Nick-seq for single-nucleotide resolution genomic maps of DNA modifications and damage

Bo Cao[1,2,3,†,*], Xiaolin Wu[2,3,4,†], Jieliang Zhou[5], Hang Wu[2,6], Michael S. DeMott[2,7], Chen Gu[2,9], Lianrong Wang[4], Delin You[8], Peter C. Dedon[2,3,7,*]

[1] College of Life Sciences, Qufu Normal University, Qufu, Shandong, 273165, China

[2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3] Singapore-MIT Alliance for Research and Technology, Antimicrobial Drug Resistance Interdisciplinary Research Group, Singapore 138602, Singapore

[4] Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Ministry of Education and School of Pharmaceutical Sciences, Wuhan University, Wuhan, Hubei, 430071, China

[5] KK Research Center, KK Women's and Children's Hospital, 229899, Singapore

[6] School of Life Sciences, Anhui University, Hefei, Anhui, 230601, China

[7] Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[8] State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University, Shanghai, 200230, China

[9] Current affiliation: Merck Research Laboratories, Merck & Co., Inc., Boston, MA 02115, USA

[†] These authors contributed equally to this work.

Corresponding authors:

Bo Cao (ORCID ID: 0000-0002-7011-1676) (caobo@qfnu.edu.cn) or

Peter C. Dedon (ORCID ID: 0000-0003-0011-3067) (pcdedon@mit.edu.)

Abstract

Here we present the Nick-seq platform for quantitative mapping of DNA modifications and damage at single-nucleotide resolution across genomes. Pre-existing breaks are blocked and DNA structures converted to strand-breaks for 3'-extension by nick-translation to produce nuclease-resistant oligonucleotides, and 3'-capture by terminal transferase tailing. Libraries from both products are subjected to NGS sequencing. Nick-seq is a generally applicable method illustrated with quantitative profiling of single-strand breaks, phosphorothioate modifications, and DNA oxidation.

Main text (1,000~1,500 word including figure legend)

Chemical and structural modifications of DNA occur continuously in all cells, including DNA damage caused by endogenous and exogenous stresses[1], epigenetic modifications[2], and intermediates of DNA replication and repair[1]. A variety of methods have been developed for mapping of specific DNA modifications and damage, such as bisulfite-dependent sequencing for high-resolution mapping of DNA methylation[3] or chromatin immunoprecipitation coupled with next-generation sequencing (ChIP-seq) for low-resolution mapping specific DNA base damage[4,5]. Direct detection of some DNA modifications has also been achieved with third-generation sequencing methods, such as single-molecule real-time sequencing (SMRT)[6] and nanopore[7] sequencing, both of which have a low sensitivity and specificity.

Here we report the development of a widely applicable platform termed Nick-seq for highly sensitive quantitative mapping of any type of DNA modification or damage that can be converted to a strand-break. The method concept and workflow are illustrated in **Figure 1a**.

Purified genomic DNA is treated with DNA endonucleases to create sequencing-compatible fragments and the resulting 3'-OH ends are blocked with dideoxyNTPs and DNA polymerase I. The DNA modification or damage of interest is then converted to a strand-break by enzymatic or chemical means, followed by capture of the 3'- and 5'-ends of the resulting strand-breaks using two complementary strategies. One portion of the DNA sample is subjected to nick translation (NT) with α-thio-dNTPs to generate 100-200 nt phosphorothioate-containing oligonucleotides that are resistant to subsequent hydrolysis of the bulk of the genomic DNA by digestion with exonuclease III and $RecJ_f$. The purified PT fragment is used to generate a sequencing library for Illumina deep sequencing, with the modification of interest positioned at the 5'-end of the PT-labeled fragment.

A second portion of the same DNA sample is used for terminal transferase (TdT)-dependent poly(dT) tailing of the 3'-end of the strand break. The poly(dT) tail is then used to create sequencing libraries by reverse transcriptase template switching[8]. Subsequent NGS sequencing positions the modification of interest 3'-end of the poly(dT) tail.

As shown in **Figure 1b**, the workflow for analysis of the sequencing data uses the NT-derived reads as the primary dataset for developing a rough modification map and the TdT-derived reads as complementary corrective data. The basis for this hybrid approach lies in the fact that NT is agnostic to the base identify at the damage site but generates a high background resulting false positive sites, while the TdT method cannot be used to localize modifications occurring at dT due to loss of the poly(dT) tail during data analysis. This hybrid approach corrects the false-positives of NT reads with the TdT reads. For example, if the NT method maps a strand break at a T at position 1000 on the genome, then the TdT reads are examined for a strand break at position 999 or 1001. This 1 position shift accommodates the removal of the

poly(dT) tail during sequencing data processing and validates the NT map result. If TdT does not call a strand break at 999 or 1001, then the NT results is considered a false positive. In other cases, if the NT-detected site is located at a G, C, or A, then this is also validated by TdT method at the same site. The use of both methods thus greatly increases the sensitivity and specificity of the resulting DNA modification map.

