		A101555A_nr-reporting-summary_2020_04_22 updated.pdf

A. Additional Supplementary Files

Type	Number If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Table	1	Tables.xlsx	Supplementary Tables 1-2

Colibactin DNA damage signature indicates mutational impact in colorectal cancer

Paulina J. Dziubańska-Kusibab[1†], Hilmar Berger[1†], Federica Battistini[2#], Britta A.M. Bouwman[3#], Amina Iftekhar[1#], Riku Katainen[4#], Tatiana Cajuso[4], Nicola Crosetto[3], Modesto Orozco[2,5], Lauri A. Aaltonen[4], and Thomas F. Meyer[1*]

† Shared first authors

Shared second authors

[1] Department of Molecular Biology, Max Planck Institute for Infection Biology, 10117 Berlin, Germany

[2] Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain

[3] Science for Life Laboratory, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 17165 Stockholm, Sweden

[4] Applied Tumor Genomics Research Program and Department of Medical and Clinical Genetics, Medicum, University of Helsinki, 00014 Helsinki, Finland

[5] Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain

* Corresponding author:

Prof. Thomas F. Meyer
Department of Molecular Biology
Max Planck Institute for Infection Biology
Charitéplatz 1
10117 Berlin
Germany
Email: tfm@mpiib-berlin.mpg.de

Abstract. The mucosal epithelium is a common target of damage by chronic bacterial infections and accompanying toxins and most cancers originate from this tissue. We investigated if colibactin, a potent genotoxin[1] associated with certain strains of *Escherichia coli*[2], creates a specific DNA damage signature in infected human colorectal cells. Notably, the genomic contexts of colibactin-induced DNA double-strand breaks (DSBs) were enriched for an AT-rich hexameric sequence motif, associated with distinct DNA shape characteristics. A survey of somatic mutations at colibactin target sites of several thousand cancer genomes revealed significant enrichment of this motif in colorectal cancers (CRCs). Moreover, the exact DSB loci corresponded with mutational hot spots in cancer genomes, reminiscent of a trinucleotide signature previously identified in healthy colorectal epithelial cells[3]. This work provides evidence for the etiological role of colibactin in human cancer.

In humans, besides the prototypical role of *Helicobacter pylori* in gastric cancer[4], several bacterial species have been proposed to have a tumorigenic role in CRC, including *Fusobacterium nucleatum* [5] and colibactin-producing strains of *E. coli* [6,7]. Mechanistic studies have indicated distinct bacteria-induced cancer-promoting features, including the activation of inflammatory and growth-promoting signaling hubs and the induction of DNA damage [8]. Genotoxins are widespread amongst bacterial species, where they participate in inter-microbial competition[9]. Several bacterial genotoxins, such as DNA alkylators duocarmycin[10], yatakemycin[11] and CC-1065 as well as polymerase inhibitors distamycin[12] and netropsin [13], show preferential binding to AT-rich motifs in the minor groove [14]. The B2 phylogenetic group of *E. coli* harbors a 54 kilobase (kb)-long *pks* genomic island that encodes a polyketide-peptide hybrid, responsible for the production of the secondary metabolite colibactin [1]. This toxin has been shown to induce double-strand breaks (DSBs) in host cells, followed by the activation of the G2-M DNA damage checkpoint pathway [1]. The chemical structure of several colibactin precursors [15-18] and the isolation of colibactin-dependent N3 adenine adducts from host DNA [19] have been recently described. The high-resolution mature structure of colibactin depicts a highly symmetrical molecule that contains identical cyclopropane warheads at each end, eliciting DNA cross-links [20]. These DNA cross-links may be processed by depurination at sites of colibactin action [21].

We set out to explore the potential specificity of colibactin-induced DNA damage by determining the characteristics of DNA breakpoints in infected human colorectal cells.

Interestingly, using a multimodal approach (**Extended Data** Fig. 1), we identified an infection signature that was preserved in the genomes of about 10% of human CRCs.

To induce damage by colibactin, we infected human colorectal adenocarcinoma Caco-2 cells with *pks+ or pks- E. coli* for 3 hrs, followed by fixation of the cells (**Extended Data** Fig. 2A). Controls included no treatment or treatment with DSB-inducing, topoisomerase II inhibitor etoposide (Fig. 1A). Following infection, colibactin-induced DSBs were identified using sBLISS[22], a sequencing-based DSB capture method, enabling unbiased identification of DSBs at a genome-wide scale with nucleotide resolution (Fig. 1A and Methods). To verify that sBLISS was able to capture known patterns of etoposide-induced damage, the breakpoint density around transcription start sites (TSS) was determined [22-24]. Indeed, an expected increase in breakpoint counts around TSSs was observed in the etoposide control (**Extended Data** Fig. 2B). Interestingly, unlike etoposide-induced DSBs, or those detected in the negative controls, DSBs induced upon *pks+ E. coli* infection failed to show any strong correlation with known particular genomic regions (**Extended Data** Fig. 2C) using Locus Overlap Analysis (LOLA) [25].

