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

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

[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] State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic and Developmental Sciences, and School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, Shanghai, 200030, China

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

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

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

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

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

[†] These authors contributed equally to this work.

Corresponding authors: B.C. (ORCID ID: 0000-0002-7011-1676) (caobo@qfnu.edu.cn), D.Y. (dlyou@sjtu.edu.cn), and P.C.D. (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 next-generation 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)

Genomic mapping of specific DNA modifications[1] and damage[2] can be achieved with methods such as bisulfite sequencing for 5-methylcytidine[3], chromatin immunoprecipitation (ChIP) coupled with next-generation sequencing (NGS)[4,5], and single-molecule real-time (SMRT)[6] and nanopore[7] sequencing. However, all are limited to specific modifications or suffer from low sensitivity and specificity. Here we describe Nick-seq for highly sensitive quantitative genomic mapping of any type of DNA modification or damage that can be converted to a strand-break. As shown in **Figure 1a**, purified genomic DNA is subjected to sequencing-compatible fragmentation and the resulting 3'-OH ends are blocked with dideoxyNTPs. The DNA modification is then converted to a strand-break by enzyme or chemical treatment, followed by capture of the 3'- and 5'-ends of resulting strand-breaks using two complementary strategies. One portion of DNA 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 exonuclease III and $RecJ_f$. The purified PT-protected fragment is used to generate an

49 NGS library with the modification of interest positioned at the 5'-end of the PT-labeled fragment.
50 A second portion of the same DNA sample is used for terminal transferase (TdT)-dependent
51 poly(dT) tailing of the 3'-end of the strand-break, with the tail used to create a sequencing library
52 by reverse transcriptase template switching[8]. Subsequent NGS positions the modification of
53 interest 5'-end of the poly(dT) tail.

54 The workflow for sequencing data processing (**Fig. 1b**) uses the NT-derived reads as the
55 primary dataset for developing a rough modification map, with TdT-derived reads as
56 complementary corrective data. This hybrid approach exploits the fact that NT is agnostic to the
57 base identify at the damage site but generates a high background of false positive sites, while TdT
58 cannot be used with modifications occurring at dT due to loss of the poly(dT) tail during data
59 analysis. The TdT reads are used to correct NT false-positive reads. For example, if NT maps a
60 strand-break at T 1000 in the genome, then TdT reads are examined for a strand-break at position
61 999 and 1001. This 1 position shift accommodates poly(dT) tail removal during data processing
62 and validates the NT map. If TdT does not call a strand-break at 999 or 1001, then the NT result
63 is considered a false positive. In other cases, if the NT-detected site occurs at a G, C, or A, then
64 TdT valdiates the same site. The use of both methods increases the sensitivity and specificity of
65 the resulting map.

66 We validated Nick-seq by mapping DNA single-strand-breaks caused by the site-specific
67 endonuclease, Nb.BsmI, at the 2,681 G/CATTC motifs in the *E.coli* genome. Purified DNA was
68 treated with Nb.BsmI and the Nick-seq-processed library sequenced using the Illumina NextSeq
69 platform with an average of 10^7 raw sequencing reads for each sample (**Fig. 2b**). Paired-end
70 sequencing confirmed that >80% of reads uniquely aligned to the *E.coli* genome (**Supplementary**
71 **Table 1**). For subequent reads enrichment (**Fig. 1b**), we calculated position-wise coverage values

using the 5'-end of sequencing reads (NT read 1, TdT read 2) 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 (**Fig. 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"), while the accuracy (identified sites/expected sites; "specificity") only increased from 98% to 99.5%. To maximize sensitivity, we chose a coverage ratio of 2, which allowed identification of 2,462 (97.5%) of the predicted Nb.BsmI sites. Another 1% of called sites (27) occurred in sequences differing from the consensus by one nucleotide (**Supplementary Table 2**). These sites showed lower average sequencing coverage (75 vs 1318) and likely represent Nb.BsmI "star" activity. Another validation experiment with Nb.BsrDI, which cuts at N/CATTGC, showed no evident 3'-end sequence bias for DNA break site detection (**Supplementary Table 3)**. Thus, Nick-seq showed high accuracy and sensitivity for single-nucleotide genomic mapping of DNA strand-breaks.