We next validated Nick-seq by mapping the DNA strand breaks introduced into the *E.coli* genome by the site-specific, single-strand endonuclease, Nb.BsmI. This enzyme targets G/CATTC sites, of which there are 2,681 in the *E. coli* genome. As illustrated in **Figure 2a**, purified DNA was treated with Nb.BsmI and the Nick-seq-processed library sequenced using the Illumina NextSeq platform with an average of 10 million raw sequencing reads obtained in each sample. Paired-end sequencing confirmed that >80% of reads uniquely aligned to the *E.coli* genome (**Supplementary Table 1**). For the subequent reads enrichment (**Fig. 1b**), we calculated position-wise coverage values using the 5'-end of sequencing reads (Read 1 in NT, Read 2 in TdT) and defined Nick-seq peaks as having >5 reads and 2-times more reads than sites located one-nucleotide up- and down-stream. We then calculated the coverage ratio of the peaks to corresponding sites in an untreated DNA control. To identify the optimal minimal coverage ratio, we varied the ratio and calculated the number of identified sites at each ratio value. As shown in **Figure 2b**, as the coverage ratio increased from 2 to 7, the number of identified sites decreased from 92% to 59% of 2,681 expected (sensitivity in **Fig. 2b**), while the accuracy (identified sites/expected sites; specificity in **Fig. 2b**) only increased from 98% to 99.5%. To maximize sensitivity in this instance, we chose a coverage ratio of 2 as the standard data mining parameter in Nick-seq, which led to the identification of 2,462 (97.5%) of the predicted Nb.BsmI sites. Another 1% of the called sites (27 sites) occurred in sequences that differed from the consensus

by one nucleotide (**Supplementary Table 2**). These sites showed lower sequencing coverage than the consensus sites (e.g., 75 vs 1318 on average) and likely represent "star" activity of the Nb.BsmI nicking enzyme. In another validation experiment, Nick-seq was found to lack 3'-end biases. Nicking of *E. coli* DNA with the nicking endonuclease Nb.BsrDI, which cuts at N/CATTGC motifs, revealed no evident 3'-end sequence bias for DNA break site detection (**Supplementary Table 3)**. Thus, the Nick-seq method showed high accuracy and sensitivity for single-nucleotide detection of DNA strand breaks on a genome scale.

The validated Nick-seq method was next applied to map the naturally-occurring phosphorothioate (PT) modifications of DNA in the *Salmonella enterica* serovar Cerro 87 genome. Here we used the well-established approach of specifically oxidizing PTs to produce DNA strand breaks[9] (**Fig. 2c**). We had previously established by single-molecule real-time (SMRT) sequencing that PTs occurred as bistranded modifications at 10-15% of the GAAC/GTTC motifs in *S. enterica*[9]. Nick-seq recognized 12,239 PT sites in the genome (**Fig. 2c, Supplementary Table 4**), of which 11,684 (96%) occurred at the expected G_{PS}AAC and G_{PS}TTC, with 8,568 (73%) modified on both strands and 27% modified on either the GAAC or GTTC strand (**Fig. 2c**). These results agree with our previous observations using an orthogonal sequencing method[9]. In addition to the GAAC/GTTC motifs, Nick-seq also revealed less abundant PT modifications at G_{PS}TAC (168), G_{PS}ATC (30), and G_{PS}AAAC or G_{PS}AAAAC (19), with half of the G_{PS}TAC and G_{PS}ATC sites modified on only one strand (**Supplementary Table 4**). These results indicate that Nick-seq has a higher sensitivity to detect rare PT modification events than other sequencing methods.[9, 10]

With the precision, accuracy, and sensitivity of the Nick-seq method established with DNA modifications with defined locations, we next sought to apply the method to DNA