While we hypothesized that sBLISS captured the preferred target sites of colibactin, we analyzed the nucleotide sequence context of induced DSBs (+/-10 nt) in *pks+ E.coli*- vs. *pks- E.coli*-infected cells, using discriminative motif discovery (DREME) [26] (Fig. 1B). One of these motifs, a heptamer sequence, was strongly enriched in all replicates, overlapped the break point position, and included the consensus pattern AAWWTT with a 3' variable, single nucleotide extension (**Extended Data** Fig. 3A). No enrichment for AAWWTT, referred to as the colibactin damage motif (CDM) could be identified when comparing etoposide-induced DSBs to untreated controls (**Extended Data** Fig. 3B). In order to evaluate how strong was the enrichment of this motif over all other motifs, we compared the enrichment of short oligonucleotide sequences (i.e., pentanucleotides or hexanucleotides) around break sites following different treatments (Methods; Fig. 1C). We found that DSBs of *pks+ E. coli*-infected cells were generally enriched in AT-rich regions. This enrichment was particularly high for the hexanucleotides AAATTT, AAAATT and AATATT, together with their complementary mates. These sequence preferences of DSBs in *pks+ E. coli*-infected cells were evident in comparison to both *pks- E. coli*-infected (Fig. 1D, top panel) and untreated cells (**Extended Data** Fig. 4A). This enrichment was independently detected with almost identical relative values in all four biological replicates (Extended Data Table S1).

Importantly, no meaningful sequence enrichments were observed in the comparison between *pks*- *E. coli*-infected cells and untreated cells (Fig. 1D, bottom panel). Hence, the preference for AT-rich sequences is directly linked to the action of colibactin, rather than to non-toxigenic *E. coli* infection. Similar enrichments were found for pentanucleotide motifs, whereby sub-patterns of AAWWTT were most strongly enriched (Extended Data Fig. 4B). Poly(dA:dT) tracts in eukaryotic genomes have been experimentally associated with genomic fragility [27]. Importantly, we found that the enrichment of other AT-rich sequences was substantially weaker than that of AAWWTT (Extended Data Fig. 4C). Moreover, AT-rich nucleosome-free regions[28,29] failed to show a distinct preference for AAWWTT, indicating that the motif preference is not driven by increased access of colibactin to unprotected DNA regions (Extended Data Fig. 4D).

Binding of small molecules to DNA is partially determined by DNA shape characteristics (reviewed by Tse et al. [30]). We predicted the DNA shape parameters for the central positions of all possible 4096 hexanucleotides and correlated them with the log2-ratios of hexanucleotide sequence enrichment at DSB positions in pks+ *E. coli*- vs. pks- *E. coli*-infected cells. Remarkably, two out of the three colibactin-specific hexanucleotide sequences (AAAATT/AATTTT and AAATTT) exhibited the narrowest minor groove widths as well as the most extreme negative electrostatic potential (Fig. 2A, Extended Data Fig. 5A). Values for DNA stiffness descriptor [31] (k_tot, see Methods) revealed that these tracts also possessed high intrinsic rigidity (Supplementary Table S1), rendering them inert to distortion. Multivariate analysis of DNA shape parameters along the full length of all possible hexanucleotides confirmed these extreme characteristics of CDM-related sequences (Fig. 2B).

In order to elucidate the binding between the DNA and colibactin, we built a molecular model using quantum mechanics calculations as a first structural prediction (see Methods). The optimized structures were hydrated and subjected to molecular dynamics simulations using state-of-the-art simulation conditions (see Methods for detailed parametrization). Colibactin appeared as a rather flexible molecule with an average end-to-end distance of around 13 Å (Extended Data Fig. 5B). The development of a putative model (see Methods) shows a very stable binding of colibactin to the minor groove (Fig. 2C). It displays excellent van der Waals contacts with all the groove's walls, with the cyclopropane rings pointing towards the adenines on opposite strands (Fig. 2D). Based on the equilibrium trajectory, we

determined that the number of base-pairs involved in the binding could fluctuate between 4 and 5, depending on the orientation of the cyclopropane ring and the carbon that alkylates the N3 of the adenines (Fig. 2D, enlargement). In all cases, colibactin fitted perfectly into the narrow minor groove of the targeted sequences and adopted a spatial arrangement with a distance of around 4-5.5 Å between N3 and the center of mass of the cyclopropane, which would facilitate N3 alkylation. These results and further analysis of sequence enrichment (Extended Data Fig. 5C) indicate that apart from the presence of adenines in a distance of 3-4 base-pairs on opposite strands and global DNA shape parameters (see above), the presence of central A/T dinucleotides also influenced colibactin binding.

Using our approach depicted in Extended Data Fig. 1C, we investigated whether a specific mutational signature associated with this sequence exists in cancers that have been experimentally connected to *pks+ E. coli* infection [6,32]. Thus, we used whole-exome sequencing data from CRC samples [33] (WXS project, n=619) and various databases (TCGA project, https://www.cancer.gov/tcga, 10,224 tumor cases in 24 cancer types, n=553 for CRCs) to test whether somatic mutations are specifically enriched in the sequences constituting the CDM (Methods). We found that mutation rates in AAWWTT motifs were enriched as compared with all other WWWWWW motifs in CRCs in both analyzed data sets (Methods, Fig. 3A, Extended Data Fig. 6A). We also observed enrichment at AAWWTT motifs in some other cancers (e.g., stomach cancer and uterine corpus endometroid cancer) and in cancers containing polymerase epsilon (*POLE*) mutations (Extended Data Fig. 6A). A thorough inspection of mutated motifs in these cancers, however, indicated that at least some of this enrichment could be attributed to other mutational processes also affecting the same motif (Extended Data Fig. 7A, B).

