Base-Resolution Analysis of Deoxyuridine at the Genome Scale Based on the Artificial Incorporation Modified Nucleobase

Yafen Wang,§ Xiong Zhang,§ Shaoqing Han,§ Wei Yang, Zonggui Chen, Fan Wu, Jizhou Liu, Xiaocheng Weng,* and Xiang Zhou*

Cite This: ACS Cent. Sci. 2021, 7, 973−979

ABSTRACT: Deamination of cytosine and dUMP misincorporation have been found to be capable of producing uracil in the genome. This study presents the AI-seq (artificial incorporation modified nucleobase for sequencing), a “base substitution”, which not only is capable of profiling uracil at single-nucleotide resolution and showing its centromeric enrichment but could also reveal that the identified uracil sites are derived from cytosine deamination. All the results indicate the potential biological significance of uracil as the epigenetic modification.

Epigenetic modifications in nucleic acids greatly increase the informational contents of the genome and play a crucial role in the gene regulation in organisms.1,2 The abnormal status of these chemical modifications has been known to cause a variety of diseases, such as cancer.3,4 Despite a plethora of technologies that have been developed toward various epigenetic modifications in recent years,5,6 significant challenges still remain, with the potential modifications in a near-dormant state, for example, largely due to the extremely low content and a lack of sensitive detection tools. Uracil, one of the canonical nucleobases present in RNA, has also been found frequently in DNA. In most mammals, uracil could be generated by the misincorporation of the dUMP during the DNA replication, spontaneous cytosine (C) deamination, or the cytosine deamination by the AID/APOBEC family proteins, which could lead to completely different biological behaviors.7,8 The misincorporation of uracil is a type of damage of the endogenous DNA resulting in genomic instability.9 Contrarily, the cytosine deamination by AID in specific regions plays vital roles in several cellular processes,10 and it is this dysregulation that is believed to cause diseases.10 As in the case of the dynamic regulation of other epigenetic modifications,11 uracil in the DNA could also be modulated and removed by several uracil DNA glycosylases (UDGs) to initiate the base excision repair (BER) pathway in mammalian cells.12 In view of the factors mentioned above, the detection of uracil could be worthy of attention. Presently, the uracil in the genome could easily be quantified by LC/MS/MS.13 Moreover, several approaches have been reported for the quantification and identification of the locations of uracil in DNA, like the semiquantitative dot-blot application,14 one-step quantitative PCR,15 differential DNA denaturation PCR,16,17 ligation-mediated PCR,18 Excision-DirSeq using excision repair enzymes,19 deletion mutation upon sequencing by UDG and APE120,21 and uracil excision-droplet digital PCR (Ex-ddPCR).22 Moreover, several different approaches were also designed for in situ detection of uracil–DNA in the cellular context using fluorescence microscopies.8,23,24 Nonetheless, the mapping of the distribution of uracil at genome-wide still remains a challenge besides impeding an in-depth view of global uracil events. dU-seq,23 UPD-seq,24 and U-DNA-Seq8 were also developed to obtain
DNA was subjected to UDG treatment to generate AP sites, which was further labeled by identifying DNA oligo could be labeled by N3 assay, and MALDI-TOF MS illustrated that the UDG-treated DNA oligo was modified by the biotin handle (Figure 2a and Figures S2–S5). In order to verify that the N3−C label-treated uracil had no sequence bias, several sequences were randomly designed containing dU. From the PAGE and MALDI-TOF-spectrum analysis, the results indicated that N3−C could efficiently label AP sites without sequence bias (Figures S6 and S7). The ability to produce the U-to-C transition at uracil sites was further verified by single-nucleotide primer extension assay, Sanger sequencing, as well as high-throughput sequencing in DNA oligos with various uracil sites (Figure 2b–d and Figures S8–S11). To evaluate the efficiency of T-to-C conversion after uracil was labeled with N3−C, DNA from U:A (thymine replacing uracils) and for Us from U:G (deaminated cytosines) sites was used for Illumina sequencing. By analyzing the proportion of T-to-C transitions in different samples, the conversion efficiency was calculated to be about 84–88% (Figure S12). The qPCR analysis confirmed that the labeled uracil had negligible effects on PCR polymerases (Figure S13). Endogenous 5fC and 5fU are also found to be present in genomic DNA.3,6,28,29 It is necessary to exclude the interference from these two modifications. Under the same conditions as the uracil marker, PAGE analysis illustrated that N3−C would not react with 5fC but could label 5fU (Figures S14a and S15). To verify that the reaction of 5fU with the N3−C would not be read as C in the PCR amplification, an 80 bp ODN containing two 5fU sites (80 bp ds ODN-5fU) was incubated with N3−C and subsequently applied in NGS. As indicated by bioinformatics analysis, the NGS data displayed negligible T-to-C conversion (Figure S14b,c) because N3−C only reacts with the 5fU but does not change its base pairing pattern during amplification. Hence, the reaction with 5fU...
The results indicated that UDG could not cleave 5fU (Figures S16 and S17). We also tested the in