modifications not previously subjected to genomic mapping: abasic sites induced by oxidative stress. Apurinic and apyrimidinic (AP or abasic) sites represent one of the most prevalent and toxic forms of DNA damage, blocking DNA replication and transcription[11, 12]. AP sites can arise as DNA repair intermediates in the base excision repair (BER) pathway, in which the damaged bases are excised by glycosylase and the resulting AP sites are then cleaved by AP endonucleases[13]. AP sites can also arise by oxidation of DNA on both the nucleobase and 2'-deoxyribose moietities[14, 15]. In addition, demethylation of m^5C epigenetic modification has been proposed to lead to an abasic site.[16] In spite of the important roles of AP sites, little is known about their formation, persistence, and distribution within genomic DNA. As shown in **Figure 2d**, we used Nick-seq to profile AP sites arising from exposure of *E. coli* to H_2O_2 at a non-lethal dose of 0.2 mM (LD_{50} ~5 mM). The AP sites in purified genomic DNA were expressed as strand breaks using endonuclease IV (Endo IV), an AP endonuclease capable of cleaving both native and oxidized AP sites[14, 15]. Nick-seq identified 1,519 Endo IV-sensitive sites in genomic DNA from H_2O_2-treated cells, as well as 82 sites in the endogenous plasmid, with an unexposed control showing 11 and 8 sites, respectively (**Fig. 2d, Supplementary Tables 5, 6**). Considering the nucleobase precursor of the abasic site, there was a weak preference for thymine (33%) followed by adenine (25%), cytosine (24%), and guanine (18%) in genomic DNA, with a similar distribution in the plasmid (**Supplementary Tables 5, 6**). However, there was a more pronounced sequence context effect. Analysis of the 15 bp up- and down-stream of the AP sites revealed a strong preference for cytosine (47%) at -1 posioin of the AP sites. As **Fig. 2d** shows, the distribution of AP sites on the plasmid was also non-random, with clustering in three regions related to DNA replication and transcription: the F1 origin, pUC origin, and Amp^R gene (**Supplementary Table 7**). The phenomenon of AP site clustering at DNA replication sites has

been observed by immunostaining in cells[17], suggesting that the transcriptionally active and single-strand DNA are vulnerable to AP site formation during oxidation stress. To test this model, we analyzed the distribution of AP sites in the *E. coli* genome relative to landmarks including origins of replication (OriC), coding sequences, and non-coding sequences. The average AP frequency across the genome was 0.32 AP sites per 1 kb (1,519 AP per 4,686,000 bp), while, in the 20 kb region around OriC, there were 0.70 Ap sites per 1 kb. Nick-seq also revealed 1401 AP sites in the 4.1×10^6 bp coding sequence region (0.34 AP/1 kb), and 118 AP sites (0.20 AP/1 kb) in the 0.58×10^6 bp non-coding region. These results suggest a preferential for AP sites formation in regions of the genome undergoing replication or transcription in *E. coli* under H_2O_2 stress.

Nick-seq thus provides an efficient, label-free approach to quantifying the landscape of DNA damage and modifications in cells, with applications in DNA damage and repair, epigenetics, restriction-modification systems, and DNA metabolism.

Methods – see the online Methods Section

Author Contributions

P.C.D. and B.C. conceived and designed the Nick-seq method. B.C., X.W., H.W. performed the Nick-seq experiments. P.C.D, B.C., X.W., and J.Z. constructed the bioinformatics pipeline. M.S.D., C.G., D.Y., and L.W., contributed vital reagents. P.C.D and B.C. wrote the draft manuscript. All authors discussed the results and contributed to the final manuscript.

Competing Interests Statement

B.C., M.S.D., and P.C.D are co-inventors on a PCT patent (PCT/US2019/013714) relating to the published work.

Data Storage

Sequencing data has been deposited in NCBI's GEO database under the accession numbers GSE138070, GSE138173, GSE138476.

References

1. Roos, W.P., Thomas, A.D. & Kaina, B. *Nat Rev Cancer* **16**, 20-33 (2016).
2. Chen, Y. et al. *Chem Soc Rev* **46**, 2844-2872 (2017).
3. Li, Q., Hermanson, P.J. & Springer, N.M. *Methods Mol Biol* **1676**, 185-196 (2018).
4. Hu, J., Selby, C.P., Adar, S., Adebali, O. & Sancar, A. *J Biol Chem* (2017).
5. Ding, Y., Fleming, A.M. & Burrows, C.J. *J Am Chem Soc* **139**, 2569-2572 (2017).
6. Clark, T.A., Spittle, K.E., Turner, S.W. & Korlach, J. *Genome Integr* **2**, 10 (2011).
7. Schibel, A.E. et al. *J Am Chem Soc* **132**, 17992-17995 (2010).
8. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. *Biotechniques* **30**, 892-897 (2001).
9. Cao, B. et al. *Nat Commun* **5** (2014).
10. Li, J. et al. *PLoS Genet* **15**, e1008026 (2019).
11. Clauson, C.L., Oestreich, K.J., Austin, J.W. & Doetsch, P.W. *Proc Natl Acad Sci U S A* **107**, 3657-3662 (2010).