We validated the findings from the WXS data in another cohort of CRC samples, assessed by whole genome sequencing (WGS) [34], and analyzed the enrichment of mutations in CDM-related sequences for 200 tumors, including 184 microsatellite-stable (MSS), 3 *POLE*-mutated and 13 microsatellite-instable (MSI) cases. This analysis was run in a similar manner to the WXS data analysis; however, each sample was analyzed separately instead of pooling the samples into sub-cohorts, so as to identify motif enrichment and mutational loads in individual samples. We found significant (Mann-Whitney-U test, $p<0.05$, FDR <20%) enrichment of mutations at colibactin-specific pentanucleotide sequences as compared with other sequences with the same length and A/T content in 3/3 *POLE*-mutated samples and

48/184 (26%) MSS cases, but not in MSI cases (Fig. 3B). We further found similar enrichments in MSS samples by probing for colibactin-specific pentanucleotide (AAATT/AAAAT) and hexanucleotide sequences (Fig. 3B and Extended Data Fig. 8A), whereby pentanucleotide sequences showed less enrichment in MSI cases. . The median number of mutations in MSS samples at the pentanucleotide motifs was 963 (range: 66-11876), corresponding to a median proportion of 6.7% (range: 3.9-44.7%).

We next compared mutations at the AAWWTT motifs to established mutational signatures (Extended Data Fig. 1C) that have been previously deduced in pan-cancer analyses [35] and normal colorectal epithelial cells [3]. To this end, we derived a theoretical single base substitution (SBS) signature based on the frequency of trinucleotide sequences, assuming a uniform distribution of T>A, T>C and T>G mutations across the motifs (Methods, Fig. 3C). Our comparison showed the highest similarity (cosine correlation coefficient 0.8) with SBSA, an SBS previously identified in healthy colonic mucosa [3]. This signature is mainly characterized by T>C mutations at ATA, ATT and TTT trinucleotides and T>G mutation at TTT [3]. A total of 18 out of 184 MSS CRC cases (9.8%) showed SBSA contributions to the total mutational load that was greater than 5%. SBS exhibiting preferential nucleotide changes at the same trinucleotides (SBS41, SBS57, SBS28) showed similarities as well (Fig. 3C). The estimated contributions of SBSA to the total mutation load in the WGS of the CRC cohort correlated significantly with the respective proportion of mutated CDMs (Fig. 3D). This finding suggests the SBSA trinucleotide signature could be a surrogate of the colibactin-specific mutational signature. Interestingly, we observed that the proportion of mutations at CDMs was particularly enriched in sigmorectal tumor sites (Figure 3E, Kruskal-Wallis test, $p < 0.001$).

To assess whether mutations in CDM could contribute to tumorigenesis by affecting pan-cancer driver genes, we searched for such mutations in the TCGA dataset. Intriguingly, CDM-associated driver gene mutations were identified in 18/286 (6.9%) of colon adenocarcinoma cases and in 8/89 (9.0%) of rectal adenocarcinoma cases, amongst which 10 out of 28 mutations were identified in the *APC* gene (Fig. 3F). Interestingly, we also observed frequent mutations in driver genes of other cancer types, of which the respective tissue might have been exposed to *E. coli* infection (Extended Data Fig. 8B and Supplementary Table 2).

We hypothesized that determining the exact ends of colibactin-induced DSBs (Extended Data Fig. 1D) could provide additional information on the resulting mutational signature. Thus, we

devised a method for detecting the authentic 5' ends of a DSB in both forward and reverse directions (Fig. 4A, Extended Data Fig. 9A), which could discriminate between blunt-ended DSBs and DSBs with 5' or 3' overhangs (Extended Data Fig. 9A). The approach was validated using published BLISS data from a cell line that contained an inducible endogenous AsiSI restriction enzyme [36] (Extended Data Fig. 9B, Methods). To identify colibactin-specific break patterns, we assessed the positional distribution of identified DSB ends at three CDM-related sequences and three control sequences of the same length and similar AT content (Fig. 4B). All sequences revealed similar profiles of DSB ends in control *pks- E. coli*-infected cells and untreated cells (Fig. 4B, C). Conversely, in *pks+ E. coli*-infected cells, we noticed a distinct positional increase of DSB ends in colibactin-specific AAWWTT motifs as compared to control sequences (Fig. 4B, C). Interestingly, the observed colibactin-specific positional end point increases suggest an action of colibactin resulting in 5' overhangs of 2 nucleotides, with a preference of end points downstream of positions 2 and 5 in the hexamer motif of the forward and reverse strands, respectively.

Next, we explored (Extended Data Fig. 1D) a possible relationship between the positions of colibactin-induced damage and mutational changes in MSS CRC patients of the WGS cohort [34]. By comparing the 20 tumors with the highest vs. lowest mutational burden at CDMs, we identified the highest number of mutations at positions 2 and 5 in the CDM-related sequences AAAATT and AATATT (Fig. 4D). For the AAATTT sequence motif, we found a broader distribution of mutated nucleotide positions, compatible with the relatively broader distribution of DSB sites (Fig. 4D). The mutational changes at positions 2 and 5 nicely correlated with the experimentally-deduced downstream (5'>3') break ends, which could be the result of single-strand breaks downstream (5'>3') to the damaged bases (Fig. 4E). Mutational changes at those positions included mainly T:A>C:G. In agreement with previous information on SBSA[3], we found that T:A>C:G mutations at CDM had a higher average allelic frequency than similar changes at other motifs (two-sided Mann-Whitney-U test, $p < 0.001$) or than mutations attributable to the age-dependent signature SBS1[35] (two-sided Mann-Whitney-U test, $p < 0.001$), indicating a mutational process happening early in carcinogenesis rather than an ongoing process.