Figure 2. Gel electrophoresis and Sanger sequencing analysis to validate that N<sub>1</sub>−C can label dU. (a) PAGE analysis of the reaction efficiency during different reaction times under optimized conditions. (b) Single-nucleotide primer extension assay with Deep Vent (exo-) DNA polymerase. Lanes 1 and 14, HEX-labeled primer marker (ODN2-Primer 1); lanes 2−5, ODN2-U; lanes 6−9, ODN2-AP; lanes 10−13, ODN2-AP labeled with N<sub>1</sub>−C. (c) Single-nucleotide prime extension assay with Deep Vent (exo-) DNA polymerase in the different fraction of dU/(T + dU)]. Lanes 1 and 14, HEX-labeled primer marker (ODN2-Primer 2). After treatment with UDG, the mixtures were incubated with 400 μM N<sub>1</sub>−C in 50 mM MES buffer (pH = 6.0) at 37 °C for 2 h. After purification, they were applied to a primer extension assay and subsequently analyzed with 20% denaturing PAGE. (d) Sanger sequencing analysis of uracil in the model ODN-containing uracil site (80 bp ds ODN-1U) before (top) and after (bottom) labeling with N<sub>1</sub>−C. The original uracil sites are marked by the surrounding red dotted lines.

does not affect the determination of the uracil sites in the genome. We also tested the influence of UDG on 5fU and found that UDG could not cleave 5fU (Figures S16 and S17). The results indicated that N<sub>1</sub>−C is capable of detecting uracil sites through base substitution and rational mutation.

Validation of the feasibility of the AI-seq in model DNA in hand, the strategy was applied to analyzing uracil in the whole genome. To exclude the original AP sites, the genomic DNA was first treated with APE1 to cleave the existing AP.<sup>20</sup> Using the base substitution strategy of AI-seq, 1064 uracil sites have been identified in the HEK293T cell at single-base resolution. The three main ways to generate uracil are dUMP random misincorporation, spontaneous C deamination, and the cytosine deamination by the AID/APOBEC family proteins. AI-seq can easily make a distinction between these kinds of origins through mutation patterns. As indicated in Figure 1c, the uracil sites from deamination of partial C sites in the input samples will be read as T in deep sequencing, resulting in the C/T mixed reads. However, in the output sample after treatment by AI-seq, the uracil sites were labeled with N<sub>1</sub>−C to generate artificial C sites. Hence, the proportion of C in the reading at these sites sees an increase in the output samples, determined as uracil sites. For dUMP misincorporation, in the input samples, in principle, the site containing uracil will be read as T, but labeling with N<sub>1</sub>−C will produce a mixed reading of T and C. Therefore, the sites with a T reading in the input sample but which converted to C/T in the mixed reading in the output sample were regarded as uracil (Figure 1c). We statistically analyzed the read ratio of cytosine and thymine in the identified uracil sites in the input and the output and found that the ratio of cytosine increased after N<sub>1</sub>−C labeling (Figure S18). Comparing the differentiation between the input sample and output sample, 67 uracil sites from uracil misincorporation and 997 from cytosine deamination were identified, respectively (Figure 3a). The uracil

Figure 3. Validation of some uracil sites by the reported method. (a) Number of uracil sites in the case of dUMP misincorporation and cytosine deamination. (b) Validation of uracil sites in genomic DNA of HEK293T cells by 3D-PCR. Agarose gel electrophoresis analysis of 3D-PCR products for the uracil site detected in AI-seq at the indicated denaturation temperature (T<sub>d</sub>) ranges. (c) Validation of uracil sites by excision repair enzymes. The HEK293T genomic DNA was treated in order with UDG, APE1, and T4 DNA ligase and subsequently amplified by site-specific PCR primers for Sanger sequencing. The red dotted line represents the position of uracil sites. The red triangle indicates the excision of uracil sites.
successfully identify the existence of uracil sites, which certify the reliability of Al-seq. It is noteworthy that Al-seq is capable of quantifying the modification fraction of uracil sites. The modification level of the several modified sites combined with $N_3-C$ treatment were measured and amplified by specific primers (Figure S21 and Table S2). The use of 3D-PCR, excision-repair enzymes, and Al-seq validated the presence of uracil in the genome. Compared with existing methods, Al-seq...
not only could detect uracil at single-base resolution but also
could assess the modification level and identify the origin of
dU.