12. Galhardo, R.S., Almeida, C.E., Leitao, A.C. & Cabral-Neto, J.B. *J Bacteriol* **182**, 1964-1968 (2000).

13. Dianov, G.L., Sleeth, K.M., Dianova, II & Allinson, S.L. *Mutat Res* **531**, 157-163 (2003).

14. Xu, Y.J., Kim, E.Y. & Demple, B. *J Biol Chem* **273**, 28837-28844 (1998).

15. Greenberg, M.M., Weledji, Y.N., Kim, J. & Bales, B.C. *Biochemistry* **43**, 8178-8183 (2004).

16. Wu, X. & Zhang, Y. *Nat Rev Genet* **18**, 517-534 (2017).

17. Chastain II, P.D. et al. *FASEB J* **24**, 3674-3680 (2010).

Figure Legends

Figure. 1. Nick-seq library preparation (a) and data analysis workflow (b).

Figure. 2. **Nick-seq validation and application**. (**a**) Mapping of single-strand breaks produced by the site-specific endonuclease, Nb.BsmI, in *E. coli* genomic DNA. The middle panel shows a representative view of the distribution of sequencing reads in one region of the genome. Red and green peaks mark reads mapped to forward and reverse strands of the genome, respectively. The region surrounding one peak is amplified in the lower panel, with read pile ups for the TdT and NT sequencing converging on the site of the strand break and thus the modification site. (**b**) The Nb.BsmI mapping data were used to define data mining parameters for the accuracy and sensitivity of the Nick-seq method. As the coverage ratio (**b**) is increased from 2 to 7, the number of identified sites decreases from 92% to 59% of the 2,681 site expected (sensitivity), while the accuracy (identified sites/expected sites; specificity) only increased from 98% to 99.5%. In

general, higher ratios yield greater accuracy but lower sensitivity. A coverage ratio of 2 was selected for subsequent work since it captured >98% of the Nb.BsmI sites. (**c**) Mapping PT modifications across the genome of *S. enterica* serovar Cerro 87 by Nick-seq. Isolated genomic DNA was treated with iodine to convert PTs to strand breaks, followed by dephosphorylation of the break ends. The DNA was then subjected to Nick-seq for library preparation. The resulting sequencing data validated the previously observed GAAC/GTTC consensus and revealed additional PT modification motifs occurring at much lower frequency. (**d**) Application of Nick-seq to quantify abasic sites generated by oxidative stress (H_2O_2 exposure) in *E. coli*. Cells were exposed to a sublethal dose of H_2O_2 (0.2 mM) and isolated DNA treated with EndoIV to convert abasic sites to strand breaks for subsequent Nick-seq mapping. Analysis of the abasic site map on plasmid reveals clustering of H_2O_2-induced abasic sites in three regions related to DNA replication and transcription: the F1 origin, pUC origin, and Amp^R gene.

Supplementary Information

Supplementary Figure 1. Mapping Nb.BsrDI induced DNA strand break sites in *E.coli* genomic DNA by Nick-seq. (a) The Nb.BsrDI cleavage motif derived from Nick-seq data. (**b**) A Venn diagram depicting the overlap of Nick-seq detected sites and Nb.BsrDI motif sites in the *E.coli* genome.

Supplementary Figure 2. Flanking sequence frequency analysis of H_2O_2-induced EndoIV-specific DNA damage sites based on Nick-seq data for genomic DNA (**a**) and plasmid (**b**) from cells exposed to a sublethal H_2O_2 dose. Site 0 represents a Nick-seq detected site.

Supplementary Figure 3. Detection of H_2O_2-induced EndoIV-specific DNA damage sites on genomic DNA (a) and an endogenous plasmid (b) in *E. coli* by Nick-seq.

Supplementary Figure 4. Distributions of H_2O_2-induced EndoIV-specific DNA damage sites on *E.coli* genomic DNA and plasmid. Outward from the center, circles represent: 0 and 0.2 mM H_2O_2 induced EndoIV-specific DNA damage sites.

Supplementary Table 1: Statistical analysis of paired-end DNA sequencing reads from Nick-seq.

Supplementary Table 2: Nb. BsmI sites identified by Nick-seq on the *E.coli* genome.

Supplementary Table 3: Nb. BsrDI sites identified by Nick-seq on the *E.coli* genome.

Supplementary Table 4: Phosphorothioate sites identified by Nick-seq on Salmonella genome.

Supplementary Table 5: AP sites identified by Nick-seq on *E.coli* genome.

Supplementary Table 6: AP sites identified by Nick-seq on plasmid in *E.coli*.

Supplementary Table 7: Regional distribution of AP sites identified by Nick-seq.

Online Methods Section

Materials. Nicking endonuclease Nb. BsmI, Nb. BspQI, Nb. BsrDI, Endonuclease IV, DNA polymerase I, OneTaq DNA polymerase, dNTPs, Nci I, Exonuclease III and $RecJ_f$ were purchased from New England Biolabs. All DNA oligos were synthesized by Integrated Device Technology, Inc. (IDT). ddNTPs and alpha-thio-dNTPs were purchased from TriLink BioTech. Agilent Bioanalyzer 2100 was used for size analysis of DNA fragments. Other chemicals were of molecular biology grade.