Using an unbiased global approach to assess the specificity of colibactin-induced DSB formation, we identified the motif AAWWTT as the preferred target. Most but not all sequences constituting the CDM exhibited extreme physical values of the DNA duplex,

characterized by a very narrow minor groove width, a highly negative electrostatic potential and a pronounced stiffness. Even though a definitive understanding of the binding conformation requires further support from 3D structural analysis, the 3D model provides valuable insight. In concert with our experimental data, this model suggests that colibactin, using its two cyclopropane groups, selectively targets two adenines on opposite strands at a distance of ~4 nt in the motif AAWWTT. This model is in perfect agreement with the observed DSB ends generated by colibactin's action and the position of mutations in the CDMs of the relevant cancer genomes. Accordingly, our data illuminate the mechanism of DSB formation by colibactin and provide a firm link to the presence of a distinct mutational signature in human cancer genomes, most specifically in colorectal carcinoma.

Recent mechanistic data suggest that, following the formation of N3 adenine adducts [19], colibactin's DNA cross-links undergo depurination via beta-elimination [21]. This may involve the action of glycosylases [37]. Our exact DSB end point determination by sBLISS suggests that cleavage occurs immediately downstream (5'>3') of positions 2 and 5 of the putatively-alkylated adenines on the opposing strands. Thus, cleavage results in staggered ends with 5' overhangs of two nucleotides each (Extended Data Fig. 1E). Most intriguingly, the identified cleavage positions next to the alkylated adenines perfectly match the location of CDM-associated mutations in CRCs. The ultimate manifestation of mutations in the damaged lesions has been suggested to involve the action of error-prone translesion polymerases in opposition to the damaged bases or apurinic sites [38]. Alternatively, other host DNA repair mechanisms might act, such as nucleotide excision of alkylated adenines [39]. This could lead to DSBs, resection of break ends, or complete repair, amongst others; however, details of the mutational process requires further investigation.

Most strikingly, we were able to demonstrate the enrichment of mutations in the CDMs of cancer genomes. Owing to the unique features of these mutations, we propose that this provides definitive evidence of a bacterial infection signature in human cancer. Further, our observation of colibactin-related driver mutations in CRCs is suggestive of a potential causal relationship. Typically for CRC, the initial driver mutations cause an intensification of Wnt signaling, consistent with mutations in *APC*[40]. In fact, our search for driver mutations in the CDM of cancers revealed enrichment in such driver genes, particularly *APC* – thus corroborating the causative role of pks+ *E. coli* in colon cancer.

An intriguing speculation is that, besides *pks+ E.coli*, other bacteria, such as *Klebsiella pneumoniae* potentially harboring the same genotoxin island[2], might be involved in human carcinogenesis as well. While current data only provide weak evidence for a broader impact of the colibactin in other cancers, this possibility remains as an important perspective for future investigation.

References

1. Nougayrede, J.P.*, et al.* Escherichia coli induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848-851 (2006).
2. Putze, J.*, et al.* Genetic structure and distribution of the colibactin genomic island among members of the family Enterobacteriaceae. *Infect Immun* **77**, 4696-4703 (2009).
3. Lee-Six, H.*, et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).
4. Watanabe, T., Tada, M., Nagai, H., Sasaki, S. & Nakao, M. Helicobacter pylori infection induces gastric cancer in mongolian gerbils. *Gastroenterology* **115**, 642-648 (1998).
5. Castellarin, M.*, et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res* **22**, 299-306 (2012).
6. Arthur, J.C.*, et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120-123 (2012).
7. Cougnoux, A.*, et al.* Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype. *Gut* **63**, 1932-1942 (2014).
8. Bleich, R.M. & Arthur, J.C. Revealing a microbial carcinogen. *Science* **363**, 689-690 (2019).
9. Hibbing, M.E., Fuqua, C., Parsek, M.R. & Peterson, S.B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**, 15-25 (2010).
10. Takahashi, I.*, et al.* Duocarmycin A, a new antitumor antibiotic from Streptomyces. *J Antibiot (Tokyo)* **41**, 1915-1917 (1988).
11. Igarashi, Y.*, et al.* Yatakemycin, a novel antifungal antibiotic produced by Streptomyces sp. TP-A0356. *J Antibiot (Tokyo)* **56**, 107-113 (2003).
12. Arcamone, F., Penco, S., Orezzi, P., Nicolella, V. & Pirelli, A. STRUCTURE AND SYNTHESIS OF DISTAMYCIN A. *Nature* **203**, 1064-1065 (1964).
13. Finlay, A., Hochstein, F., Sobin, B. & Murphy, F. Netropsin, a new antibiotic produced by a Streptomyces. *Journal of the American Chemical Society* **73**, 341-343 (1951).
14. Boger, D.L. & Johnson, D.S. CC-1065 and the duocarmycins: unraveling the keys to a new class of naturally derived DNA alkylating agents. *Proc Natl Acad Sci U S A* **92**, 3642-3649 (1995).
15. Zha, L.*, et al.* Colibactin assembly line enzymes use S-adenosylmethionine to build a cyclopropane ring. *Nat Chem Biol* **13**, 1063-1065 (2017).
16. Healy, A.R., Vizcaino, M.I., Crawford, J.M. & Herzon, S.B. Convergent and Modular Synthesis of Candidate Precolibactins. Structural Revision of Precolibactin A. *J Am Chem Soc* **138**, 5426-5432 (2016).
17. Zha, L., Wilson, M.R., Brotherton, C.A. & Balskus, E.P. Characterization of Polyketide Synthase Machinery from the pks Island Facilitates Isolation of a Candidate Precolibactin. *ACS Chem Biol* **11**, 1287-1295 (2016).
18. Vizcaino, M.I. & Crawford, J.M. The colibactin warhead crosslinks DNA. *Nat Chem* **7**, 411-417 (2015).
19. Wilson, M.R.*, et al.* The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**(2019).

