Volume (µl) per sample	Final conc.
Step 3.2	50	
Water (to 60 µl)	2.8	
Prop-dGTP (2 mM)	6	200 µM
NEBuffer 2.1	1	1 x
Therminator IX (10 U/µl)	0.2	0.033 U/µl

3.4 Incubate the reactions in a thermal cycler for 10 min at 60 °C.

3.5 The products were purified using Monarch PCR & DNA Cleanup Kit (NEB) according to the manufacturer's instructions and dried by speedvac concentrator.

4. Ligation of tag sequence via click reaction

4.1 Prepare the following stock solutions: 30 mM BTTAA ligand in water, 5 mM CuSO4 in water, 50 mM aminoguanidine in water, 25 mM sodium ascorbate in water, 1 M potassium phosphate buffer (pH 7), 250 µM azido-modified oligonucleotides (Bar-N3-mix) in 10 mM Tris buffer.

4.2 Add the following components to the tubes with dried alkyne-modified DNA fragments in the following order:

Reagent	Volume (µl) per sample	Final conc.
Bar-N3-mix(Biotin labelled)	8	100 µM
potassium phosphate buffer	2	100 mM
aminoguanidine	2	5 mM
sodium ascorbate	2	2.5 mM
DMSO	2	10%
Premixed CuSO4:BTTAA (1:6)	4	0.5 mM (Conc. of Cu^{2+})

4.3 Replace the oxygen in the tube with inert gas and close the tube, mix by inverting the tube several times, spin the tubes briefly in a microcentrifuge. Allow the reaction to proceed for 30 mins at 37 °C.

4.4 The products were purified using Monarch PCR & DNA Cleanup Kit (NEB) according to the manufacturer's instructions. The trace amount of adaptor in the mixture were removed by the second round purification with ProNex Size-Selective Purification System (2:1). DNA concentration was estimated using the Quantus fluorometer.

5. Immobilization of ligation products on beads

5.1 Resuspend the stock of Dynabeads MyOne Streptavidin C1 beads by vortexing. For each sample, transfer 5 µl of the bead suspension into a 0.5 mL tube (Sigmacoat pre-treated). Wash the beads twice with 200 µl of bead-binding buffer. Resuspend the beads in 50 µl bead-binding buffer.

5.2. Incubate the DNA (50 µl in TE buffer) from Step 4.4 for 2 min at 95 °C in a thermal shaker and quickly transfer the tubes into an ice bath. Let the reaction mixture cool down for at least 2 min. Spin the tubes briefly in a microcentrifuge and add the DNA to the bead suspensions prepared in Step 5.1.

5.3 Rotate the tubes for 30 min at room temperature.

5.4 Spin the tubes briefly in a microcentrifuge. Pellet the beads using a magnetic rack and discard the supernatant. Wash the beads twice with wash buffer.

6. Phosphorylation with T4 PNK

6.1 Prepare the following reaction mixture with a total volume of 49 µl per sample. Mix by flicking the tubes with a finger and spin the tubes briefly in a microcentrifuge.

Reagent	Volume (µl) per sample	Final conc.
Water (to 49 µl)	39	
T4 PNK reaction buffer (10×)	5	1 x
ATP (10 mM)	5	1 mM

6.2 Pellet the beads using a magnetic rack and discard the wash buffer. Add 49 µl of the reaction mixture from Step 6.1 to the pelleted beads and resuspend the beads by pipetting. Add 1 µl of T4 PNK (10 units, NEB). Mix the contents briefly by pipetting.

6.3 Incubate the reaction mixtures for 0.5 h at 37 °C.

6.4 Pellet the beads using a magnetic rack and discard the supernatant. Wash the beads twice with wash buffer.

7. Ligation of second adapter and library elution

7.1 Prepare the following reaction mixture with a total volume of 49 µl per sample. Mix by votexing and spin the tubes briefly in a microcentrifuge.

