ARTICLES
https://doi.org/10.1038/s41589-018-0065-9
nature chemical biology

Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA

Xiaoting Shu[1,2,4], Menghao Liu[1,2,4], Zhike Lu[1,4], Chenxu Zhu[1], Haowei Meng[1], Sihao Huang[1], Xiaoxue Zhang[2] and Chengqi Yi[1,3]*

Uracil in DNA can be generated by cytosine deamination or dUMP misincorporation; however, its distribution in the human genome is poorly understood. Here we present a selective labeling and pull-down technology for genome-wide uracil profiling and identify thousands of uracil peaks in three different human cell lines. Surprisingly, uracil is highly enriched at the centromere of the human genome. Using mass spectrometry, we demonstrate that human centromeric DNA contains a higher level of uracil. We also directly verify the presence of uracil within two centromeric uracil peaks on chromosomes 6 and 11. Moreover, centromeric uracil is preferentially localized within the binding regions of the centromere-specific histone CENP-A and can be excised by human uracil-DNA glycosylase UNG. Collectively, our approaches allow comprehensive analysis of uracil in the human genome and provide robust tools for mapping and future functional studies of uracil in DNA.

Chemical modifications to deoxynucleotides have profound influences on various cellular processes[1–3]. They can be installed by endogenous modification machineries and hence play critical roles in genome function (for example, 5-methylcytosine and N^6-methyladenine)[3,4]; alternatively, they can be generated by exogenous factors and therefore detrimental to the cells (for example, pyrimidine dimers and nucleobase oxidation)[5]. However, a global picture of these modifications in the genome is often missing because of the lack of sensitive and genome-wide detection methods[6].

Uracil in DNA is special in that it can be beneficial or harmful to the cells depending on the biological context[7,8]. Uracil can be generated by the AID/APOBEC family proteins, and it is recognized as a key intermediate in diverse cellular processes including somatic hypermutation and class switch recombination in B cells[9,10], intrinsic immunity against viral infection[11] and inhibition of retrotransposition of endogenous retroelements[12]. Dysregulation of the AID/APOBEC family deaminases has been shown to correlate with various types of cancers[13,14]. In addition, uracil in DNA can result from spontaneous deamination of cytosine and misincorporation of dUMP[7,8]. Though cytosine deamination creates U:G mispairs, dUMP misincorporation during replication results in U:A pairs.

Uracil in DNA can be removed by at least four different DNA glycosylases in human cells, which initiate base excision repair (BER) by cleaving the *N*-glycosidic bond connecting the uracil base and the deoxyribose[15,16]. Uracil-DNA *N*-glycosylase (UNG), which is encoded by the *UNG* gene, is considered the major uracil DNA glycosylase: the nuclear UNG2 efficiently removes misincorporated dUMP in replication foci, whereas the mitochondrial UNG1 is the only uracil DNA glycosylase in mitochondria[17,18]. UNG deficiency in primary mouse hematopoietic cells leads to abnormal telomere lengthening, indicating a crucial role for UNG-initiated BER in the maintenance of telomere integrity[19]. Interestingly, uracil-containing DNA is also reported to be involved in development and metamorphosis of *Drosophila melanogaster*, which accumulates high levels of uracil in genomic DNA as a result of a lack of *UNG* gene[20]. Single-strand selective monofunctional uracil DNA glycosylase 1 (SMUG1) has a broader substrate specificity than UNG and may serve as an efficient backup for UNG in repair of U:G mismatches[21]. Thymine DNA glycosylase (TDG) and methyl-CpG binding domain protein 4 (MBD4) may have specialized roles in removing mismatched uracil in double-stranded DNA[22,23].

To explore the locations of uracil in DNA, various methods have been developed. Differential DNA denaturation PCR (termed '3D-PCR') detects uracil in the context of U:G mismatches on the basis of the differential denaturation temperatures of PCR amplicons[24,25]. Ligation-mediated PCR captures the presence of short DNA fragments that originate from uracil-containing DNA and are excised at the sites of uracil by UNG and APE1 (ref. [26]). A more recent study used T4 DNA ligase to seal the gap generated by uracil-DNA glycosylase (UDG) and APE1 and detected uracil as a deletion mutation upon sequencing[27]. In addition, in situ detection of uracil in DNA was also achieved with catalytically inactive UNG sensor fusion proteins in $UNG^{-/-}$ MEF cells[28]. Recently, genome-wide uracil mapping technology has also been reported. In 'excision-seq', UDG and endonuclease IV were used in combination to create double-strand breaks at uracil-rich DNA regions; the resulting small DNA fragments were then subjected to high-throughput sequencing. Excision-seq revealed significant variation in uracil content in *Escherichia coli* and budding yeast[29]. Moreover, quantitative methods for targeted uracil detection have also been reported[30,31]. A real-time-PCR-based method, using the *Cq* shift between uracil-containing and uracil-free DNAs when amplified with *Pfu* or mutant *Pfu* DNA polymerases, measured uracil content within selected genomic segments in *E. coli* and MEF cells[30]. Ex-ddPCR (uracil excision-droplet digital PCR) exploited the amplification difference between UNG-digested and mock-digested DNAs to calculate the percentage of uracil-containing DNA and identified abundant uracil levels across the HIV genome during infection of monocyte-derived macrophages containing high cellular dUTP levels[31]. However, no

[1]State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. [2]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. [3]Department of Chemical Biology and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. [4]These authors contributed equally: Xiaoting Shu, Menghao Liu, Zhike Lu. *e-mail: chengqi.yi@pku.edu.cn

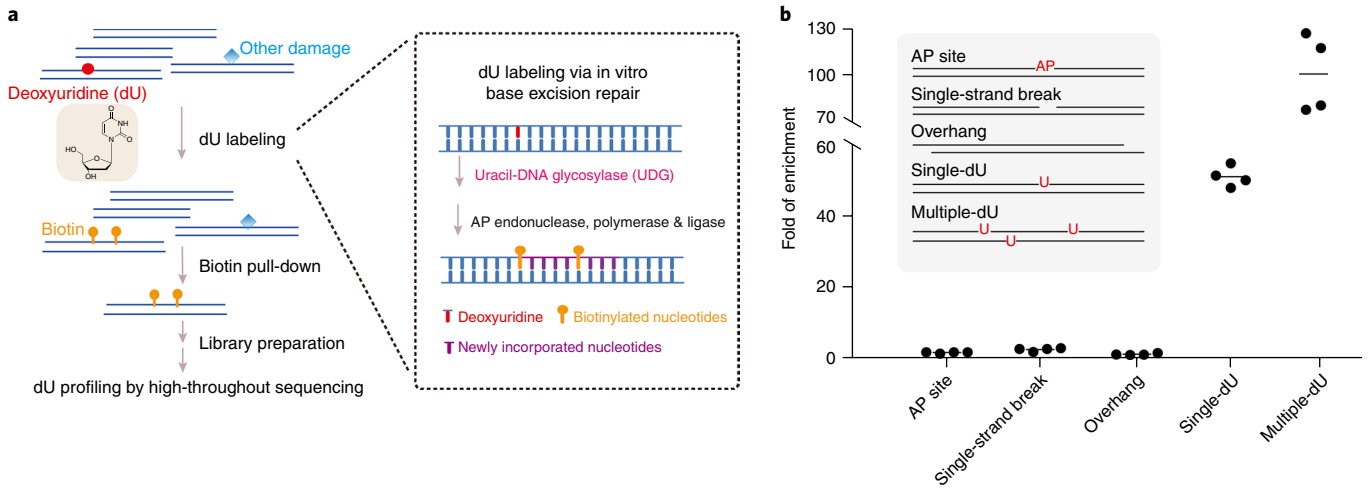

Fig. 1 | dU-seq is a selective labeling and pull-down technology for genome-wide uracil profiling. a, dU-seq specifically labels and enriches deoxyuridine via an in vitro base excision repair (BER) reaction. DNA fragments are subjected to an in vitro BER reaction, in which a uracil-DNA glycosylase (UDG), an AP endonuclease, a DNA polymerase and a DNA ligase are present. Biotin-dUTP is used in the in vitro BER reaction (replacing dTTP) and is incorporated into the uracil-containing DNA for pull-down and high-throughput sequencing. **b**, dU-seq efficiently enriches uracil-containing model sequences. Values represent fold of enrichment over the input, normalized to a reference sequence consisting of canonical bases. Each dot represents an independent experiment; the mean values are indicated with black horizontal lines ($n = 4$). Detailed sequence information can be found in Supplementary Table.

genome-wide uracil detection data has been reported for mammalian genomes so far[6]. Not only are the mammalian genomes much larger in size, but the uracil content is also very low (lower than 10 deoxyuridines per 10^6 nucleotides as measured by mass spectrometry)[32,33]. Thus, a highly sensitive method is required for genome-wide detection of uracil in the mammalian genome.