A chromosome-level analysis indicated that the uracil sites
were significantly enriched on chromosomes 1 and 7 (Figure
S22). After considering the different length of each
chromosome, a rather higher distribution in chromosomes 1,
7, 16, and 18 was observed (Figure 4a). Among the different
genomic element groups, around 79.2% of the sites were
located in the intergenic regions, 19.3% in the intron, and 1.5%
in other regions, such as 3' UTR, transcriptional termination
sites (TTTs), and promoter (Figure 4b,c). The uracil sites were
further examined in the repetitive elements, and the results
revealed that uracil sites were highly enriched at satellite
repeats and, nevertheless, were depleted at long terminal
repeats (LTRs) (Figure 4d).

By mapping uracil sites to introns, the lengths of introns
containing uracil modifications were compared against all the
introns in dU-modified genes. The lengths of dU-containing
introns were significantly longer than all introns taken together
(\(P < 2.2 \times 10^{-16}\), Wilcoxon test; Figure 4e). Surprisingly, it was
observed that most of the uracil sites in introns are located in the
first few introns of the corresponding gene, especially the
first one (Figure 4f). Moreover, in order to exclude the
possibility that the number of introns on the genes is mostly
very low, the number of introns containing the uracil locus on
the gene was calculated. Most genes with high counts for uracil
sites contain intron numbers mainly ranging from 3 to 10
(Figure S23). Reports illustrated that the first intron is closely
related to transcriptional regulation.25 The presence of uracil in
the first intron on the gene may be one of the transcriptional
regulatory elements affecting physiological processes, which
deserves further attention.

Recent findings by Shu et al. revealed that uracil peaks are
enriched in the human centromeric DNA.23 Centromeres are
highly repetitive segments presenting serious problems in NGS
data analysis, and these are blacklisted in many studies. To
avoid these problems, a blacklist region was established
according to the reported method.35 After filtering out these
controversial regions, the proportion of uracil in the
centromere region accounted for 31.4% of all the uracil sites in the
gene of the HEK293T cell line (Figure 5a–c and Figure
S24), and in contrast, the length of the total centromeres only
comprised about 2% of the whole human genome (the
centromeres data is downloaded from http://genome.ucsc.
edu/). The relative enrichment of uracil sites for the
centromere of each chromosome was analyzed. Uracil was
enriched on centromeres of most chromosomes, excluding the
chromosomes 13, 14, 15, 21, X, and Y (Figure 5d and Figure
S25). It is worth mentioning that the in situ \(R_{15,26}\) methods may
be helpful for the identification of uracil in some complex
fragments. It was reported that the uracil sites were colocalized
with CENP-A binding regions, which is essential for the
determination of centromere identity and the key to
centromere assembly.31 From the results of AI-seq, the
correlation of the uracil sites in the centromere with CENP-
A binding targets could be observed closely (Figure S26).

In summary, we have developed a novel technology, named
AI-seq, based on the concept of “base substitution” for
the whole-genome localization of uracil at single-base resolution.
Consistent with the dU-seq, our results confirmed that uracil
was enriched in the centromeric DNA and colocalized with
CENP-A binding regions.25 AI-seq also showed the advantages
to distinguish the origin of uracil based on mutation pattern,
which cannot be achieved with the existing methods due to the
low resolution. The base substitution strategy can produce site-
specific mutations, which is far more convenient for
bioinformatics analysis than methods based on termination
or random mutation in the process of PCR. We believe that
our base substitution strategy has broad applications for
profiling high-resolution maps of other epigenetic modificatios,
such as N’-methylguanosine (m'G) (Figure S27). The advantages of “base substitution” are anticipated to be a new
and urgently useful strategy to promote further research on
nucleic acid epigenetics.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge at
https://pubs.acs.org/doi/10.1021/acscentsci.0c01504.