20. Xue, M.*, et al.* Structure elucidation of colibactin and its DNA cross-links. *Science* **365**(2019).
21. Xue, M., Wernke, K. & Herzon, S.B. Depurination of colibactin-derived interstrand cross-links. *Biochemistry* (2020).
22. Yan, W.X.*, et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat Commun* **8**, 15058 (2017).
23. Canela, A.*, et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170**, 507-521.e518 (2017).
24. Yang, F., Kemp, C.J. & Henikoff, S. Anthracyclines induce double-strand DNA breaks at active gene promoters. *Mutat Res* **773**, 9-15 (2015).
25. Sheffield, N.C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-589 (2016).
26. Bailey, T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653-1659 (2011).
27. Tubbs, A.*, et al.* Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse. *Cell* **174**, 1127-1142 e1119 (2018).
28. Nelson, H.C., Finch, J.T., Luisi, B.F. & Klug, A. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature* **330**, 221-226 (1987).
29. Yuan, G.C.*, et al.* Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626-630 (2005).
30. Tse, W.C. & Boger, D.L. Sequence-selective DNA recognition: natural products and nature's lessons. *Chem Biol* **11**, 1607-1617 (2004).
31. Drsata, T.*, et al.* Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Res* **42**, 7383-7394 (2014).
32. Buc, E.*, et al.* High prevalence of mucosa-associated E. coli producing cyclomodulin and genotoxin in colon cancer. *PLoS One* **8**, e56964 (2013).
33. Giannakis, M.*, et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep* **15**, 857-865 (2016).
34. Katainen, R.*, et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-821 (2015).
35. Alexandrov, L.B.*, et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
36. Iannelli, F.*, et al.* A damaged genome's transcriptional landscape through multilayered expression profiling around in situ-mapped DNA double-strand breaks. *Nat Commun* **8**, 15656 (2017).
37. Iyama, T. & Wilson, D.M., 3rd. DNA repair mechanisms in dividing and non-dividing cells. *DNA Repair (Amst)* **12**, 620-636 (2013).
38. Choi, J.Y., Lim, S., Kim, E.J., Jo, A. & Guengerich, F.P. Translesion synthesis across abasic lesions by human B-family and Y-family DNA polymerases alpha, delta, eta, iota, kappa, and REV1. *J Mol Biol* **404**, 34-44 (2010).
39. Martin, L.P., Hamilton, T.C. & Schilder, R.J. Platinum resistance: the role of DNA repair pathways. *Clin Cancer Res* **14**, 1291-1295 (2008).
40. Morin, P.J.*, et al.* Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science* **275**, 1787-1790 (1997).

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors would like to thank Silvano Garnerone from N.C. laboratory for processing raw sBLISS data, Pablo D. Dans from Biophysical Chemistry Lab, Department of Biological Science (CENUR Litoral Norte), UdelaR, UY, for constructive discussions about the theoretical model of colibactin, Ulrich Dobrindt from University of Münster for providing *E. coli* strains and Kfir Lapid and Rike Zietlow for editing the manuscript. We also acknowledge the computational resources provided by the CSC–IT Center for Science, Finland.

Funding

P.J.D.K. was supported by IMPRS-IDI graduate school. B.A.M.B was supported by a Rubicon fellowship from the Netherlands Organisation for Scientific Research (NWO). M.O. is an ICREA (Institució Catalana de Recerca i Estudis Avancats) academia researcher. This work was supported by the German Research Foundation (DFG, grant number ME705/18-1) to T.F.M., by the Spanish Ministry of Science (grants BIO2015-64802-R, BFU2014-61670-EXP and BFU2014-52864-R), the Catalan Government (grant 2014-SGR), the Instituto de Salud Carlos III-Instituto Nacional de Bioinformática (ISCIII PT 13/000/0030), the Biomolecular and Bioinformatics Resources Platform and the EU Horizon 2020 research and innovation program (grants Elixir-Excelerate 676559 and BioExcel2:823830), and the MINECO Severo Ochoa Award of Excellence to the IRB Barcelona, and by the Ragnar Söderberg Foundation, the Swedish Foundation for Strategic Research (BD15-0095) and the Strategic Research Programme in Cancer (StratCan) at Karolinska Institutet to N.C. L.A.A. was supported by grants from the Academy of Finland (Finnish Center of Excellence No. 312041, Academy Professor grants 319083 and 320149, iCAN Flagship 320185), Jane and Aatos Erkko Foundation, Sigrid Juselius Foundation, Helsinki Institute of Life Science, and the Cancer Society of Finland.