Cell growth and preparation of DNA. The PT-containing strain *Salmonella enterica* serovar Cerro 87 and its genomic DNA were prepared as described previously[9]. *E.coli* DH10B was used for nicking enzyme and H_2O_2-induced DNA damage mapping studies. A single colony of *E. coli* DH10B was grown in 5 mL LB medium overnight at 37 °C. 1mL cells were harvested by centrifuge at ambient temperature (unless indicated otherwise). Cells were resuspended and diluted with fresh 10 LB medium to a starting optical density at 600 nm (OD600) of 0.1, followed by growth at 37°C, 230 rpm until OD600=0.8 for DNA extraction or H_2O_2 treatment. 10 μL diluted H_2O_2 solution were added to the culture with a final concentration 0.1, 0.5, 1 and 2 mM. As un-exposed control, 10 μL sterile water was used instead of H_2O_2. After sitting at ambient temperature for 30 min, 10 μL of the cells were used for lethal dose (LD) analysis by counting the colony formation unit on LB agar plate. All the rest cells were harvested for DNA extraction with OMEGA bacterial genomic DNA or plasmid isolation kit by following the manufacture's protocol.

Mapping of modification/damage sites on DNA by NT-dependent method. These studies were initiated by random fragmentation of purified genomic DNA (1 μg) in each of three separate digestions with NciI, or HindIII and XhoI, or SalI, XbaI and NdeI. RNase A was also added to each reaction to remove trace of contaminating RNA. After digestion, the DNA was purified using a Qiagen PCR Purification Kit. The three purified DNA samples were mixed for the blocking step. Blocking of pre-existing strand break sites was achieved in a reaction mixture (40 μl) containing 4 μl of reaction buffer (NEBcutsmart buffer), 1 μl of shrimp alkaline phosphatase (NEB), and 1 μg of template genomic DNA, with incubation at 37 °C for 30 min to remove phosphate at 3' end of the strand breaks. The phosphatase was then inactivated by heating at 70°C for 10 min. After cooling, 2 μl of ddNTPs (2.5 mM each, TriLink) and 1 μl of

DNA polymerase I (10 U, NEB) was added to the reaction with incubation at 37 °C for 40 min to block any pre-existing strand break sites. Shrimp alkaline phosphatase (1 μl) was then added at 37 °C for 30 min to degraded excess ddNTPs and the reaction was terminated by heating at 75 °C for 10 min. Following de-salting using a DyeEx column (QIAGEN), the DNA was ready for one of the following nick creation or conversion procedures.

Nicking *E. coli* DH10B genomic DNA with Nb. BsmI and Nb. BsrDI was accomplished in a 50 μl reaction mixture containing 1 μl Nb. BsmI (10 U, NEB), 1 μl Nb. BsrDI (10 U, NEB), 1 μg genomic DNA, and 1X NEBcutsmart buffer incubated at 65 °C for 1 h. The reaction was terminated by heating at 80 °C for 20 min and cooled down to 4 °C at the rate of 0.1 °C/s. The reaction product was used for NT- or TdT- reactions as described below with no further purification.

For mapping PT modifications, 40 μl of blocked DNA from *S. enterica* was mixed with 5 μl of dibasic sodium phosphate buffer (500 mM, pH 9.0) and 2 μl of iodine solution (0.1 N, FLUKA). After incubation at 65 °C for 5 min and cooling to 4 °C, the reaction product was purified using a DyeEX column (QIAGEN) to remove salts and iodine. The purified product was treated with shrimp alkaline phosphatase by adding 5 μl of NEBcutsmart buffer and 1 μl of phosphatase to remove 3'-phosphates arising from iodine cleavage. After incubation at 37 °C for 20 min and 75 °C for another 10 min, the product was kept on ice for the following NT- or TdT- reactions with no additional purification.

For mapping H_2O_2-induced AP sites, genomic DNA was extracted from H_2O_2-treated *E. coli* and the AP sites converted to strand breaks in a 50 μl reaction mixture containing 1 μg genomic DNA, endonuclease IV (20 U), and 1X NEBcutsmart buffer, with incubation at 37 °C for 60 min. The reaction mixture was then kept on ice and used for the following NT- or TdT-

reactions with no further inactivation or purification.