The validated Nick-seq was applied to map the naturally-occurring phosphorothioate (PT) DNA modifications in *Salmonella enterica* serovar Cerro 87, using iodine to oxidize PTs to produce DNA strand-breaks[9] (**Fig. 2c**). We previously established by 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 (**Fig. 2c, Supplementary Table 4**), of which 11,684 (96%) occurred at G_{PS}AAC/G_{PS}TTC, with 8,568 (73%) modified on both strands and 27% modified on one strand (**Fig. 2c**). This agrees with our previous observations using an orthogonal sequencing method[9]. In addition to GAAC/GTTC motifs, Nick-seq also revealed less abundant PTs at G_{PS}TAC

95 (168), G_{PS}ATC (30), and G_{PS}AAAC or G_{PS}AAAAC (19), with half of G_{PS}TAC and G_{PS}ATC sites
96 modified on only one strand (**Supplementary Table 4**). These results indicate that Nick-seq has a
97 higher sensitivity to detect rare PT modifications than other methods.[9, 10]

98 Finally, we applied Nick-seq to DNA modifications not previously subjected to genomic
99 mapping: oxiatively-induced abasic sites. Apurinic and apyrimidinic (AP) sites represent a
100 prevalent and toxic form of DNA damage that blocks DNA replication and transcription[11, 12]. AP
101 sites arise as intermediates in base excision DNA repair, in which damaged bases are excised by
102 glycosylases and the resulting AP sites cleaved by AP endonucleases[13]. AP sites can also arise
103 by oxidation of DNA on both the nucleobase and 2'-deoxyribose moietities[14, 15], as well as by
104 demethylation of m^5C epigenetic marks.[16] In spite of the importance of AP sites, little is known
105 about their formation, persistence, and distribution in genomic DNA. Here we used Nick-seq to
106 profile AP sites in *E. coli* exposed to H_2O_2 at a non-lethal dose of 0.2 mM (LD_{50} ~5 mM) (**Fig.**
107 **2d**). Following DNA purification, AP sites were expressed as strand-breaks using endonuclease
108 IV (EndoIV), which cleaves both native and oxidized AP sites[14, 15]. Nick-seq identified 1,519
109 EndoIV-sensitive sites, as well as 82 sites in the endogenous plasmid, with an unexposed control
110 showing 11 and 8 sites, respectively (**Fig. 2d, Supplementary Tables 5, 6**). Considering the
111 nucleobase precursor of the AP site, there was a weak preference for thymine (33%) followed by
112 adenine (25%), cytosine (24%), and guanine (18%), with a similar distribution in the plasmid
113 (**Supplementary Tables 5, 6**). This suggests that H_2O_2-derived DNA oxidizing agents either do
114 not selectively oxidize guanine as predicted[17] or that the predominant form of damage is DNA
115 sugar oxidation. However, there was a more pronounced sequence context effect. Analysis of 15
116 bp up- and down-stream of the AP sites revealed a strong preference for cytosine (47%) at -1
117 relative to the AP sites. The distribution of AP sites on the plasmid was also non-random (**Fig.**