Author contributions

P.J.D.-K., H.B. and T.F.M. conceived the project and designed experiments, N.C. and B.A.M.B. designed sBLISS experiments, P.J.D.-K. and B.A.M.B. performed sBLISS experiments, A.I. performed E.coli infection and immunostaining experiments. P.J.D.-K. and A.I. prepared samples for sBLISS. Bioinformatics analysis were performed by P.J.D.-K. and H.B. Theoretical model of colibactin was built by F.B. and M.O. R.K., T.C. and L.A.A. provided and analyzed WGS CRC data. The manuscript was written by P.J.D.-K., H.B. and T.F.M.

Competing interests

The authors declare no competing interests

Data and materials availability

Input FASTA files have been submitted to GEO under accession GSE145594 and analysis scripts are available from https://github.com/MPIIB-Department-TFMeyer/Dziubanska-Kusibab_et_al._Colibactin. All other data is available in the main text or supplementary materials.

Figure Legends

Fig.1. Colibactin preferentially damages DNA in specific AT-rich motifs

(A) Illustration of the sBLISS protocol for the identification of colibactin-induced DSBs. (B) We employed DREME to analyze motif enrichment around DSBs in *pks+* vs. *pks- E. coli*-infected cells. The sequence AAWWTT was identified as the preferred CDM among the top three enriched motifs in four independent replicates. (C) Relative enrichment of hexanucleotide sequences in close proximity to the DSB positions (±7 nt) upon different treatments was determined as the ratio of counts of a given sequence between two conditions. (D) Bars represent the mean enrichment of the top 50 DSB-associated hexanucleotide motifs (log2 ratios of DSBs at each motif comparing two conditions) in *pks+* vs. *pks- E. coli*-infected cells (top) or *pks- E. coli*-infected cells vs. nontreated (NT) cells (bottom), n=4 independent experiments. Insets - results for all hexanucleotides. Error bars denote the 95% confidence interval (CI) around the mean log2 ratios.

Fig. 2. Colibactin-specific binding motif correspond to extreme DNA shape parameter values and electrostatic potential

(A) Relations between hexanucleotide sequence enrichments (log2 ratio standardized to 1; mean of 4 independent experiments) in *pks+* vs*. pks- E.coli*-infected cells and values of predicted DNA shape parameters for the central nucleotides of those motifs. Error bars denote the 95% CI around mean scaled log2 ratios for each hexanucleotide sequence, n=4.

(B) 3D principal component analysis (PCA) visualization of the first three principal components from predicted DNA shape values for all positions in all possible 6 nt motifs. The motifs that match AAWWTT and/or show strong enrichment are highlighted: red – significant AAWWTT sequences; blue – significant non-AAWWTT sequences; grey – other sequences. Statistical significance is determined by an upper 95% CI limit of mean log2 ratio > 0.2, n=4 independent experiments.

(C-D) Graphical representations of the theoretical docking of the predicted colibactin structure into its preferred sequence motif (central sequence AAATTT), showing the insertion of colibactin into the minor groove, with the double-stranded DNA as a surface (C)

or in atomic details (D). The enlargement shows a zoomed-in image of the proximity of the cyclopropane to one of the N3 atoms of the adenine, highlighting the possibility to alkylate the subsequent base-pair, depending on the carbon involved in the alkylation.

Fig. 3. Several cancers show enrichment of mutations in CDMs

Enrichment of SBS mutations at colibactin-specific AAATTT/AAAATT sequences in exomes of 619 CRC cases [33] (A) or at AAATT/AAAAT pentanucleotides in whole-genome data from CRC[34] (B, n=200). (B) Top - differences in log2(mutations/base-pair) between colibactin-specific and all other WWWWW motifs. Middle - total mutation count at CDMs. Bottom - proportions of total mutations that overlap with CDMs in MSS, MSI, and *POLE*-mutated cases. (A,B) Asterisks - significant difference (two-sided Mann-Whitney-U test, $p < 0.05$ and FDR < 20%) between CDMs and all WWWW(W) motifs. Numbers under asterisks – number of mutations overlapping AAWWTT / all mutations, [number of samples per cohort]. Error bars = 0 ± 2MAD intervals of mutation rates (mutations/base-pair) of WWWWW(W) sequences, excluding CDMs, after subtracting their mean. Dots - mutation rates for AATTT/AAAAT or AAATTT/AAAATT after mean subtraction.. Crosses - means of the CDMs. Q1-Q4 cohorts defined by the quantiles of total SNV numbers. Outliers - samples with SNV numbers significantly larger than the 95% CI in each cancer entity. POLE - polymerase epsilon mutated cases. (C) Theoretical trinucleotide change signature for uniform mutations at AAWWTT motifs (top) and the matching signature SBSA (bottom).(D) Correlation between the estimated contribution of SBSA signature per sample in 184 MSS CRCs and the proportion of total SNVs overlapping colibactin-specific pentanucleotides. (E) The proportion of mutations in CDMs depending on the tumor location (n=184 MSS CRC, Kruskal-Wallis test, $p < 0.001$). (F) *APC* exonic somatic mutations in CDM in CRC cases from TCGA and the COSMIC database.