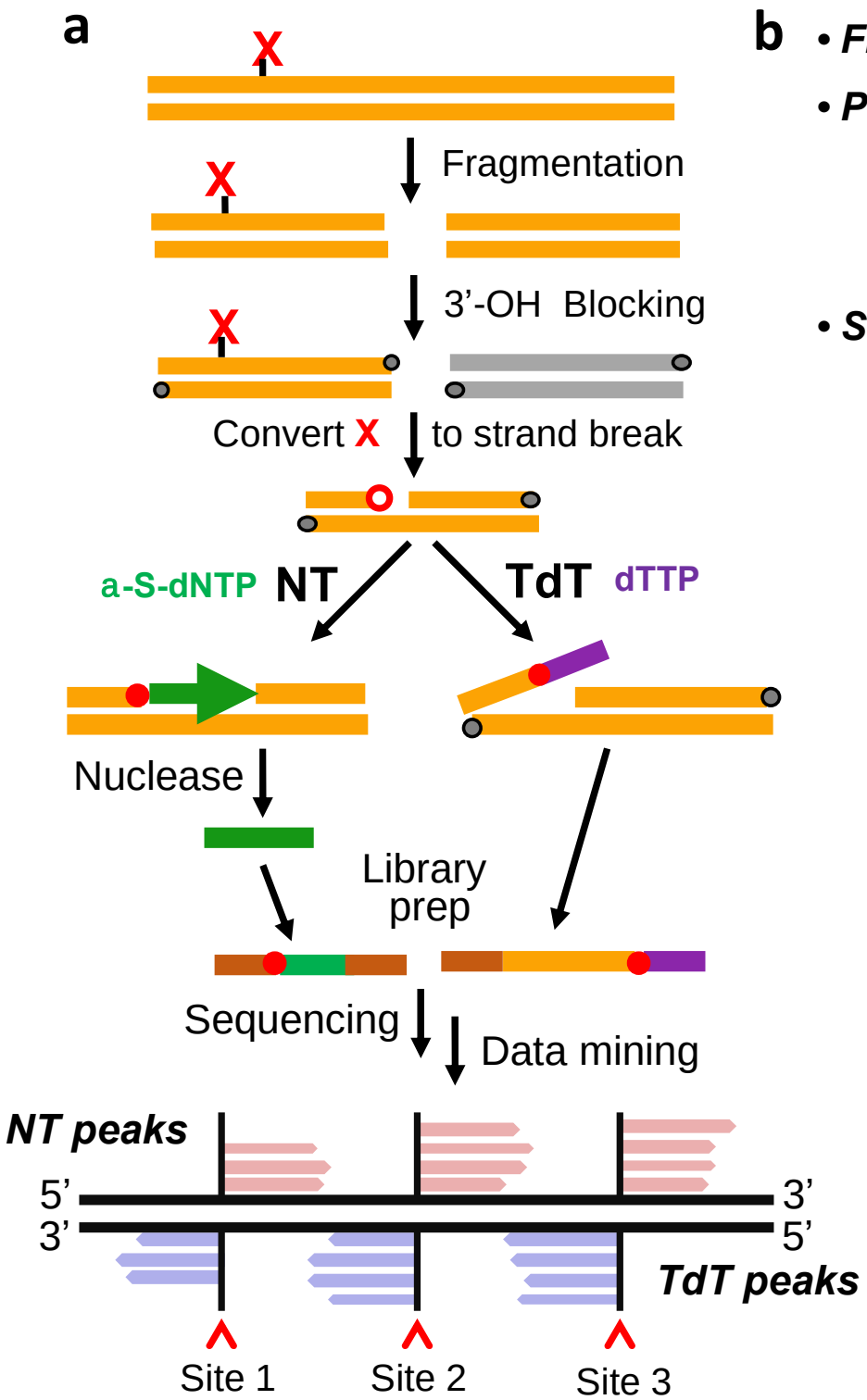
Following splitting of the nicked sample into two portions for NT- and TdT- reactions, the NT- reaction was achieved by further splitting the DNA sample into two parts: one for NT-reaction and the other as a negative control. NT-reaction was performed in a 50 μl reaction system containing 2 μl of α-thio-dNTPs (2.5 mM each, TRILINK), 1X of NEBcutsmart buffer, 2 μl of DNA polymerase I, and the DNA template. The negative control consisted of H_2O instead of DNA polymerase I. The reaction mixture was incubated at 15 °C for 90 min and then terminated by heating at 75 °C for 20 min. The product was ready of template DNA digestion after the purification by DyeEx. The template DNA digestion reaction was performed in a 50 μl reaction system containing 200 U of exonuclease III, 5 μl NEBcutsmart buffer and DNA sample by incubating at 37 °C for 60 min. The DNA was then denatured by heating at 95 °C for 3 min and crashing on ice. $RecJ_f$ (60 U) was then added to the reaction mixture with incubation at 37 °C for 60 min. For some DNA samples with high complexity of structure and/or modifications, digestion with an additional 60 units of $RecJ_f$ might be necessary. After digestion, the enzymes were inactivated by incubation at 80 °C for 10 min. The DNA product was then purified using a Zymo Oligo Clean & Concentrator kit (Zymo) following the manufacturer's protocol. The purified product was ready for Illumina library preparation.