Materials and methods and additional data and figures
including structures, gel electrophoresis, dot blot, HPLC
analysis, pull-down assay, DNA MALDI-TOF mass
spectra, LC/MS/MS analysis, PAGE analyses, single-
nucleotide primer extension assay, Sanger sequencing
analyses, Illumina sequencing, qPCR analysis, and uracil
site distributions (PDF)

**AUTHOR INFORMATION**

**Corresponding Authors**

Xiaocheng Weng — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China; orcid.org/0000-0002-6065-8495;
Email: xcweng@whu.edu.cn

Xiang Zhou — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China; orcid.org/0000-0002-1829-9368;
Email: xzhou@whu.edu.cn

**Authors**

Yafen Wang — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China

Xiong Zhang — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China

Shaoqing Han — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China

Wei Yang — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China

Zonggui Chen — State Key Laboratory of Freshwater Ecology
and Biotecnology, Institute of Hydrobiology, Innovation
Academy for Seed Design, Chinese Academy of Sciences,
Wuhan 430072, China

Fan Wu — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China

Jizhou Liu — College of Chemistry and Molecular Sciences,
Key Laboratory of Biomedical Polymers of Ministry of Education,
Wuhan University, Wuhan 430072, Hubei, China

Complete contact information is available at:
Author Contributions
Y. Wang, X. Zhang, and S. Han contributed equally.

Notes
The authors declare no competing financial interest.

Data Availability: Sequencing data have been deposited into the Gene Expression Omnibus (GEO). The accession number is GSE142071.

ACKNOWLEDGMENTS
The authors are thankful for all of their schoolfellows who have focused so much of their attention on discussions of the authors work. The numerical calculations in this Article were performed on the supercomputing system in the Supercomputing Center of Wuhan University. We thank Dr. Haifang Li (Analysis Center, Tsinghua University) who provided DNA MALDI-TOF test instructions. This work was supported financially by the National Natural Science Foundation of China (91753201 and 21721005 to X. Zhou; 21822704, 91940304, and 21778040 to Y. Weng; 21907077 to Y. Wang), the Postdoctoral Innovative Talent Support Program of China (BX20180228 to Y. Wang), and China Postdoctoral Science Foundation (2019M652691 to Y. Wang). S. Frank MA contributed to the editing.