Here we present 'dU-seq', a genome-wide method to detect uracil in the entire human genome. dU-seq utilizes an in vitro BER reaction to specifically label uracil in the genome and allows enrichment of uracil-containing DNA before high-throughput sequencing. dU-seq identifies thousands of dU peaks in three different human cell lines and shows that uracil is highly enriched in the centromere regions of the human genome. Within the centromeric DNA, uracil preferentially localizes in the binding regions of CENP-A, a centromere-specific histone H3 variant. Lastly, we show that centromeric uracil can be excised by human uracil-DNA glycosylase UNG.

Results

Biotin labeling of uracil-containing DNA via in vitro BER. Previous uracil-detection methods rely on the conversion of a deoxyuridine into a single-stranded break, which is then captured and used as the sequencing readout[26,27,29]. In contrast to these methods, dU-seq is based on biotin labeling of uracil-containing DNA. Biotin incorporation is achieved via an in vitro BER reaction in which a UDG, an AP endonuclease, a DNA polymerase and a DNA ligase were added into one test tube to mimic the BER reaction in the cellular context. Instead of using regular dNTPs for DNA synthesis, we replaced one dNTP with a biotinylated dNTP during the in vitro BER reaction (Methods) so that uracil-containing DNA can be labeled with biotin (Fig. 1a).

To ensure the efficient incorporation of biotin, we carefully considered the choice of repair proteins in our in vitro BER reaction. First, to reconstitute the entire repair process under in vitro conditions, we selected four repair proteins that are compatible with each other, meaning that the repair product from the previous step is a preferred substrate of the next enzyme[16]. Second, we made sure that the desired polymerase possessed a double-strand specific 5′→3′ exonuclease activity but lack a 3′→5′ exonuclease activity. The 5′→3′ exonuclease activity guarantees that the deoxyribose 5′-phosphate created by the AP endonuclease can be removed, whereas the absence of the 3′→5′ exonuclease activity prevents nonspecific biotin labeling at the ends of DNA fragments. After testing different conditions of repair enzymes (Supplementary Fig. 1), we found that UDG and endonuclease IV from *E. coli*, DNA polymerase from *Bacillus stearothermophilus* (*Bst*), and *Taq* ligase are very compatible with each other and can efficiently incorporate the biotinylated nucleotide into the uracil-containing model DNA sequences.

dU-seq specifically enriches uracil-containing DNA. The specificity of dU-seq is achieved through multiple precautionary measures (Supplementary Fig. 2). First, we used the *E. coli* UDG because it is highly specific and removes uracil from both U:G and U:A pairs[34]. Hence, dU-seq labels uracil in dsDNA in an unbiased manner, without cross-reactivity to other types of damage in the genome. Second, we used a commercial enzymatic digestion method instead of sonication, which has been shown to introduce artificial DNA damage during the fragmentation process[35,36]. Because the DNA overhangs generated during the fragmentation step could also be mistakenly filled in with biotinylated dNTP by DNA polymerase, we used an 'end repair' step to convert overhangs into blunt DNA ends (Supplementary Fig. 1b). Third, because apurinic/apyrimidinic sites (AP sites) and single-stranded breaks, which are intermediates in our in vitro BER reaction, are also naturally occurring DNA damage in the genome, we added a 'damage repair' step before uracil labeling to eliminate these lesions (Supplementary Fig. 2). Indeed, we found that the damage repair step is necessary and sufficient to remove the AP sites, which would otherwise significantly interfere with uracil labeling (Supplementary Fig. 1c). Collectively, these approaches ensure that only uracil, but not other types of DNA lesions in the genome, is specifically labeled by dU-seq.

Multiple biotins can be incorporated during dU-seq even for DNA with one single uracil site. This is due to the so-called 'nick translation' activity of the *Bst* DNA polymerase (Supplementary Fig. 3a). We found that under our experimental conditions, the *Bst* DNA polymerase consistently replaces ~10 nucleotides 3′ to the damaged site and incorporates one biotin-dUTP when it encounters an adenosine in the template strand (Supplementary Fig. 3b). Therefore, the nick-translation process also enables uracil labeling in the context of both U:A and U:G pairs (in the case of U:G, as long as there is at least one adenosine within the range of nick translation

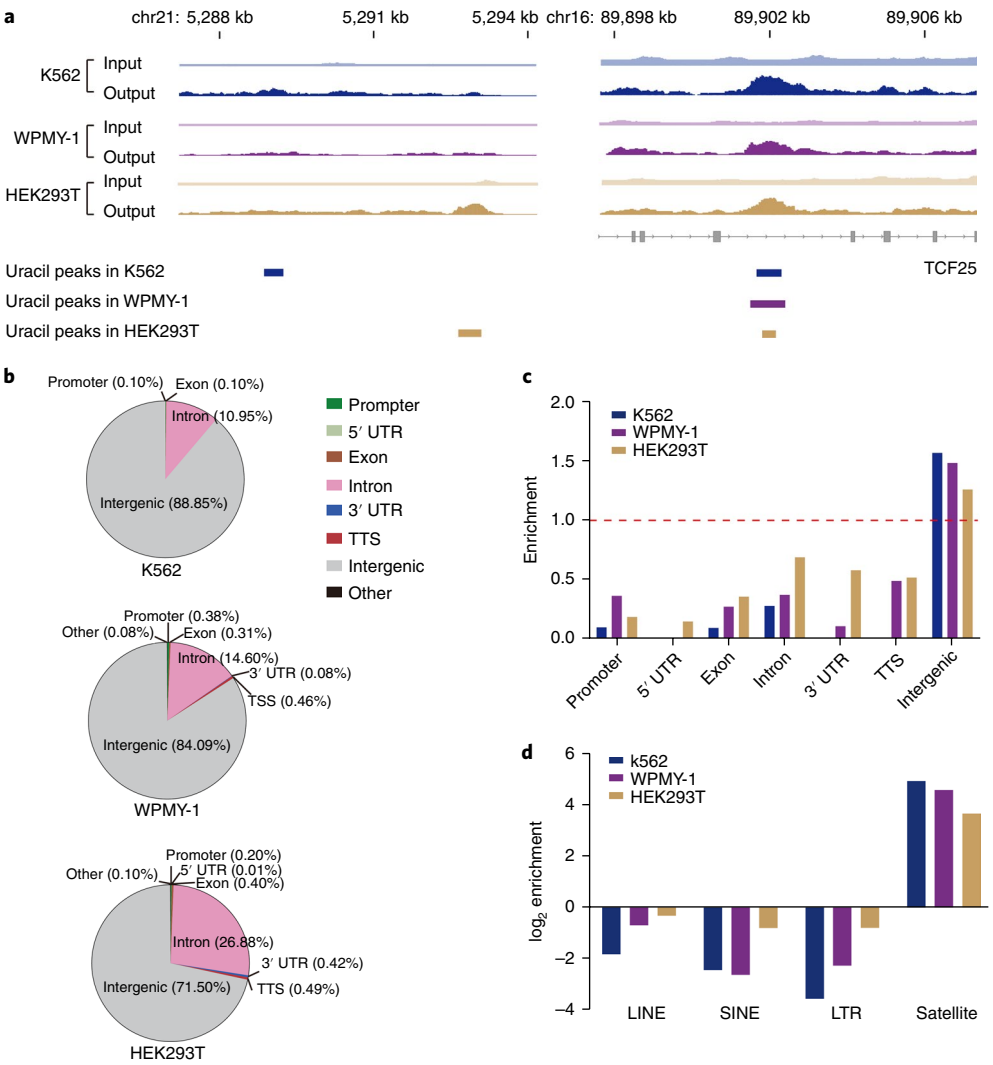

Fig. 2 | dU-seq reveals the genome-wide profile of uracil in K562, WPMY-1 and HEK293T cells, respectively. a, IGV (Integrative Genomics Viewer) views of representative uracil peaks. Similar results were observed within two independent replicates. **b**, Overall distribution of uracil peaks in the human genome. **c**, Relative enrichment of uracil peaks in different genomic elements. The dashed red line represents enrichment fold = 1.0. TSS, transcription start site; TTS, transcription termination site. **d**, Relative enrichment of uracil peaks in the major classes of repetitive elements. For all the three cell lines, combined data sets of two replicates (cell culture) were analyzed and shown here; the patterns of individual replicates are the same.