Illumina library preparation was performed by the Clontech SMART ChIp-seq kit (Clontech) by following the manufacturer's protocol. 12 cycles were used in the final step of PCR amplification. The PCR product of each sample was combined with its corresponding negative control and then size selected using AMPure XP beads (NEB). The purified library was submitted to Illumina NextSeq 500 instrument for 75 bp paired-end sequencing.

Mapping of modification and damage sites by TdT-dependent method. The steps of DNA fragmentation, blocking, and nick conversion are the same as described above in NT-method.100 ng of the nick-converted DNA was denatured by heating at 95 °C for 3 min in 20 µl of H_2O, followed by adding A poly(T) tail to the ssDNA in a 30 µl reaction system containing 3 µl DNA SMART buffer (Clontech), 1 µl Terminal Deoxynucleotidyl Transferase (TdT, Clontech) and 1 µl DNA SMART T-Tailing Mix (Clontech) by incubating at 37 °C for 20 min and terminating the reaction at 70 °C for 10 min. The primer annealing and template switching reaction was then performed with the Clontech SMART ChIp-seq kit (Clontech) by following the manufacturer's protocol. The final step of PCR was perfomed using the Illumina primers provided in ChIp-seq kit and 12 cycles were used for amplification. The PCR product of each sample, with unique sequencing barcode, was combined with its corresponding negative control and then size selected using AMPure XP beads (NEB). The purified library was submitted to Illumina NextSeq 500 instrument for 75 bp paired-end sequencing.

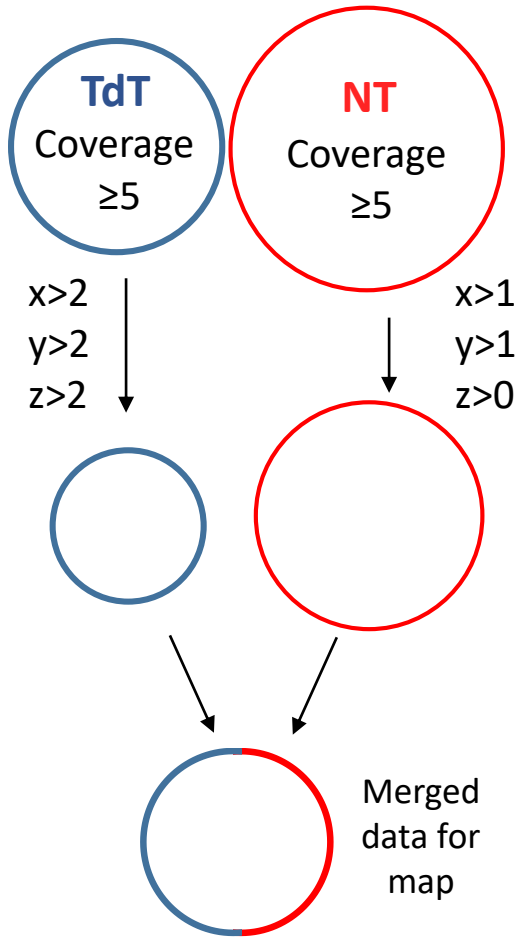
Data analysis. Sequencing results were processed on the Galaxy web platform (https://usegalaxy.org/). Initially, the paired end reads were pre-processed by trim galore to remove adapters and the first 3 bp on the 5' of read 1 were also trimmed. All the reads were aligned to the corresponding genome using Bowtie 2. A custom method for peak calling of sequencing data was developed with BamTools, BEDTools and Rstudio. Briefly, the bam results were filtered based on R1 (selected for NT data) or R2 (selected for TdT data). The 5' coverage (experiment sample) or full coverage (controls) on each position were calculated based on the filtered bam result by BEDTools (positive and negative strand separately). The coverage of position N (sample)/ coverage of position N-1(sample), coverage of position N(sample)/