118 **2d**), with clustering in three regions related to DNA replication and transcription: the F1 origin,
119 pUC origin, and Amp^R gene (**Supplementary Table 7**). AP site clustering near DNA replication
120 sites was observed previously by immunostaining[18], suggesting that the transcriptionally active
121 and single-strand DNA are vulnerable to oxidatively-induced AP sites. We tested this by analyzing
122 the distribution of AP sites in the *E. coli* genome relative to origins of replication (OriC), coding
123 sequences, and non-coding sequences. While there was an average of 0.32 Aps/kbp (1,519 AP per
124 4,686,000 bp), the 20 kb region around OriC showed 0.70 Aps/kbp. Nick-seq also revealed 1401
125 APs in the 4.1×10^6 bp coding sequence region (0.34 AP/kbp), and 118 APs (0.20 AP/kbp) in the
126 0.58×10^6 bp non-coding region. These results suggest a preference for AP sites in DNA
127 undergoing replication or transcription during H_2O_2 stress.
128 Nick-seq thus provides an efficient, label-free approach to quantive mapping of DNA
129 damage and modifications in genomes, with applications in DNA damage and repair, epigenetics,
130 restriction-modification systems, and DNA metabolism.
131
132 **Methods – see the online Methods Section**
133
134 **Acknowledgements**
135 The authors thank the MIT BioMicro Center, MIT Center for Environmental Health Science,
136 Singapore-MIT Alliance for Research and Technology (SMART) for use of their facilities, and
137 funding support from the National Natural Science Foundation of China (31630002), National
138 Science Foundation of the USA (CHE-1709364), National Research Foundation of Singapore
139 through the SMART Infectious Disease and Antimicrobial IRGs, National Institute of
140 Environmental Health Sciences (P30-ES002109), and Fundamental Research Funds for the

Central Universities of China (2015306020202). X.W. was supported by a fellowship from the China Scholarship Council (201606270163).

Author Contributions

P.C.D. and B.C. designed Nick-seq. B.C., X.W., H.W. performed 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. 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) and US Patent (US 2019/0284624 A1) relating to the published work.

Data Availability

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

Software and Code Availability

Custom scripts for processing the sequencing data are described in Methods and are available at https://github.com/BoCao2019/Nick-seq: .gitignore, NT_negative_strand.R, NT_positive_strand.R, TdT_negative+NT_positive.R, TdT_negative_strand.R, TdT_positive+NT_negative.R, and TdT_positive_strand.R.

164 **References**

165 1. Chen, Y. et al. *Chem Soc Rev* **46**, 2844-2872 (2017).

166 2. Roos, W.P., Thomas, A.D. & Kaina, B. *Nat Rev Cancer* **16**, 20-33 (2016).

167 3. Li, Q., Hermanson, P.J. & Springer, N.M. *Methods Mol Biol* **1676**, 185-196 (2018).

168 4. Hu, J., Selby, C.P., Adar, S., Adebali, O. & Sancar, A. *J Biol Chem* (2017).

169 5. Ding, Y., Fleming, A.M. & Burrows, C.J. *J Am Chem Soc* **139**, 2569-2572 (2017).

170 6. Clark, T.A., Spittle, K.E., Turner, S.W. & Korlach, J. *Genome Integr* **2**, 10 (2011).

171 7. Schibel, A.E. et al. *J Am Chem Soc* **132**, 17992-17995 (2010).

172 8. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. *Biotechniques* **30**, 892-
173 897 (2001).

174 9. Cao, B. et al. *Nat Commun* **5** (2014).

175 10. Li, J. et al. *PLoS Genet* **15**, e1008026 (2019).

176 11. Clauson, C.L., Oestreich, K.J., Austin, J.W. & Doetsch, P.W. *Proc Natl Acad Sci U S A*
177 **107**, 3657-3662 (2010).

178 12. Galhardo, R.S., Almeida, C.E., Leitao, A.C. & Cabral-Neto, J.B. *J Bacteriol* **182**, 1964-
179 1968 (2000).

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

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

182 15. Greenberg, M.M., Weledji, Y.N., Kim, J. & Bales, B.C. *Biochemistry* **43**, 8178-8183
183 (2004).

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

185 17. Dedon, P.C. & Tannenbaum, S.R. *Arch Biochem Biophys* **423**, 12-22 (2004).

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

187 **Figure Legends**

188

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

190

191 **Figure. 2**. **Nick-seq validation and application**. (**a**) Mapping single-strand breaks produced by
192 Nb.BsmI in *E. coli* genomic DNA. Middle panel: representative view of sequencing reads
193 distirbuted in one genomic region. Red and green peaks mark reads mapped to forward and reverse
194 strands of the genome, respectively. Lower panel: amplification of the region surrounding one peak,
195 with read pile ups for TdT and NT sequencing converging on the site of the strand-break. (**b**)
196 Nb.BsmI mapping data were used to define data mining parameters for accuracy and sensitivity of
197 Nick-seq. In general, higher ratios yield greater accuracy but lower sensitivity. (**c**) Mapping PTs
198 across the *S. enterica* genome by Nick-seq. (**d**) Application of Nick-seq to quantify abasic sites
199 generated by H_2O_2 exposure in *E. coli*.