of the template strand; Supplementary Fig. 4). With our optimized dU-seq condition, a model DNA sequence containing one single deoxyuridine can be enriched by ~50-fold, and a sequence with multiple uracils can be enriched by ~100-fold (Fig. 1b and Supplementary Fig. 5). Importantly, no enrichment was observed for sequences containing an AP sites, single-strand breaks or overhangs at DNA ends (Fig. 1b and Supplementary Fig. 5). Thus, dU-seq integrates both specific dU labeling and efficient biotin pull-down to enrich the uracil-containing DNA for detection.

dU-seq identifies genome-wide uracil peaks in human cells. We next applied dU-seq to the genomic DNA of three different human cell lines (Supplementary Fig. 6). We identified 968, 1,301 and 8,186 uracil peaks for K562, WPMY-1 and HEK293T cells, respectively. Whole-genome views of uracil peaks and two representative uracil peaks are shown in Supplementary Fig. 6 and Fig. 2a, respectively. Within different functional regions, most of the uracil peaks were located and enriched in the intergenic regions of the human genome, whereas uracil peaks were mostly depleted from the gene body region (Fig. 2b,c). Because a large proportion of intergenic regions are repetitive elements, we then examined whether deoxyuridine is enriched at specific types of repeats. We found that uracil is highly enriched at simple repeats (such as satellite repeats) but is depleted at transposable elements including long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs) and long-terminal repeats (LTRs) (Fig. 2d).

Uracil is enriched in the human centromeric DNA. We next analyzed the distribution pattern of uracil peaks along each chromosome. Surprisingly, we found that uracil peaks are strongly concentrated in the centromere regions (Fig. 3a–c): approximately 30% of uracil peaks are located at the centromere in HEK293T cells, whereas more than 50% of uracil peaks are located at the centromere in K562 cells and WPMY-1 cells (Fig. 3b) despite of the fact that centromeres comprise only ~3% of the human genome (according to the GRCh38/hg38 assembly). We then calculated the relative enrichment of uracil peaks for the centromere of each chromosome. Almost all of the centromeres were highly enriched with uracil peaks, and such enrichment pattern was observed for all three different cell lines used in the study (Fig. 3d). We also performed

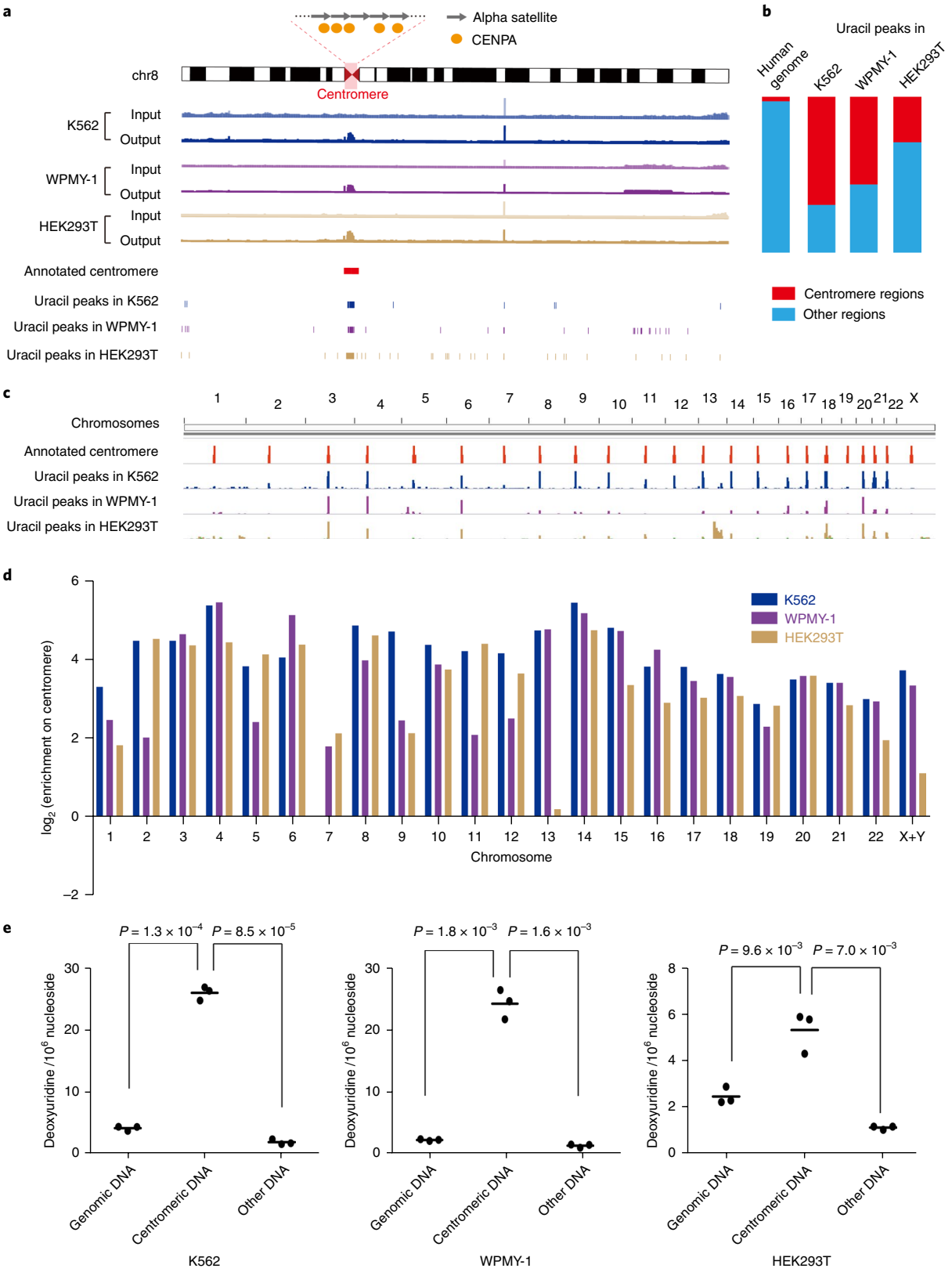

Fig. 3 | Uracil is enriched in the human centromeric DNA. a, IGV views of sequencing data and uracil peaks on chromosome 8. **b**, Proportions of uracil peaks in the centromere regions and other genomic regions. **c**, IGV views of uracil peaks for all the chromosomes of K562, WPMY-1 and HEK293T cells. Due to the scale of this figure, please refer to **d** for more accurate information of dU enrichment at each centromere. **d**, Relative uracil enrichment for the centromere of each chromosome. Enrichments were normalized by the lengths of each chromosome and the corresponding centromere. For all data in **a**–**d**, combined data sets of two replicates (cell culture) were analyzed and shown here; the patterns of individual replicates are the same. **e**, Quantification of uracil in the genomic DNA, centromeric DNA and the remaining DNA by LC–MS/MS. The mean values are indicated with black horizontal lines ($n = 3$); P values arise from a one-tailed t-test.