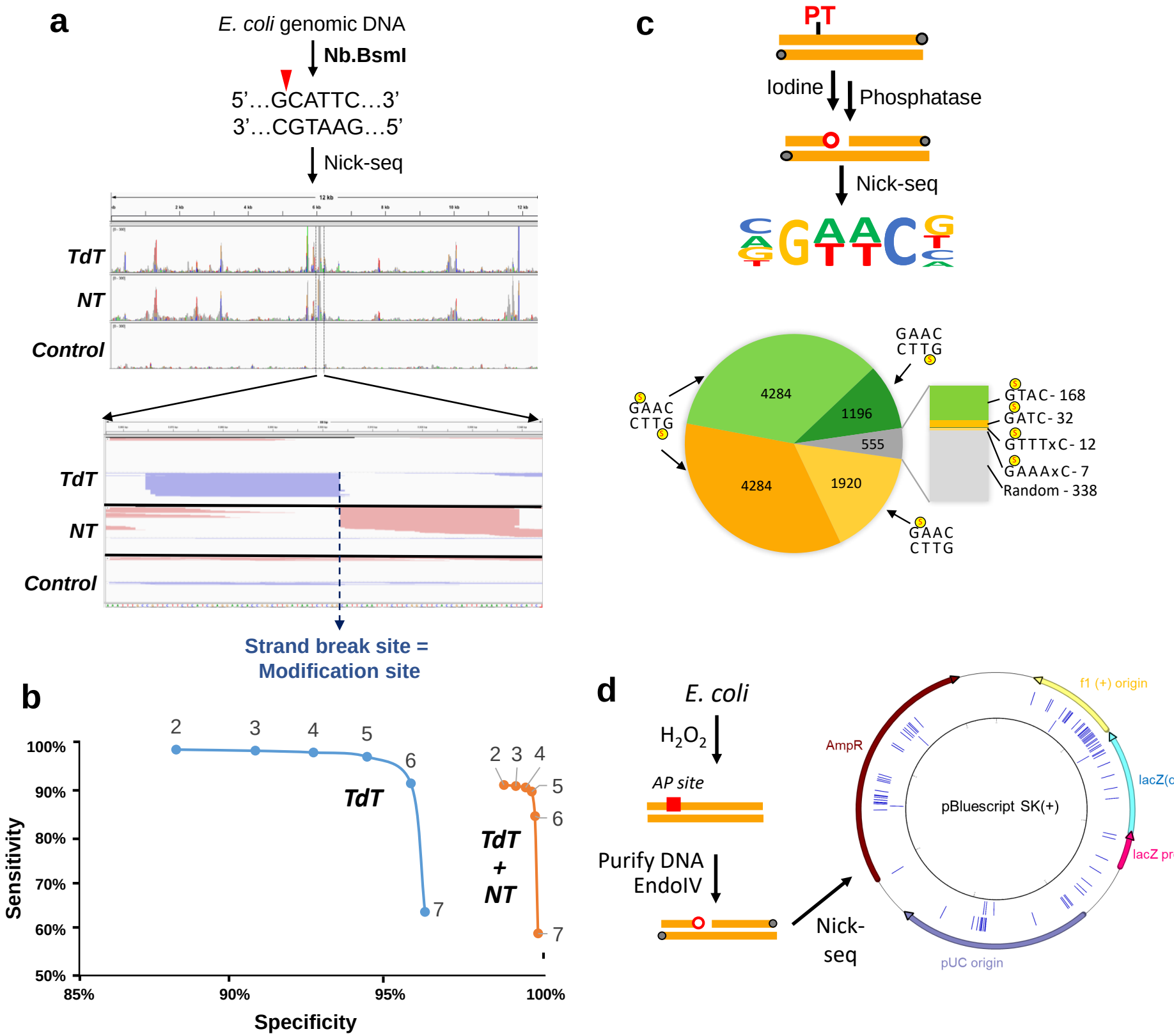
coverage of position N+1(sample) and coverage of position N(sample)/ coverage of position N(control) ratio was calculated on each position by RStudio for modification site calling.



b

- ***Filter raw reads for seq coverage ≥5***
- ***Parameters for refining filtered reads***
  - x = coverage at N/coverage at N-1
  - y = coverage at N/coverage at N+1
  - z = coverage at N/coverage at N neg ctl
- ***Steps for refining the filtered reads***
  1. Filter NT data for x>1, y>1, z>0
  2. Filter TdT data for x>2, y>2, z>2
  3. Intersect step #1 and #2 datasets
  4. Merged data used for final map




a
E. coli genomic DNA
Nb.BsmI
5'…GCATTC…3'
3'…CGTAAG…5'
Nick-seq
TdT
NT
Control
TdT
NT
Control
Strand break site =
Modification site
b
Sensitivity
Specificity
100%
90%
80%
70%
60%
50%
85%
90%
95%
100%
2
3
4
5
6
7
TdT
TdT
+
NT
c
PT
Iodine
Phosphatase
Nick-seq
GAAC
CTTG
4284
1196
555
1920
4284
GTAC - 168
GATC - 32
GTTTxC - 12
GAAAxC - 7
Random - 338
d
E. coli
H2O2
AP site
Purify DNA
EndoIV
Nick-
seq
AmpR
f1 (+) origin
lacZ(α
lacZ pr
pBluescript SK(+)
pUC origin