200

201 **Supplementary Information**

202

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

207

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

211

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

214

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

218

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

220

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

222

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

224

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

226

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

228

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

230

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

232

233 **Online Methods Section**

234

235 **Materials.** Nicking endonuclease Nb. BsmI, Nb. BspQI, Nb. BsrDI, Endonuclease IV, DNA
236 polymerase I, OneTaq DNA polymerase, dNTPs, Nci I, Exonuclease III and $RecJ_f$ were purchased
237 from New England Biolabs. All DNA oligos were synthesized by Integrated Device Technology,
238 Inc. (IDT). ddNTPs and alpha-thio-dNTPs were purchased from TriLink BioTech. Agilent
239 Bioanalyzer 2100 was used for size analysis of DNA fragments. Other chemicals were of
240 molecular biology grade. All cell lines used in this work are readily available from the authors.

241 **Cell growth and preparation of DNA.** The PT-containing strain *Salmonella enterica*
242 serovar Cerro 87 and its genomic DNA were prepared as described previously[9]. *E.coli* DH10B
243 was used for nicking enzyme and H_2O_2-induced DNA damage mapping studies. A single colony
244 of *E. coli* DH10B was grown in 5 mL LB medium overnight at 37 °C. 1mL cells were harvested
245 by centrifuge at ambient temperature (unless indicated otherwise). Cells were resuspended and
246 diluted with fresh 10 LB medium to a starting optical density at 600 nm (OD600) of 0.1, followed

247 by growth at 37°C, 230 rpm until OD600=0.8 for DNA extraction or H_2O_2 treatment. 10 μL diluted
248 H_2O_2 solution were added to the culture with a final concentration 0.1, 0.5, 1 and 2 mM. As un-
249 exposed control, 10 μL sterile water was used instead of H_2O_2. After sitting at ambient temperature
250 for 30 min, 10 μL of the cells were used for lethal dose (LD) analysis by counting the colony
251 formation unit on LB agar plate. All the rest cells were harvested for DNA extraction with
252 OMEGA bacterial genomic DNA or plasmid isolation kit by following the manufacture's protocol.
253 **Mapping of modification/damage sites on DNA by NT-dependent method.** These
254 studies were initiated by random fragmentation of purified genomic DNA (1 μg) in each of three
255 separate digestions with NciI, or HindIII and XhoI, or SalI, XbaI and NdeI. RNase A was also
256 added to each reaction to remove trace of contaminating RNA. After digestion, the DNA was
257 purified using a Qiagen PCR Purification Kit. The three purified DNA samples were mixed for the
258 blocking step. Blocking of pre-existing strand-break sites was achieved in a reaction mixture
259 (40 μl) containing 4 μl of reaction buffer (NEBcutsmart buffer), 1 μl of shrimp alkaline
260 phosphatase (NEB), and 1 μg of template genomic DNA, with incubation at 37 °C for 30 min to
261 remove phosphate at 3' end of the strand-breaks. The phosphatase was then inactivated by heating
262 at 70°C for 10 min. After cooling, 2 μl of ddNTPs (2.5 mM each, TriLink) and 1 μl of DNA
263 polymerase I (10 U, NEB) was added to the reaction with incubation at 37 °C for 40 min to block
264 any pre-existing strand-break sites. Shrimp alkaline phosphatase (1 μl) was then added at 37 °C
265 for 30 min to degraded excess ddNTPs and the reaction was terminated by heating at 75 °C for
266 10 min. Following de-salting using a DyeEx column (QIAGEN), the DNA was ready for one of
267 the following nick creation or conversion procedures.
268 Nicking *E. coli* DH10B genomic DNA with Nb. BsmI and Nb. BsrDI was accomplished
269 in a 50 μl reaction mixture containing 1 μl Nb. BsmI (10 U, NEB), 1 μl Nb. BsrDI (10 U, NEB),

270 1 μg genomic DNA, and 1X NEBcutsmart buffer incubated at 65 °C for 1 h. The reaction was
271 terminated by heating at 80 °C for 20 min and cooled down to 4 °C at the rate of 0.1 °C/s. The
272 reaction product was used for NT- or TdT- reactions as described below with no further
273 purification.