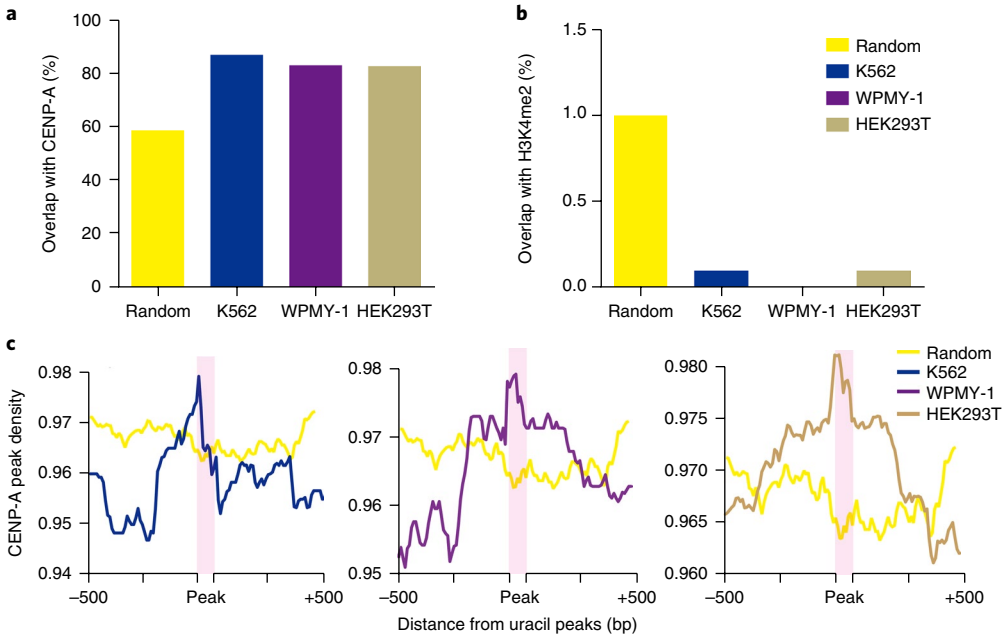

Fig. 4 | Uracil peaks colocalize with the CENP-A binding regions at the centromere. a, Percentages of uracil peaks that overlap with CENP-A peaks at the centromere. "Random," 2,000 random peaks in the centromere regions were selected as a control. **b**, Percentages of uracil peaks that overlap with H3K4me2 peaks at the centromere. **c**, CENP-A peak densities across the uracil peaks in the centromere region. For all the three cell lines, combined data sets of two replicates (cell culture) were analyzed and shown here; the patterns of individual replicates are the same.

excision-seq as an independent means to confirm this observation[29]. Although Excision-seq relies on the presence of densely located uracil in both DNA strands, we still found enrichment of uracil in the centromeres of both K562 and WPMY-1 cells (Supplementary Fig. 7). Moreover, to exclude the possibility that the centromeric enrichment of uracil is due to its high A/T-rich nature, we calculated the AT contents of LINEs, SINEs and LTRs, which are depleted of uracil. We found that the AT contents in these elements are comparable to that of the centromeres (Supplementary Fig. 8), indicating that enrichment of uracil at the centromere is not caused by its A/T-rich sequence context. Furthermore, we also performed biotin labeling using biotin-dCTP instead of biotin-dUTP and observed a similar enrichment of uracil in the centromere for both HEK293T and K562 cells, strongly supporting the observed enrichment of centromeric uracil (Supplementary Fig. 9). Centromeric DNA is composed of different classes of α-satellite monomers[37]; we found that uracil is enriched in all types of monomers, especially in the D1 and D2 classes (Supplementary Fig. 10). This observation may also explain the enrichment of uracil peaks at satellite repeats shown in Fig. 2d.

Mass spectrometry confirms a higher centromeric dU level. To further validate the enrichment of uracil at the centromeres, we sought to utilize mass spectrometry (MS) to quantitatively measure the level of uracil on centromeres. Because it is challenging to separate the centromeric DNA from the rest of the human genome, we first established a restriction-enzyme-based protocol to enrich the human centromeric DNA. We found that a special restriction endonuclease site (5′-NGCATTC-3′) is frequently occurring within the human centromeric regions but is rarely present in the rest of genomic regions (Methods). Digestion of genomic DNA using endonuclease BsmI, which recognizes this restriction site, would primarily result in two types of DNA fragments: fragments less than 250 bp that are mainly composed of centromeric DNA and those larger than 2,000 bp that are mainly from other genomic regions (Supplementary Fig. 11a). Hence, centromeric DNA can be readily separated and isolated by size selection following restriction endonuclease digestion. To further prove that the shorter fragments are enriched with centromeric DNA sequences, we subjected them to high-throughput sequencing (Supplementary Fig. 11b). We found that the shorter fragments consist of ~50–65% centromeric DNA sequences, thereby representing approximately 20-fold enrichment compared to the original genomic DNA.

We then analyzed the uracil content of the genomic DNA, the enriched centromeric DNA and the remaining DNA regions by LC–MS/MS[32,33]. For all three cell lines, the uracil content of the centromeric DNA is significantly higher than that of the genomic DNA (Fig. 3e and Supplementary Fig. 12); as expected, the remaining DNA regions, which are depleted of centromere DNA, contain the lowest level of uracil. Using K562 cells as an example, the uracil level is about 4.1 p.p.m., 26.0 p.p.m. and 1.8 p.p.m. for the three DNA samples, respectively; hence, the uracil level of the enriched centromeric DNA sample is approximately 6.5-fold higher relative to that of the genomic DNA sample. Considering that the enriched centromeric DNA samples contain approximately half of the centromere sequences (Supplementary Fig. 11b), the actual uracil level at the centromere of K562 cells is estimated to be ~50 p.p.m., which is comparable to the 5-formylcytosine level in the genomic DNA[38,39]. Moreover, we found a higher uracil level for the centromeric DNA of K562 and WPMY-1 cells than that in HEK293T cells; this is also consistent with our observation that the proportion of centromere uracil peaks is higher for the K562 and WPMY-1 cells (Fig. 3b).

We further demonstrated the presence of uracil at the centromere using the 3D-PCR technology[24,25]. Due to the repetitive nature of the centromere DNA, we first chose the centromeric peaks that allow the design of specific PCR primers. Two regions—one on the centromere of chromosome 6 and one on the centromere of chromosome 11—met the criteria and could be specifically amplified (Supplementary Table 4). We then performed 3D-PCR using K562 genomic DNA, which is either untreated or treated with a USER enzyme mix (containing UDG and endonuclease VIII). When a U:G pair is present, it will give rise to both a C:G and an A:T pair

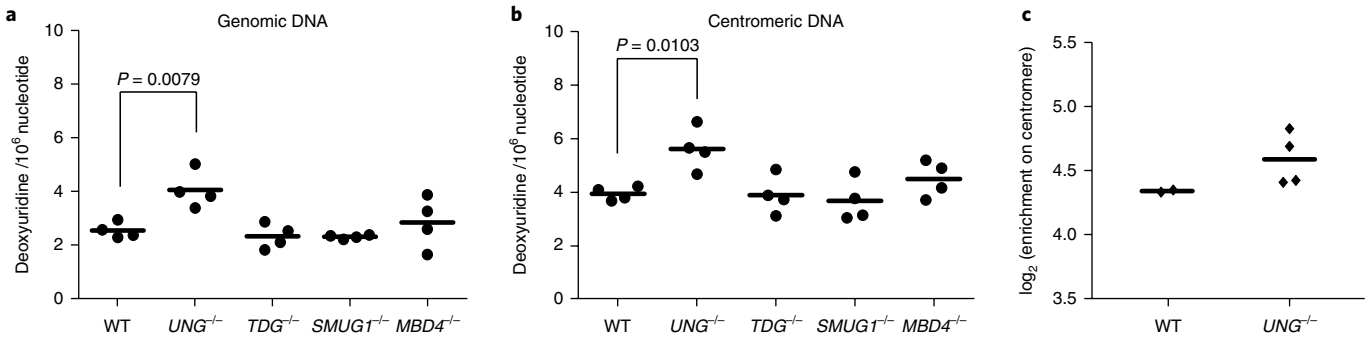

Fig. 5 | Uracil at the centromere can be excised by UNG. a, Deoxyuridine content in the genomic DNA of wild-type (WT), *UNG*$^{-/-}$, *TDG*$^{-/-}$, *SMUG1*$^{-/-}$ and *MBD4*$^{-/-}$ HEK293T cells quantified by LC–MS/MS. Each dot represents an independent experiment ($n = 4$, one-tailed *t*-test). **b**, Deoxyuridine content in the centromeric DNA of WT, *UNG*$^{-/-}$, *TDG*$^{-/-}$, *SMUG1*$^{-/-}$ and *MBD4*$^{-/-}$ HEK293T cells quantified by LC–MS/MS. Each dot represents an independent experiment ($n = 4$; one-tailed *t*-test). **c**, Relative enrichment of dU peaks on centromere of WT and *UNG*$^{-/-}$ HEK293T cells. Values represent independent experiments ($n = 2$ for WT and $n = 4$ for *UNG*$^{-/-}$).

in the PCR amplicons; the A:T-containing amplicons have a lower denaturation temperature and will be present only in the untreated sample, hence serving as the detection readout of 3D-PCR[24,25]. Indeed, we found that the regions on chromosome 6 (spanning satellite monomers R1 and R2) and chromosome 11 (spanning satellite monomers W5 and W1) both contain one or more U:G pair (Supplementary Fig. 13), directly demonstrating the existence of uracil in the dU-seq-identified peaks. Collectively, our quantitative MS analysis and 3D-PCR experiments unambiguously validated the presence of uracil in the human centromere.