Reagent	Volume (µl) per sample	Final conc.
Water	36	
T4 DNA ligase buffer (10×)	5	1 x
PEG-4000 (50%)	5	5%
Tween 20 (1%)	0.5	0.01%
Double-stranded adapter (40 µM)	2.5	2 µM

7.2 Pellet the beads using a magnetic rack and discard the wash buffer. Add 49 µl of the reaction mixture from Step 7.1 to the pelleted beads and resuspend the beads by pipetting. Add 1 µl of T4 DNA ligase (400 U, NEB). Mix the contents briefly by pipetting.

7.3 Incubate the reaction mixtures for 2 h at room temperature. Keep the beads suspended during incubation by rotating.

7.4 Pellet the beads using a magnetic rack and discard the supernatant. Wash the beads twice with wash buffer and once with water. Add 20 µl of water to the pelleted beads, resuspend the beads by vortexing.

8. Triazole bypass by Vent exo-

8.1 Prepare the following PCR mix.

Reagent	Volume (µl) per sample	Final conc.
Water (to 50 µl)	17.5	
ThermoPol Buffer (10 x)	5	1 x
dNTPs (2 mM)	5	200 µM
Ex_Bar_mix (10 µM)	1	0.2 µM
Pr_P7_23nt (10 µM)	1	0.2 µM
Beads from 7.4	20	
Vent (exo-) DNA Polymerase	0.5	0.02 U/µl

8.2 Incubate the reactions in a thermal cycler with the following thermal profile. Initial denaturation should be carried out at 95 °C for 2 min. Follow this by 6 PCR cycles, involving denaturation for 20 s at 95 °C, annealing for 20 s at 59 °C and primer extension for 60 s at 72 °C, final extension for 5 min at 72 °C, and hold at 4 °C.

8.3 Thoroughly shake the ProNex Chemistry bottle to resuspend the beads. Add 100 µl of ProNex Chemistry into the PCR mix (a 2:1 (v/v) ratio of ProNex Chemistry to sample volume) and mix by pipetting 10 times. Incubate the sample at room temperature for 10 minutes. Place the sample on a magnetic stand for 2 minutes. Carefully remove and discard the supernatant. Leaving the sample on the magnetic stand, add 200µl of Wash Buffer to the sample and allow it to incubate for 60 seconds. Remove and discard the Wash Buffer. Repeat the wash step. Allow the sample to air-dry for 5 minutes or longer. Remove the sample from the magnetic stand. Add 30µl of Elution Buffer and resuspend the beads by pipetting. Incubate the samples at room temperature for 5 minutes to elute the DNA. Return the sample to the magnetic stand for 1 minute, then carefully transfer the supernatant containing the DNA to a clean tube.

9. Library amplification

9.1 Prepare the following PCR mix.

Reagent	Volume (µl) per sample	Final conc.
Water (to 100 µl)	x	
Q5 Reaction Buffer (5 x)	20	1 x
dNTPs (2 mM)	10	200 µM
P5-universal (10 µM)	5	0.5 µM
P7 index primer (10 µM)	5	0.5 µM
Library	x	
Q5 DNA Polymerase (2 U/µl)	1	0.02 U/µl

9.2 Incubate the reactions in a thermal cycler with the following thermal profile. Initial denaturation should be carried out at 98 °C for 30 s. Follow this by a selected number of PCR cycles, involving denaturation for 10 s at 98 °C, annealing for 20 s at 67 °C and primer extension for 30 s at 72 °C, final extension for 2 min at 72 °C, and hold at 4 °C. The optimal number of PCR cycles for each sample should be determined from the amplification plots obtained by qPCR.

9.3 Purify amplified libraries using the Monarch PCR purification kit or ProNex size-selective DNA purification system according to the manufacturer's instructions. Elute the DNA in 20 µl of TE buffer.

9.4 Determine the fragment size distributions and concentrations of the DNA libraries by running the Agilent 2200 Tapestation with high sensitivity D1000 screen tape.