274 For mapping PT modifications, 40 μl of blocked DNA from *S. enterica* was mixed with 5
275 μl of dibasic sodium phosphate buffer (500 mM, pH 9.0) and 2 μl of iodine solution (0.1 N,
276 FLUKA). After incubation at 65 °C for 5 min and cooling to 4 °C, the reaction product was purified
277 using a DyeEX column (QIAGEN) to remove salts and iodine. The purified product was treated
278 with shrimp alkaline phosphatase by adding 5 μl of NEBcutsmart buffer and 1 μl of phosphatase
279 to remove 3'-phosphates arising from iodine cleavage. After incubation at 37 °C for 20 min and
280 75 °C for another 10 min, the product was kept on ice for the following NT- or TdT- reactions with
281 no additional purification.

282 For mapping H_2O_2-induced AP sites, genomic DNA was extracted from H_2O_2-treated *E.*
283 *coli* and the AP sites converted to strand-breaks in a 50 μl reaction mixture containing 1 μg
284 genomic DNA, endonuclease IV (20 U), and 1X NEBcutsmart buffer, with incubation at 37 °C for
285 60 min. The reaction mixture was then kept on ice and used for the following NT- or TdT- reactions
286 with no further inactivation or purification.

287 Following splitting of the nicked sample into two portions for NT- and TdT- reactions, the
288 NT- reaction was achieved by further splitting the DNA sample into two parts: one for NT-reaction
289 and the other as a negative control. NT-reaction was performed in a 50 μl reaction system
290 containing 2 μl of α-thio-dNTPs (2.5 mM each, TRILINK), 1X of NEBcutsmart buffer, 2 μl of
291 DNA polymerase I, and the DNA template. The negative control consisted of H_2O instead of DNA
292 polymerase I. The reaction mixture was incubated at 15 °C for 90 min and then terminated by

293 heating at 75 °C for 20 min. The product was ready of template DNA digestion after the
294 purification by DyeEx. The template DNA digestion reaction was performed in a 50 µl reaction
295 system containing 200 U of exonuclease III, 5 µl NEBcutsmart buffer and DNA sample by
296 incubating at 37 °C for 60 min. The DNA was then denatured by heating at 95 °C for 3 min and
297 crashing on ice. $RecJ_f$ (60 U) was then added to the reaction mixture with incubation at 37 °C for
298 60 min. For some DNA samples with high complexity of structure and/or modifications, digestion
299 with an additional 60 units of $RecJ_f$ might be necessary. After digestion, the enzymes were
300 inactivated by incubation at 80 °C for 10 min. The DNA product was then purified using a Zymo
301 Oligo Clean & Concentrator kit (Zymo) following the manufacturer's protocol. The purified
302 product was ready for Illumina library preparation.