Centromeric dU co-localizes with CENP-A binding regions. We next analyzed the potential correlation of the distribution of centromeric uracil with that of histone proteins at the centromere. There are CENP-A (a histone H3 variant) and interspersed H3 domains in the human centromeres: CENP-A is the key determinant of centromere identity and is essential for kinetochore assembly, whereas the H3 domains contain a H3K4me2 modification but lack a H3K9me modification[40]. We then calculated the percentage of centromeric uracil peaks that overlap with CENP-A or H3K4me2 binding sites using the published chromatin immunoprecipitation sequencing (ChIP-seq) data for CENP-A and H3K4me2 (Methods). Interestingly, uracil preferentially occurs in the CENP-A binding regions (Fig. 4a). In contrast, uracil peaks are depleted in the regions with H3K4me2 modification (Fig. 4b). We also calculated the CENP-A peak density around the centromeric uracil peaks and observed higher CENP-A signals in the uracil peaks (Fig. 4c). We concluded that within the centromeres, uracil peaks preferentially colocalize with the CENP-A binding regions.

Uracil at the centromere can be excised by human UNG. To examine whether or not centromeric uracil can be excised by human DNA glycosylases, we generated *UNG*$^{-/-}$, *TDG*$^{-/-}$, *SMUG1*$^{-/-}$ and *MBD4*$^{-/-}$ HEK293T cell lines using the CRISPR–cas9 system (Supplementary Fig. 14). We first showed that on the global level, only the *UNG* knockout cell exhibits a higher level of uracil in the genomic DNA (Fig. 5a), which is consistent with previous findings[41]. We then enriched the centromeric DNA sequences using BsmI digestion and measured the centromeric uracil content in these knockout cells. Compared to the wild-type cells, *UNG*$^{-/-}$ cells showed an increased level of centromeric uracil, whereas *TDG*$^{-/-}$, *SMUG1*$^{-/-}$ and *MBD4*$^{-/-}$ cells did not demonstrate a significant difference (Fig. 5b). In addition, we treated both HEK293T and *UNG*$^{-/-}$ cells with 5-fluoro-2′-deoxyuridine (5FdUR), a commonly used thymidylate synthase inhibitor, and found that the centromeric uracil level of the *UNG*$^{-/-}$ cells is significantly higher than that of HEK293T cells (Supplementary Fig. 15a). More importantly, we overexpressed UNG2 in wild-type and *UNG*$^{-/-}$ (as a rescue experiment) HEK293T cells and found a decrease of centromeric uracil levels in both cell lines (Supplementary Fig. 15b,c). Furthermore, UNG2 overexpression also reduced the uracil level at the centromere when these cells were treated with 5FdUR (Supplementary Fig. 15d). We next performed dU-seq for the *UNG*$^{-/-}$ cells and observed a higher percentage of uracil peaks on centromeres compared to wild-type HEK293T cells (Fig. 5c). Because the *UNG* gene encodes both the mitochondrial UNG1 and the nuclear UNG2, we also analyzed uracil modification in mitochondrial DNA. For both wild-type and *UNG*$^{-/-}$ 293T cells, dU-seq did not identify any mitochondrial uracil peaks, suggesting an absence of uracil hotspot in mitochondrial DNA. Collectively, our results showed that uracil in centromeric DNA can be excised by the UNG2 glycosylase in human cells.

Discussion

In this study, we report a genome-wide method for the detection of uracil in human cells. dU-seq has multiple advantages: it requires only commercially available enzymes and reagents, needs accessible starting materials (~1–2 μg genomic DNA as input) and uses existing bioinformatics tools for data processing (Methods). Previously, excision-seq has been developed to detect the genome-wide uracil in *E. coli* and yeast[29]. Both excision-seq and dU-seq positively enrich uracil-containing DNA; yet, they have different requirements for enzyme cleavage and uracil density (Supplementary Fig. 7a and Fig. 1a). Furthermore, the concept of dU-seq may be applied to the detection of other modifications in the human genome. For instance, specific and robust DNA glycosylases for 7,8-dihydro-8-oxoguanine (8-oxoguanine) and cyclobutane pyrimidine dimers (CPDs) are well characterized[15,42,43]; by optimizing the conditions for labeling, the procedure of dU-seq could be adapted to detect these modifications as well. Thus, dU-seq is sensitive, convenient and has broad potential for finding modified nucleotides in the human genome.

Our results show that instead of being randomly distributed along chromosomes, uracil is enriched in the human centromere. Modifications with a clear pattern, for instance 5-methylcytosine and its oxidative derivatives, are catalyzed by dedicated enzymes and are tightly regulated[1]; the recently reported 6-methyladenosine in the higher eukaryotic genomes may also have a gene regulatory function[4]. Hence, it is tempting to speculate that the centromeric enrichment of uracil is indicative of regulated biogenesis and potential functions. Because dU-seq detects both U:A and U:G, the centromeric deoxyuridine identified by dU-seq can come from either

dUMP misincorporation or dC deamination. The former situation would require replication machinery to frequently incorporate dUMP at the centromere, whereas the latter case would involve a high occurrence of spontaneous or enzymatic dC deamination in the centromeric region. Upon 5FdUR treatment, we observed a more pronounced increase in uracil level in the centromeric DNA than in the genomic DNA, indicating a preferred incorporation of uracil at centromeres under elevated dUTP conditions. However, whether deoxyuridine is preferentially incorporated into centromeric DNA under the physiological condition remains to be examined. It is necessary to point out that treatment with 5FdUR may cause certain amount of 5FdU incorporation in genomic DNA; nevertheless, 5FdU is much less efficiently removed by UDG[21,44] and also has a different molecular weight in quantitative MS analysis compared to dU. On the other hand, based on our 3D-PCR results, U:G pairs are present at the centromeres of chromosomes 6 and 11, suggesting that dC deamination is one source of centromeric uracil. Although spontaneous dC deamination may occur throughout the genome, more than ten AID/APOBEC family proteins have been identified in human cells, and the genomic targets of many of these proteins are currently unknown[12]. Lastly, inefficient removal of uracil could also contribute to the centromeric enrichment. We observed that UNG overexpression led to a decrease in centromeric uracil level, indicating insufficient repair of uracil at the centromere. Presumably, the compact structure of the centromere could influence the accessibility and repair of UNG2, ultimately resulting in the accumulation of centromeric uracil.

Centromeres allow chromosomes to associate with spindle microtubules and segregate chromosomes to daughter cells[40]. In human cells, although centromeres are composed of repetitive α-satellites, they do not seem to be defined by the primary DNA sequences; the identity of centromeres is epigenetically determined by CENP-A localization[40,45]. In this study, we showed that uracil is enriched on centromeres, is at a higher level at centromeric DNA (comparable to that of 5-formylcytosine in the genomic DNA), colocalizes with the CENP-A binding regions and can be removed by UNG2. Thus, it is appealing to hypothesize that uracil on centromeres could be a mark for CENP-A binding. Alternatively, repair-coupled chromatin remodeling has been suggested to stimulate CENP-A deposition[40], and thus uracil on centromeres could also function as an intermediate during CENP-A assembly. For instance, a previous study has revealed that an unidentified deoxycytidine deaminase and UNG2 are involved in CENP-A assembly in *Xenopus* extracts[46]. Also, UNG2 has been found to colocalize with CENP-A at centromeres in normally cycling cells by an immunofluorescence approach[47]. Additionally, repair of APOBEC3B-induced C-to-U mutations at estrogen receptor target genes has been demonstrated to facilitate chromatin remodeling[48]. Hence, direct and in vivo evidence in the future may demonstrate whether or not centromeric uracil could facilitate the centromere-specific assembly of CENP-A.

In summary, our study demonstrated the centromeric enrichment of uracil in the human genome. Our integrated approaches revealed the abundance of uracil at the human centromeres and also enabled the evaluation of uracil regulation. Our selective labeling and pull-down method allowed the genome-wide profiling of uracil in human cells, providing a reference and tool for future investigations of this DNA modification.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41589-018-0065-9.