REFERENCES
(1) Frye, M.; Harada, B. T.; Behm, M.; He, C. RNA modifications modulate gene expression during development. Science 2018, 361 (6409), 1346–1349.
(2) Chen, Y.; Hong, T.; Wang, S.; Mo, J.; Tian, T.; Zhou, X. Epigenetic modification of nucleic acids: from basic studies to medical applications. Chem. Soc. Rev. 2017, 46 (10), 2844–2872.
(3) Flavahan, W. A.; Gaskell, E.; Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. Science 2017, 357 (6348), eaal2380.
(4) Delaunay, S.; Frye, M. RNA modifications regulating cell fate in cancer. Nat. Cell Biol. 2019, 21 (5), S52–S59.
(5) Berney, M.; McGouran, J. F. Methods for detection of cytosine and thymine modifications in DNA. Nature Reviews Chemistry 2018, 2 (11), 332–348.
(6) Wang, Y.; Zhang, X.; Zou, G.; Peng, S.; Liu, C.; Zhou, X. Detection and Application of 5-Formylcytosine and 5-Formyluracil in DNA. Acc. Chem. Res. 2019, 52 (4), 1016–1024.
(7) Chon, J.; Field, M. S.; Stover, P. J. Deoxycytosine in DNA and disease: Genomic signal or managed situation? DNA Repair 2019, 77, 36–44.
(8) Palinkas, H. L.; Bekesi, A.; Rona, G.; Pongor, L.; Papp, G.; Tihanyi, G.; Holub, E.; Poti, A.; Gemma, C.; Ali, S.; Morten, M. J.; Rothenberg, E.; Pagano, M.; Szuts, D.; Gyorffy, B.; Vertessy, B. G. Genome-wide alterations of uracil distribution patterns in human DNA upon chemotherapeutic treatments. eLife 2020, 9, e60498.
(9) Methot, S. P.; Di Noia, J. M. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. In Adv. Immunol.; Alt, F. W., Ed; Academic Press: 2017, Chapter 2, Vol. 133, pp 37–87.
(10) Sirivardena, S. U.; Chen, K.; Bhagwat, A. S. Functions and Malfunctions of Mammalian DNA-Cytosine Deaminases. Chem. Rev. 2016, 116 (20), 12688–12710.
(11) Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat. Rev. Genet. 2018, 19 (2), 81–92.
(12) Healing, E.; Charlier, C. F.; Meira, L. B.; Elliott, R. M. A panel of colorimetric assays to measure enzymatic activity in the base excision DNA repair pathway. Nucleic Acids Res. 2019, 47, No. e61.
(13) Galashevskaya, A.; Sarro, A.; Vagba, C. B.; Aas, P. A.; Hagen, L.; Slupphaug, G.; Krokan, H. E. A robust, sensitive assay for genomic uracil detection by LC/MS/MS reveals lower levels than previously reported. DNA Repair 2013, 12 (9), 699–706.
(14) Rona, G.; Scheer, I.; Nagy, K.; Palinkas, H. L.; Tihanyi, G.; Boros, M.; Bekesi, A.; Vertessy, B. G. Detection of uracil within DNA using a sensitive labeling method for in vitro and cellular applications. Nucleic Acids Res. 2016, 44 (3), No. e28.
(15) Horváth, A.; Vertessy, B. G. A one-step method for quantitative determination of uracil in DNA by real-time PCR. Nucleic Acids Res. 2010, 38 (21), No. e196.
(16) Suspen, R.; Henry, M.; Guillot, S.; Wain-Hobson, S.; Vartanian, J. P. Recovery of APOBEC3-edited human immunodeficiency virus G-A hypermutants by differential DNA denaturation PCR. J. Gen. Virol. 2005, 86 (1), 125–129.
(17) Stenglein, M. D.; Burns, M. B.; Li, M.; Lengyl, J.; Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. Nat. Struct. Mol. Biol. 2010, 17 (2), 222–229.
(18) Maul, R. W.; Saribasak, H.; Martomo, S. A.; McClure, R. L.; Yang, W.; Vaisman, A.; Gramlich, H. S.; Schatz, D. G.; Woodgate, R.; Wilson, D. M.; 3rd; Gearhart, P. J. Uracil residues dependent on the deaminase AID in immunoglobulin gene variable and switch regions. Nat. Immunol. 2011, 12 (1), 70–76.
(19) 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.
(20) 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.
(21) Riedl, J.; Fleming, A. M.; Burrows, C. J. Sequencing of DNA Lesions Facilitated by Site-Specific Excision via Base Excision Repair DNA Glycosylases Yielding Ligated Gaps. J. Am. Chem. Soc. 2016, 138 (2), 491–494.
(22) Hansen, E. C.; Ransom, M.; Hesselberth, J. R.; Hosmane, N. N.; Capoferri, A. A.; Bruner, K. M.; Pollack, R. A.; Zhang, H.; Drummond, M. B.; Siliciano, J. M.; Siliciano, R. J.; Stivers, J. T. Diverse fates of uracilized HIV-1 DNA during infection of myeloid lineage cells. eLife 2016, 5, e18447.
(23) Shu, X.; Liu, M.; Lu, Z.; Zhu, C.; Meng, H.; Huang; S.; Zhang, X.; Yi, C. Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA. Nat. Chem. Biol. 2018, 14, 680–687.
(24) Sakhtemani, R.; Senevirathne, V.; Stewart, J.; Perera, M. L. W.; Pique-Regi, R.; Lawrence, M. S.; Bhagwat, A. S. Genome-wide mapping of regions preferentially targeted by the human DNA-cytosine deaminase APOBEC3A using uracil-DNA pulldown and sequencing. J. Biol. Chem. 2019, 294 (41), 15037–15051.
(25) Datta, M.; Aroli, S.; Karmakar, K.; Dutta, S.; Chakravortty, D.; Varshney, U. Development of mCherry tagged UdgX as a highly sensitive molecular probe for specific detection of uracils in DNA. Biochem. Biophys. Res. Commun. 2019, 518 (1), 38–43.
(26) Stewart, J. A.; Schauer, G.; Bhagwat, A. S. Visualization of uracils created by APOBEC3A using UdgX shows colocalization with RPA at stalled replication forks. Nucleic Acids Res. 2020, 48 (20), No. e118.
(27) Liu, Z. J.; Martinez Cuesta, S.; van Delft, P.; Balasubramanian, S. Sequencing absic sites in DNA at single-nucleotide resolution. Nat. Chem. 2019, 11 (7), 629–637.
(28) Zhu, C.; Gao, Y.; Guo, H.; Xia, B.; Song, J.; Wu, X.; Zeng, H.; Kee, K.; Tang, F.; Yi, C. Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. Cell Stem Cell 2017, 20 (5), 720–731.
(29) Hofer, A.; Liu, Z. J.; Balasubramanian, S. Detection, Structure and Function of Modified DNA Bases. J. Am. Chem. Soc. 2019, 141 (16), 6420–6429.
(30) Bornstein, P.; McKay, J.; Morishima, J. K.; Devarayalu, S.; Gélinas, R. E. Regulatory elements in the first intron contribute to transcriptional control of the human alpha 1(I) collagen gene. Proc. Natl. Acad. Sci. U. S. A. 1987, 84 (24), 8869–8873.
(31) De Rop, V.; Padeganeh, A.; Maddox, P. S. CENP-A: the key player behind centromere identity, propagation, and kinetochore assembly. Chromosoma 2012, 121 (6), 527–538.
Amemiya, H. M.; Kundaje, A.; Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **2019**, *9* (1), 9354.