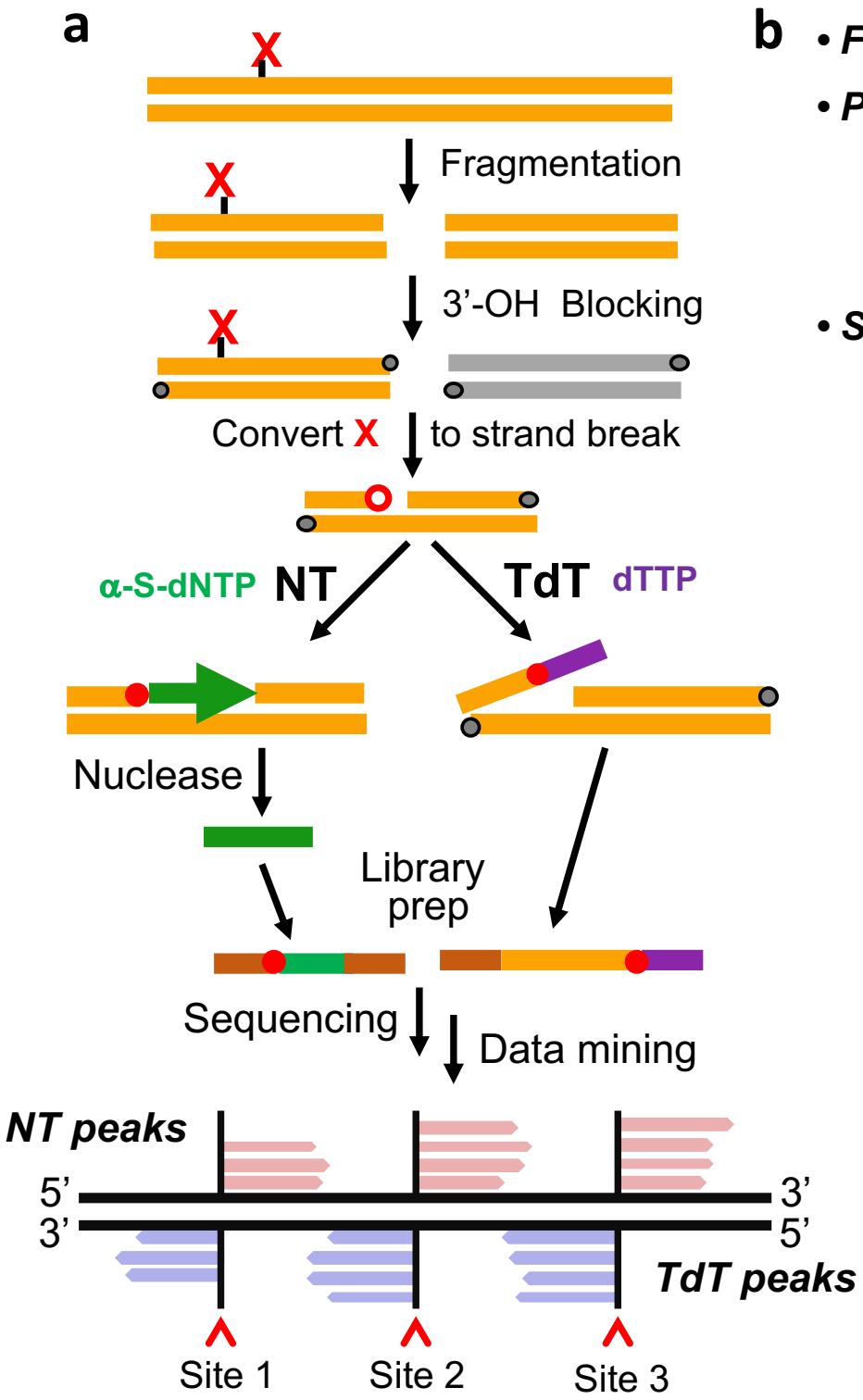
303 Illumina library preparation was performed by the Clontech SMART ChIp-seq kit
304 (Clontech) by following the manufacturer's protocol. 12 cycles were used in the final step of PCR
305 amplification. The PCR product of each sample was combined with its corresponding negative
306 control and then size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

308 **Mapping of modification and damage sites by TdT-dependent method**. The steps of
309 DNA fragmentation, blocking, and nick conversion are the same as described above in NT-
310 method.100 ng of the nick-converted DNA was denatured by heating at 95 °C for 3 min in 20 µl
311 of H_2O, followed by adding A poly(T) tail to the ssDNA in a 30 µl reaction system containing 3
312 µl DNA SMART buffer (Clontech), 1 µl Terminal Deoxynucleotidyl Transferase (TdT, Clontech)
313 and 1 µl DNA SMART T-Tailing Mix (Clontech) by incubating at 37 °C for 20 min and
314 terminating the reaction at 70 °C for 10 min. The primer annealing and template switching reaction
315 was then performed with the Clontech SMART ChIp-seq kit (Clontech) by following the

316 manufacturer's protocol. The final step of PCR was perfomed using the Illumina primers provided
317 in ChIp-seq kit and 12 cycles were used for amplification. The PCR product of each sample, with
318 unique sequencing barcode, was combined with its corresponding negative control and then size
319 selected using AMPure XP beads (NEB). The purified library was submitted to Illumina NextSeq
320 500 instrument for 75 bp paired-end sequencing.