Received: 15 May 2017; Accepted: 19 March 2018;
Published online: 21 May 2018

References

1. Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* **9**, 465–476 (2008).
2. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461**, 1071–1078 (2009).
3. Shen, L., Song, C. X., He, C. & Zhang, Y. Mechanism and function of oxidative reversal of DNA and RNA methylation. *Annu. Rev. Biochem.* **83**, 585–614 (2014).
4. Luo, G. Z., Blanco, M. A., Greer, E. L., He, C. & Shi, Y. DNA N^6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.* **16**, 705–710 (2015).
5. David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* **447**, 941–950 (2007).
6. Wyrick, J. J. & Roberts, S. A. Genomic approaches to DNA repair and mutagenesis. *DNA Repair (Amst.)* **36**, 146–155 (2015).
7. Krokan, H. E., Drabløs, F. & Slupphaug, G. Uracil in DNA—occurrence, consequences and repair. *Oncogene* **21**, 8935–8948 (2002).
8. Kavli, B., Otterlei, M., Slupphaug, G. & Krokan, H. E. Uracil in DNA—general mutagen, but normal intermediate in acquired immunity. *DNA Repair (Amst.)* **6**, 505–516 (2007).
9. Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26**, 261–292 (2008).
10. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).
11. Harris, R. S. & Dudley, J. P. APOBECs and virus restriction. *Virology* **479-480**, 131–145 (2015).
12. Siriwardena, S. U., Chen, K. & Bhagwat, A. S. Functions and malfunctions of mammalian DNA-cytosine deaminases. *Chem. Rev.* **116**, 12688–12710 (2016).
13. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).
14. Matsumoto, Y. et al. Up-regulation of activation-induced cytidine deaminase causes genetic aberrations at the CDKN2b-CDKN2a in gastric cancer. *Gastroenterology* **139**, 1984–1994 (2010).
15. David, S. S. & Williams, S. D. Chemistry of glycosylases and endonucleases involved in base-excision repair. *Chem. Rev.* **98**, 1221–1262 (1998).
16. Krokan, H. E. & Bjørås, M. Base excision repair. *Cold Spring Harb. Perspect. Biol.* **5**, a012583 (2013).
17. Nilsen, H. et al. Nuclear and mitochondrial uracil-DNA glycosylases are generated by alternative splicing and transcription from different positions in the UNG gene. *Nucleic Acids Res.* **25**, 750–755 (1997).
18. Otterlei, M. et al. Post-replicative base excision repair in replication foci. *EMBO J.* **18**, 3834–3844 (1999).
19. Vallabhaneni, H. et al. Defective repair of uracil causes telomere defects in mouse hematopoietic cells. *J. Biol. Chem.* **290**, 5502–5511 (2015).
20. Muha, V. et al. Uracil-containing DNA in *Drosophila*: stability, stage-specific accumulation, and developmental involvement. *PLoS Genet.* **8**, e1002738 (2012).
21. Kavli, B. et al. hUNG2 is the major repair enzyme for removal of uracil from U:A matches, U:G mismatches, and U in single-stranded DNA, with hSMUG1 as a broad specificity backup. *J. Biol. Chem.* **277**, 39926–39936 (2002).
22. Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301–304 (1999).
23. Neddermann, P. & Jiricny, J. Efficient removal of uracil from G.U mispairs by the mismatch-specific thymine DNA glycosylase from HeLa cells. *Proc. Natl Acad. Sci. USA* **91**, 1642–1646 (1994).
24. Suspène, 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.* **86**, 125–129 (2005).
25. Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nat. Struct. Mol. Biol.* **17**, 222–229 (2010).
26. Maul, R. W. et al. Uracil residues dependent on the deaminase AID in immunoglobulin gene variable and switch regions. *Nat. Immunol.* **12**, 70–76 (2011).
27. Riedl, J., Fleming, A. M. & Burrows, C. J. Sequencing of DNA lesions facilitated by site-specific excision via base excision repair DNA glycosylases yielding ligatable gaps. *J. Am. Chem. Soc.* **138**, 491–494 (2016).
28. Róna, G. et al. Detection of uracil within DNA using a sensitive labeling method for in vitro and cellular applications. *Nucleic Acids Res.* **44**, e28 (2016).
29. 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.* **24**, 1534–1542 (2014).
30. Horváth, A. & Vértessy, B. G. A one-step method for quantitative determination of uracil in DNA by real-time PCR. *Nucleic Acids Res.* **38**, e196 (2010).

31. Hansen, E. C. et al. Diverse fates of uracilated HIV-1 DNA during infection of myeloid lineage cells. *eLife* **5**, e18447 (2016).
32. Galashevskaya, A. et al. A robust, sensitive assay for genomic uracil determination by LC/MS/MS reveals lower levels than previously reported. *DNA Repair (Amst.)* **12**, 699–706 (2013).
33. Bulgar, A. D. et al. Removal of uracil by uracil DNA glycosylase limits pemetrexed cytotoxicity: overriding the limit with methoxyamine to inhibit base excision repair. *Cell Death Dis.* **3**, e252 (2012).
34. Xiao, G. et al. Crystal structure of *Escherichia coli* uracil DNA glycosylase and its complexes with uracil and glycerol: structure and glycosylase mechanism revisited. *Proteins* **35**, 13–24 (1999).
35. Ali, M. H., Al-Saad, K. A. & Ali, C. M. Biophysical studies of the effect of high power ultrasound on the DNA solution. *Phys. Med.* **30**, 221–227 (2014).
36. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
37. Shepelev, V. A. et al. Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom. Data* **5**, 139–146 (2015).
38. Pfaffeneder, T. et al. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Edn. Engl.* **50**, 7008–7012 (2011).
39. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
40. Allshire, R. C. & Karpen, G. H. Epigenetic regulation of centromeric chromatin: old dogs, new tricks? *Nat. Rev. Genet.* **9**, 923–937 (2008).
41. Nilsen, H. et al. Uracil-DNA glycosylase (UNG)-deficient mice reveal a primary role of the enzyme during DNA replication. *Mol. Cell* **5**, 1059–1065 (2000).
42. Lloyd, R. S. Investigations of pyrimidine dimer glycosylases–a paradigm for DNA base excision repair enzymology. *Mutat. Res.* **577**, 77–91 (2005).
43. Boiteux, S., Coste, F. & Castaing, B. Repair of 8-oxo-7,8-dihydroguanine in prokaryotic and eukaryotic cells: properties and biological roles of the Fpg and OGG1 DNA N-glycosylases. *Free Radic. Biol. Med.* **107**, 179–201 (2017).
44. Grogan, B. C., Parker, J. B., Guminski, A. F. & Stivers, J. T. Effect of the thymidylate synthase inhibitors on dUTP and TTP pool levels and the activities of DNA repair glycosylases on uracil and 5-fluorouracil in DNA. *Biochemistry* **50**, 618–627 (2011).
45. Müller, S. & Almouzni, G. Chromatin dynamics during the cell cycle at centromeres. *Nat. Rev. Genet.* **18**, 192–208 (2017).
46. Zeitlin, S. G., Patel, S., Kavli, B. & Slupphaug, G. *Xenopus* CENP-A assembly into chromatin requires base excision repair proteins. *DNA Repair (Amst.)* **4**, 760–772 (2005).
47. Zeitlin, S. G. et al. Uracil DNA N-glycosylase promotes assembly of human centromere protein A. *PLoS One* **6**, e17151 (2011).
48. Periyasamy, M. et al. APOBEC3B-mediated cytidine deamination is required for estrogen receptor action in breast cancer. *Cell Reports* **13**, 108–121 (2015).

Acknowledgements

The authors would like to thank G. Liu and H. Li for measurements with LC–MS/MS; B. Xia, X. Li and X. Xiong for technical advice and discussions. This work was supported by the National Natural Science Foundation of China (nos. 21522201 and 21472009 to C.Y.), the National Basic Research Foundation of China (nos. 2016YFC0900301 and 2014CB964900 to C.Y.) and the Fok Ying Tung Education Foundation (no. 161018 to C.Y.).

Author contributions

X.S. and C.Y. conceived the project; X.S., M.L. and C.Y. designed the experiments and wrote the manuscript with the help of Z.L.; X.S., M.L., C.Z., S.H. and X.Z. performed the experiments; Z.L. and H.M. performed bioinformatics analysis. All authors commented on and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41589-018-0065-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.Y.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell culture. K562, WPMY-1, HEK293T were used for analysis in this study. K562 was maintained in RPMI 1640 medium (Gibco) supplemented with 10% FBS and 1% penicillin/streptomycin. WPMY-1 and HEK293T were maintained in DMEM medium (Gibco) supplemented with 10% FBS and 1% penicillin/streptomycin. Mycoplasma contamination tests were performed routinely using GMyc-PCR Mycoplasma Test Kit from YEASEN (cat. #40601). To treat the cells with thymidylate synthase inhibitor, cells were plated and grew to a density of 50% to 70%. Then, 5 μM 5FdUR (Sigma, F0503) was added, and cells were harvested after 48 h of treatment.