Fig. 4 Analysis of colibactin-induced DSB break ends and mutational patterns in CDMs indicate damaged positions

(A) Schematic representation of the break end determination of an arbitrary DSB configuration based on sBLISS data (see Fig. S5 for additional examples). (B) Mean proportions of break ends by the position and direction of a motif in reads (red - forward;

green - reverse) for colibactin-specific (leftmost three columns) and control motifs (rightmost three columns) in *pks+ E. coli*-infected cells, *pks- E. coli*-infected cells and control conditions, n=4 independent experiments (dots). Negative values denote mean proportions on the reverse strand of the motif. Error bars: 95% CI around mean proportions. (C) The mean differences in break end proportions (taken from Fig. 4B) between *pks+ E. coli*-infected cells and *pks- E. coli*-infected cells, n=4. Negative values denote mean differences on the reverse strand. Error bars: 95% CI around mean differences. Only increased break end proportions in *pks+ E. coli*-infected cells are shown for each position. (D) The positional frequencies of nucleotide changes in CDMs found in the CRC WGS cohort. The 20 cases with the highest and lowest proportion of mutations at AAWWTT are shown. (E) The model of colibactin-induced DSB and damaged bases as derived from the data shown in this figure. Red-colored nucleotides denote damaged bases, vertical bars are single-strand breaks, which were detected by sBLISS, and the broken line denotes the inter-strand crosslink caused by colibactin action.

Materials and Methods

Cell line, bacterial strains, *E. coli* infection and etoposide-treatment

Caco-2 cells (from ATCC® HTB-37™) were cultured at 37 °C in a humidified 5% CO_2 atmosphere, in DMEM medium (Life Technologies, cat. number: 10938-025), supplemented with 20% FCS (Biochrom, cat. number: S0115). Contamination of Mycoplasma spp. in this immortalized cell line was excluded using the Venor®GeM OneStep PCR kit (Minerva Biolabs®, cat. number: 11-8250). To infect Caco-2 cells, an overnight liquid culture of *E. coli* strain M1/5 (streptomycin-resistant and colibactin-positive) and *E. coli* strain M1/5::ΔclbR (streptomycin-resistant and colibactin-negative) was set up. Bacteria were inoculated in 5 ml of Luria broth (LB) medium and incubated overnight at 37 °C in a shaking incubator. The overnight inoculum was diluted 1:33 in infection medium (DMEM + 10% FCS + 10 mM HEPES (Life Technologies, cat. number: 15630-056)) to obtain OD600=1 after 3 hr of incubation to give 1.5 x 10^9 bacteria/ml. The prepared bacteria inoculum was further diluted to reach MOI 20, added to Caco-2 cells seeded previously and incubated for 3 hr at 37 °C. Medium was then aspirated and cells fixed according to the protocol for immunofluorescence or sBLISS.

For every biological replicate, positive (etoposide-treatment) and negative (no treatment) controls were included. Etoposide powder (Sigma Aldrich, cat. number: E1383) was diluted in DMSO to 50 mM and aliquots stored at -20 °C. Final drug dilutions to the concentration of 50 µM were performed in pre-warmed infection medium prior to each treatment. Treatment was conducted for 3 hr at 37 °C, after which the medium was aspirated, and etoposide-treated cells were fixed in the same way as *E. coli*-infected cells.

Immunofluorescence staining

Caco-2 cells grown and infected on MatTek glass-bottom dishes were washed three times with PBS (Life Technologies, cat. number: 14190-094) and fixed with 3.7% paraformaldehyde (Sigma Aldrich, cat. number: P6148) for 1 h. The cells were kept overnight in blocking buffer (3% BSA, Biomol, cat. number: 01400.100), 1% saponin (Sigma Aldrich, cat. number: 84510), 2% Triton X-100 (Carl Roth, cat. number: 3051.2) and 0.02% sodium azide (Sigma Aldrich, cat. number: S2002). Blocking was followed by overnight incubation with γH2AX antibody (Phospho-Histone H2A.X (Ser139) Antibody, Cell Signaling, cat. number: 2577, 1:500 dilution) at 4 °C. The next day, the MatTek dishes were washed three times with blocking buffer followed by overnight incubation with secondary antibody (Dianova, cat. number: 711-035-152, 1:250 dilution) diluted in blocking buffer. Phalloidin 546 (Invitrogen, cat. number: A22283, 1:200 dilution) and Hoechst (Sigma, cat. number: H6024, 1:10000 dilution) were added for staining actin filaments and DNA, respectively. The next day, cells were washed three times with blocking buffer and coverslipped using Vectashield® Antifade Mounting Medium (Vector Laboratories, cat. number: H-1000). Images were acquired using a Leica TCS SP-8 confocal microscope and processed using ImageJ.

sBLISS, an adaptation of the BLISS method

DSBs were identified using the suspension-cell BLISS (sBLISS) method [41], which is an adaptation of the previously published BLISS protocol [42,43]. In contrast to BLISS, where DSBs are labeled in fixed cells immobilized on microscope slides, in sBLISS DSBs are labeled in fixed cell suspensions. In brief, cells were treated/infected in culture dishes, after which they were trypsinized, counted, centrifuged and resuspended in pre-warmed medium to obtain 10^6 cells/ml. Then, cells were fixed by adding 16% PFA (Electron Microscopy Sciences, cat.

number: 15710) to reach a final concentration of 4%. After 10 minutes, 2 M glycine (Molecular Dimensions, cat. #MD2-100-105) was added to a final concentration of 125 mM in order to block unreacted aldehydes. This was followed by two 5 min incubations, first at room temperature and then on ice, followed by two washes in ice-cold PBS. Cross-linked cells were stored in PBS at 4 °C until further processing.