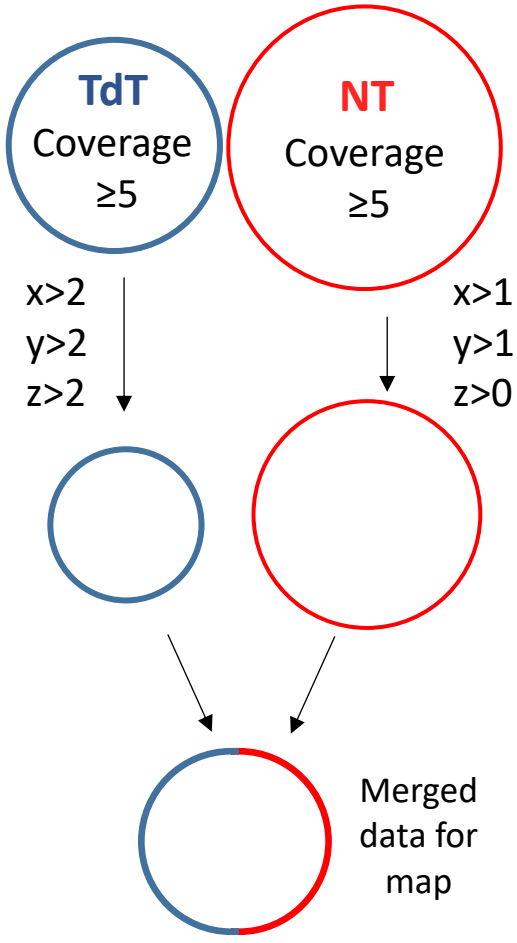
321 **Data analysis.** Sequencing results were processed on the Galaxy web platform
322 (https://usegalaxy.org/). Initially, the paired-end reads were pre-processed by Trim Galore! to
323 remove adapters, as well as trimming the first 3 bp on the 5' end of read 1. All the reads were
324 aligned to the corresponding genome using Bowtie 2. A custom method for peak calling of
325 sequencing data was developed with BamTools, BEDTools and Rstudio. Briefly, the BamTools
326 results were filtered based on R1 (selected for NT data) or R2 (selected for TdT data). The 5'
327 coverage (experiment sample and controls) or full coverage (controls) on each position were
328 calculated based on the filtered BamTools results by BEDTools (positive and negative strand
329 separately). For both NT or TdT data on each strand, three ".tabular" files containing the genome
330 position and their corresponding read coverage (sample_coverage_5.tabular,
331 control_coverage_5.tabular, control_coverage_full.tabular) were analyzed and exported from
332 Galaxy after the previous BEDTools. These data were used to normalize the read coverage by the
333 sequencing depth and then calculate the read coverage ratio of specific position compared to its
334 up- downstream position in the same sample and the same sample in negative controls. Three ratios
335 were calculated at each position by RStudio for modification site calling: coverage of position N
336 (sample)/coverage of position N-1(sample), coverage of position N(sample)/coverage of position
337 N+1(sample), and coverage of position N(sample)/coverage of position N(control). Positions with
338 a ratio>1 were retained using the following R scripts: TdT_positive_strand.R