Antibodies. Monoclonal rabbit anti-UNG antibody was purchased from Abcam (ab109214). Polyclonal rabbit anti-MBD4 antibody was purchased from Abcam (ab191037). Monoclonal rabbit anti-SMUG1 antibody was purchased from Abcam (ab 192240). Polyclonal rabbit anti-TDG antibody was purchased from Sigma-Aldrich (HPA052263). Monoclonal mouse anti-FLAG antibody was purchased from Sigma-Aldrich (F1804). Monoclonal mouse anti-β actin antibody was purchased from CWBiotech (CW0096). The secondary antibodies used were anti-mouse-IgG-HRP (CW0102; CWBiotech) and anti-rabbit-IgG-HRP (CW0103; CWBiotech).

Spike-in model DNA sequences preparation. The single-dU sequence (A:U pair) was generated from a short double-strand DNA containing one dU (purchased from Takara) and another ~200 bp DNA duplex by sticky end ligation. The multiple-dU spiked-in was PCR amplified from lambda DNA by EASYTaq DNA Polymerase (Transgene) with a cocktail of dATP/dGTP/dCTP and 10% dUTP (Promega) and 90% dTTP. The single G:U mispair sequence was annealed by a 58-mer short primer with a dU to a long single strand DNA and primer extended to form double strand DNA. The AP site sequence was generated from a G:U mispair sequence by uracil excision with UDG at 37 °C for 30 min. The single-strand break sequence was annealed by three oligonucleotides. The overhang sequence was annealed by two oligonucleotides. The control spiked-in was PCR amplified from lambda DNA by EASYTaq DNA Polymerase with dATP/dGTP/dCTP/dTTP. All spike-in sequences were stored at −80 °C. Detailed sequences were in the Supplementary Note 2.

dU-seq. Cells were harvested, and genomic DNA was extracted by Universal Genomic DNA Kit (CWbiotech, CW2298). Genomic DNA was digested to 100–500 bp with NEBNext dsDNA Fragmentase (NEB, M0348). DNA fragments were purified with 1.8 × Agencourt AMPure XP beads (Beckman Coulter) following Spin-6 column (Bio-Rad) purification. 2 μg DNA fragments and 15 pg spike_in sequences were used for end repair with NEBNext End Repair Module (NEB, E6050) and 1 μl *E. coli* ligase was added to repair nicks in DNA. DNA was then purified with 1.8 × AMPure XP beads. dA was added to the 3′ end of double-stranded DNA by NEBNext dA-Tailing Module (NEB, E6053) and purified with 1.8 × AMPure XP beads.

Damages that may interfere the following labeling step were repaired in a mixture of 2 μl endonuclease IV (NEB, M0304), 1 μl *Bst* full-length polymerase (NEB, M0328), 2 μl Taq ligase (NEB, M0208), 1 μl NAD^+, 1 μl dNTP (2.5 mM each) in NEBuffer 3 for 40 min at 37 °C and 60 min at 45 °C. DNA was purified and subjected to in vitro BER labeling in a mixture of 1 μl UDG (NEB, M0280), 1.5 μl endonuclease IV, 0.8 μl *Bst* full-length polymerase, 1.7 μl Taq ligase, 1 μl NAD^+, [200 nM biotin-dUTP, 200 nM dATP, 200 nM dCTP and 200 nM dGTP] or [200 nM biotin-dCTP, 200 nM dATP, 200 nM dTTP and 200 nM dGTP], in NEBuffer 3 for 40 min at 37 °C. After 1.8 × XP bead purification, fragments labeled with biotin were enriched by streptavidin C1 beads (Invitrogen) following the manufacturer's guidelines.

Y adaptor (see Supplementary Note 1) was ligated (NEBNext Quick Ligation Module, E6056) to double-stranded DNA on streptavidin C1 beads so that free adaptor can be removed by beads washing. DNA was then eluted from beads by 95 °C 3 min using deionized water. Eluted DNA was subjected to PCR amplification. Sequencing was performed by Illumina Hiseq X.

Calculating enrichment of spike_in sequences by real-time PCR. 15 pg of each spike_in sequence with different qPCR primer (Supplementary Table 1) was spiked into 2 μg fragmented genomic DNA. Before DNA labeling, 5% of the mixed DNA was separated as input. Enrichment was performed as procedures in dU-seq above. After DNA purification, SYBR Premix Ex Taq^T II (Takara) was used to perform real-time PCR with LightCycler 480 Real-Time PCR System (Roche). The "enrichment" represents the fold change in modified DNA/unmodified DNA relative to the input sample; calculated by comparing pull-down sample to input sample using real-time PCR assay.

dU-seq data processing and analysis. 150 bp pair-end reads were first sent for adaptor and quality trimming using trim_galore, and reads shorter than 25 nt after trimming were excluded. For spike-in DNA sequences calculation, bowtie2 (ref. [49]) was used to reads mapping to corresponding sequences. For genome-wide dU sites identification, processed reads mapped to spike-in sequence were discarded, and then mapped to human reference genome (hg38) using bowtie2. For multiple alignments, bowtie2 searches and reports a random one in the best matches. After alignment, peaks were called using model-based analysis of ChIP-Seq (MACS2)[50] using nonredundant reads. Peaks overlap with that in the control sample were removed. Genomic annotations were performed using Homer software. Gene annotations (RefSeq) were download from UCSC. Reads visualization was done by IGV. Intersection between bed files was performed using BEDTools. The CENP-A and H3K4me2 sequencing data were downloaded from GEO databases (GSE45497; GSM733651; GSM733780) and mapped to hg38 reference genome.

Excision-seq library preparation and data analysis. The excision-seq library preparation procedures were based on the pre-digestion method of excision-seq for uracil[29] with some modifications. Cells were harvested and genomic DNA was extracted by Universal Genomic DNA Kit (CWbiotech, CW2298). High-molecular-weight genomic DNA was first selected by 0.35 × AMPure XP beads and treated with 5 units of UDG and 10 units of endonuclease VIII (NEB, M0299) for 2 h at 37 °C. Undigested large genomic DNA was removed by 0.35 × AMPure XP beads twice. The remaining short DNA fragments were subjected to library preparation (NEB, E7645) and sequencing by Illumina Hiseq X. Sequences were analyzed by alignment to human reference genome (hg38) using bowtie2. Reads visualization was done by IGV.

Statistical analyses. One-tailed Student's *t*-test were used to analyze the data from independent experiments for significance test of statistical hypothesis using significance levels: **P* value < 0.05; **P value < 0.01; *n* indicates the number of independent experimental replicates.

Generation of *UNG*/*MBD4*/*TDG*/*SMUG1* knockout 293T cells. CRISPR–Cas9 was used to generate the four knockout cell lines. The plasmid PX330 containing the individual guide RNA sequence was transfected into HEK293T cells using Lipofectamine 2000 (Life Technologies). The guide RNA was designed according to the method described before[51], and the target DNA sequences corresponding to the target gene are listed in Supplementary Table 2. Cells were diluted and seeded into a 96-well dish 60 h after transfection at the concentration of 0.5 cell per well. After 2 weeks, single colonies were transferred to a 24-well dish. Genotyping of each colony was carried out by PCR and enzyme digestion. After the first-round selection, western blot was performed to select for colonies no longer expressing the target gene. The edited gene sequences of successful colonies were examined by TA cloning and Sanger sequencing.

Generation of UNG2 overexpression cell lines in HEK293T and *UNG*$^{-/-}$ cells. The coding sequence of *UNG2* gene was amplified with *TransStart FastPfu* DNA Polymerase (Transgene, AP221) using HEK293T cDNA as template and was cloned into pLentiCMV vector. Then, the vector was transfected into cells using Lipofectamine 2000 (Invitrogen, 11668019) to obtain the packaged lentivirus, which was used to infect the HEK293T and *UNG*$^{-/-}$ cells for 12 h and twice, respectively. The infected cells were maintained in the growth medium supplemented with 12 μM Blasticidin S HCl (Invitrogen, A1113903), and the stable UNG2 overexpression cell lines were generated after screening for 2 weeks. Finally, western blot was performed to validate whether these stable cell lines overexpress the UNG2 protein. The primers used for amplification of the *UNG2* coding sequence are listed in Supplementary Table 3).

Separation of centromere DNA from other genomic DNA. For centromeric DNA, a restriction endonuclease site (BsmI: 5′-NGCATTC-3′) occurs frequently so that the centromeric DNA can be digested to less than 250 bp. For other genomic regions, the frequency of this sequence is very low, and those DNA are digested to more than 2 K bp, based on a tool that calculates the theoretical restriction site distances in the human genome (http://tools.neb.com/~posfai/TheoFrag/TheoreticalDigest.human.html). 200 μg Genomic DNA was digested with 200 U restriction endonuclease BsmI (NEB, R0134L) in 400 μl at 37 °C for 3 h. 0.45 × AMPure XP beads were used to bind most of large DNA fragments, and the supernatant was transferred to another new tube. Then 2.4 × AMPure XP beads were added to the new tube to recover all the leftover DNA, and DNA was eluted in 50 μl (small volume) ddH_2O. 0.45 × AMPure XP beads were used to remove the large DNA fragments again. The supernatant was transferred to another new tube, and new beads (without buffer) were added to bind the large DNA fragments once more. Finally, the small fragments (centromere DNA) was recovered by 2.4× AMPure XP beads.

Quantification of uracil in genomic DNA by LC–MS/MS. Cells were harvested and genomic DNA was extracted by Universal Genomic DNA Kit (CWbiotech, CW2298). Genomic DNA (~25 μg) was incubated with 10 U of purified UDG (NEB, M0280) in 50 μl of reaction buffer at 37 °C for 2 h. Three volumes of ice-cold acetonitrile were then added and centrifuged (16,100 × *g* for 20 min at 4 °C). The supernatants were transferred to new tubes and vacuum centrifuged until they were dry at room temperature (20–25 °C). The samples were dissolved in 20–30 μl ddH_2O and filtrated by Millex-GV Syringe Filter Unit (0.22 μm, PVDF, 4 mm). 10 μl of the solution was injected into LC–MS/MS, separated by ultra-performance LC on a C18 column and then detected by triple-quadrupole MS (Agilent UPLC

1290 - MS/MS 6495) using negative electrospray ionization, monitoring the mass transitions of 111.0→42.0 for uracil[32,33]. The uracil concentration was quantified according to the standard curve running for the same batch of samples. The DNA concentration was quantified by Qubit dsDNA HS; 0.3 µg DNA is calculated as 1 nmol nucleotide. To quantify the dT content, the same genomic DNA (100 ng) was treated with 5U DNA Degradase Plus (Zymo research, E2021) in 25 µl total volume at 37 °C for 2 h. Then, 975 µl ddH_2O was added and centrifuged (14,800 r.p.m. for 20 min at 4 °C). 2 µl of the supernatant was analyzed by LC–MS/MS as mentioned above except using positive electrospray ionization and monitoring the mass transitions 243.0→127.0 for dT.

Validation of centromeric uracil by 3D-PCR. The K562 genomic DNA was incubated with USER Enzyme (NEB, M5505) at 37 °C for 0.5 h. Then, 50 ng treated DNA or control DNA (without USER enzyme digestion) was used separately as input for PCR amplification using 2 × Es Taq MasterMix (CWbiotech, CW0690) and primers listed in Supplementary Table 4), which amplify the specific centromeric region at chromosome 6 and chromosome 11 (28 cycles: 94 °C, 30 s; 57 °C, 30 s; and 72 °C, 30 s). The PCR products were analyzed by agarose gel and recovered with EasyPure Quick Gel Extraction Kit (TransGen Biotech, EG101). 2 ng of the recovered DNA was used for another PCR amplification under different denaturation temperatures (T_d) with the same primer and polymerase above (28 cycles: T_d gradient as indicated in Supplementary Fig. 13; 57 °C, 30 s; and 72 °C, 30 s). Finally, these PCR products were visualized on agarose gel.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. Sequencing data have been deposited into the Gene Expression Omnibus (GEO). The accession number is GSE99011.

References

49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
51. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

nature research

Corresponding author(s): Chengqi Yi

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical method was used to predetermine the sample size. A minimum of triplicates was chosen in our quantitative MS/MS and qPCR experiments, which is sufficient for us to perform statistical tests when needed. Two independent sample replicates were used in our high-throughput sequencing experiments, because larger sample size could provide very limited extra information but increase the costs.

2. Data exclusions

Describe any data exclusions.

No data were excluded from our qPCR, high-throughput sequencing and QQQ results.

3. Replication

Describe whether the experimental findings were reliably reproduced.

All experimental findings were reliably reproduced.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

When drugs (5FdUR) was used to treat the cells, the samples were equally divided into two parts and one was randomly selected for drug treatment. No other specific method was used to randomize the allocation of samples and all the experiments were performed in parallel.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No animal subjects or human research participants were involved in this study, so blinding was not relevant. Uniform data processing was applied to all the samples in analysis of sequencing data regardless of whether they were control or pull-down samples. Since the results reported are either entirely quantitative (i.e., quantitative MS/MS results, enrichment quantified by qPCR) or unambiguous in nature (e.g., DNA agrose gel images), blinding was not necessary for the experiments performed.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

bowtie2 was used to reads mapping to corresponding sequences. Peaks were called using model-based analysis of ChIP-Seq (MACS2) using nonredundant reads. Reads visualization was done by IGV. Intersection between bed files was performed using BEDTools. Graphs were plotted with GraphPad Prism 7.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All reagents and enzymes used were obtained from commercial suppliers. Plasmids and cell lines would be available upon request.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Monoclonal rabbit anti-UNG antibody was purchased from abcam (ab109214), which was validated by western blot and immunohistochemistry by manufacturer (http://www.abcam.cn/ung-antibody-epr4371-ab109214.html).
Polyclonal rabbit anti-MBD4 antibody was purchased from abcam (ab191037), which was validated by western blot by manufacturer (http://www.abcam.cn/mbd4-antibody-c-terminal-ab191037.html).
Monoclonal rabbit anti-SMUG1 antibody was purchased from abcam (ab192240), which was validated by western blot by manufacturer (http://www.abcam.cn/smug1-antibody-epr15624-ab192240.html).
Polyclonal rabbit anti-TDG antibody was purchased from Sigma-Aldrich (HPA052263), which was validated by western blot, immunocytochemistry and immunocytochemistry by manufacturer (https://www.proteinatlas.org/ENSG00000139372-TDG/antibody#antibody_summary).
Monoclonal mouse anti-FLAG antibody was purchased from Sigma-Aldrich (F1804), which was validated by western blot by manufacturer (https://www.proteinatlas.org/ENSG00000139372-TDG/tissue).
Monoclonal mouse anti-β-actin antibody was purchased from CWBiotech (CW0096), which was validated by western blot by manufacturer (http://www.cwbiotech.com/upload/image/201709/97712f17-4ac7-4630-92c6-4370437679d8.pdf).
The secondary antibodies used were anti-mouse-IgG-HRP (CW0102; CWBiotech, http://www.cwbiotech.com/goods/content/201707/10118.html) and anti-rabbit-IgG-HRP (CW0103; CWBiotech, http://www.cwbiotech.com/goods/content/201707/10119.html).

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

HEK293T was purchased from China Infrastructure of Cell Line Resources (3111C0001CCC000212);
WPMY-1 was a gift from Dr. Min Fang.
Fang, M. et al. The ER UDPase ENTPD5 promotes protein N-glycosylation, the Warburg effect, and proliferation in the PTEN pathway. Cell 143, 711-724 (2010)
K562 was a gift from Prof. Zhengfan Jiang.
Sun, W. et al. ERIS, an endoplasmic reticulum IFN stimulator, activates innate immune signaling through dimerization. Proc Natl Acad Sci U S A. 106(21):8653-8. (2009)

b. Describe the method of cell line authentication used.

All the three cell lines used were authenticated by Short Tandem Repeat (STR) profiling methods.

c. Report whether the cell lines were tested for mycoplasma contamination.

Yes, mycoplasma contamination tests were performed routinely using GMyc-PCR Mycoplasma Test Kit from YEASEN (CAT #40601). See Cell culture part of Online Methods section.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

HEK293T cells were used because of their high transfection efficiency and easy accessibility to laboratories world-wide, facilitating replication of our experiments. As a cancer cell line, K562 cells were widely used and these are many sequencing data sets available for further bioinformatics analysis. Thus we adopt this cell line in this study.
WPMY-1 cells have the same karyotype with normal human cells and were used as a control for HEK293T cells which are described as hypotriploid.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This study did not involve human research participants.