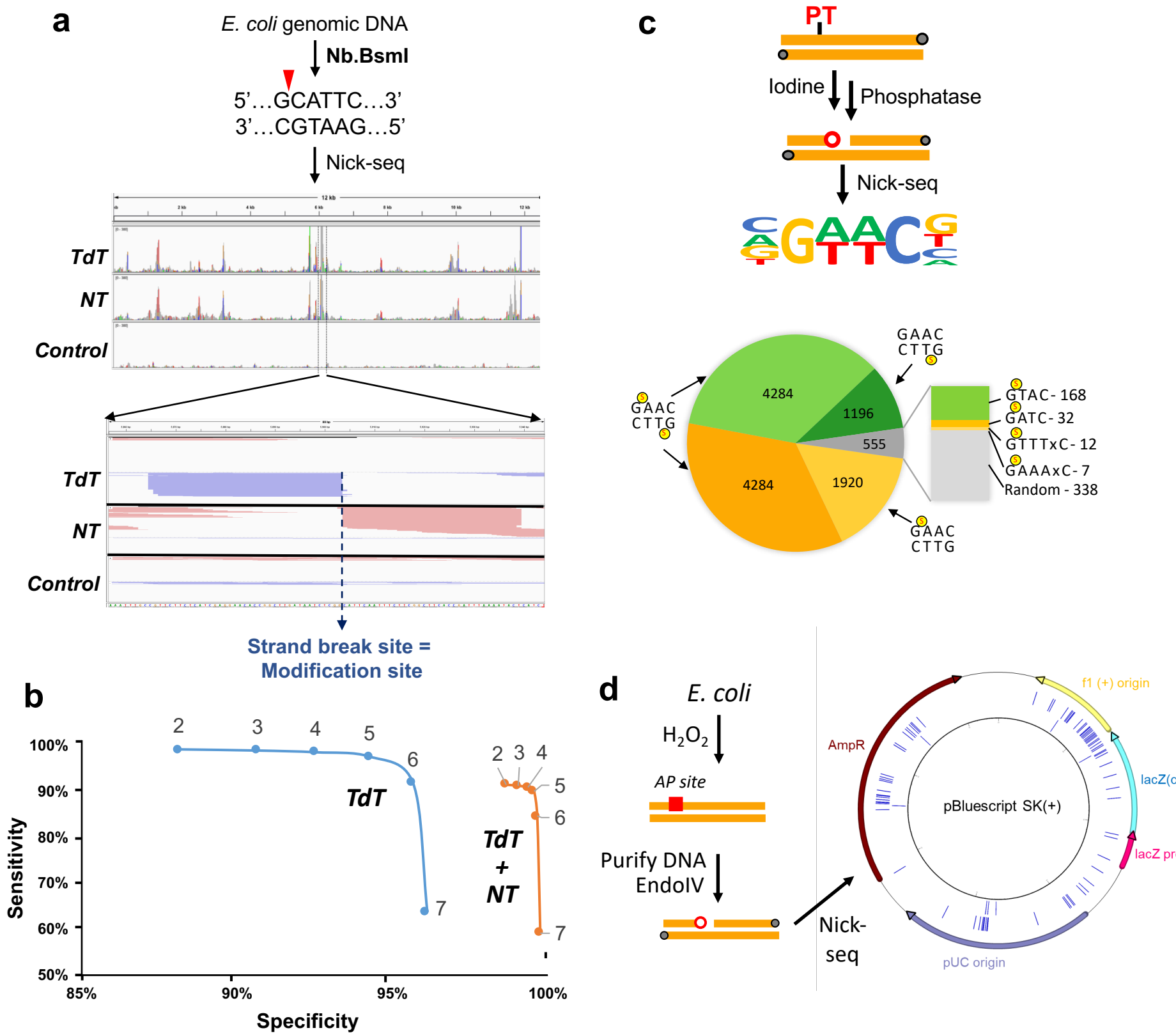
339 TdT_negative_strand.R NT_positive_strand.R NT_negative_strand.R From these datasets, the
340 intersection of the datasets from the NT and TdT methods were calculated using the following R
341 scripts: TdT_positive+NT_negative.R TdT_negative+NT_positive.R The output files (CSV files;
342 Excel format) contain the read coverage ratio information for the putative nick sites. The ratio
343 cutoffs can be varied in the Excel spreadsheet as needed. For example, for site-specific nicking by
344 Nb. BsmI and Nb. BsrDI, we determined that a ratio >2 was adequate to capture nearly all sites,
345 while for variable sites (PT) or unknown samples (H_2O_2), the ratio was increased to 5-10.
346
347



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