Next, the BLISS template was prepared. This includes (1) Cell lysis in 10 mM Tris-HCl, 10 mM NaCl, 1 mM EDTA, and 0.2% Triton X-100 (pH 8) buffer, followed by lysis in buffer containing 10 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, and 0.3% SDS (pH 8); (2) DSB blunting with NEB's Quick Blunting Kit (NEB, cat. number: E1201); (3) *In situ* BLISS adapter ligation using T4 DNA Ligase (ThermoFisher Scientific, cat. number: EL0011). Each BLISS adapter contained a T7 promoter sequence for IVT, the RA5 Illumina RNA adapter sequence, a random 8 nt long sequence referred to as Unique Molecular Identifier (UMI) and an 8 nt long sample barcode; (4) Phenol:chloroform-based extraction of gDNA; (5) Fragmentation of isolated genomic DNA (400-600bp) using BioRuptor Plus (Diagenode). Obtained BLISS templates were stored at -20 °C.

The final step of the BLISS protocol was *in vitro* transcription (IVT) followed by NGS library preparation. First, 100 ng of purified, sonicated and differentially-barcoded BLISS template of 1) etoposide-treated and non-treated cells, or 2) cells infected with *pks+ E. coli* or with *pks- E. coli* were pooled into one reaction, respectively. IVT was performed using MEGAscript T7 Transcription Kit (ThermoFisher, cat. number: AMB13345) for 14 hr at 37 °C in the presence of RiboSafe RNAse Inhibitor (Bioline, cat. number BIO-65028). Next, gDNA was removed using DNase I (ThermoFisher, cat. number: AM2222), and the remaining RNA was purified with Agencourt RNAClean XP beads (Beckman Coulter). The Illumina RA3 adapter sequence was ligated to the purified RNA using T4 RNA Ligase 2 (NEB, cat. number: M0242) for 2 hr at 25 °C and reverse transcription was performed with Reverse Transcription Primer (Illumina sequence) using SuperScript IV Reverse Transcriptase (ThermoFisher, cat. number: 18090050) for 50 minutes at 50 °C. This was followed by heat inactivation for 10 minutes at 80 °C. Finally, libraries were amplified with NEBNext High-Fidelity 2x PCR Master Mix (NEB, cat. number: M0541), the RP1 common primer and a uniquely selected index primer. 12 PCR cycles were conducted, and after that, libraries were purified according to the two-sided AMPure XP bead purification protocol (Beckman Coulter). Profiles of the libraries were

quantified on a BioAnalyzer High Sensitivity DNA chip. Libraries were sequenced as single-end (1x75) reads on the NextSeq platform.

Pre-processing of sequencing data

Raw sequencing data were pre-processed as previously described [42]. In brief, only reads which contained the expected prefix of UMI and sample barcode were kept using SAMtools [44]. One mismatch in the barcode sequence was allowed. Further, prefixes were trimmed, and the remaining sequences were aligned to the GRCh37/hg19 reference genome using BWA-MEM [45]. Reads with mapping quality scores ≤ 30 and those that were determined as PCR duplicates were removed. Finally, a BED file containing a list of unique DSB locations was generated. DSBs which fell into ENCODE blacklist regions [46], high coverage regions and low mappability regions [47] were removed. Kept positions of DSBs were further used in downstream analysis. For BLISS data of AsiSI-induced breaks [48] a similar procedure with simplified filtering was applied.

Locus Overlap Analysis

To identify significant overlaps of DNA DSB with genomic region sets, we used LOLA [11]. We first defined whole genome as a Universe Set, which was next divided into tiles of equal lengths (1,000 nt). For each created tile, we next searched for overlaps with DSBs captured by sBLISS using the findOverlap() function. All tiles containing ≥ 10 breaks were used as a Query Set. The runLOLA() function was executed with LOLA Core databases (reduced by Tissue clustered DNase hypersensitive sites) as well as LOLA Extended databases and custom database containing non-B-DNA regions (https://nonb-abcc.ncifcrf.gov/apps/site/references). Fisher's exact test was used with a FDR ≤ 5%.

Pattern identification around breakpoints

We screened the +/-10 bp contexts of DSB positions produced by sBLISS using DREME[49], comparing all breakpoints from *pks+ E. coli*-infected cells to *pks- E. coli*-infected cells from the same experiment, allowing for a maximum of 10 motifs per experiment with a minimum E value of less than 0.05. A similar analysis was performed for etoposide vs. untreated

control cells but allowing for up to 50 motifs per experiment. Most of the top motifs identified for *pks+ E.coli*- vs. *pks*- *E.coli*-infected cells were of 7 nt length; therefore the subsequent computation of enrichment log ratios for all possible oligonucleotides was done on windows of DSB breakpoints +/-7 bp and for oligonucleotide lengths of 5-7 nt. We computed the proportion of all DSB breakpoints in each sample that contained a given pattern and determined the log2 ratio of the proportion of each pattern in the test compared to control conditions of the same experiment. We observed that several of the patterns enriched only in specific experiments were very similar to the library barcode from the same experiments, indicating a priming activity of unbound sBLISS linkers. We therefore excluded all patterns matching library-specific barcodes with at most 1 mismatch from further analysis. Enriched heptamers were, in agreement with the DREME analysis, mostly the top enriched hexamers with different flanking nucleotides and are therefore not reported here. To visualize the proportion of each nucleotide per position the package seqLogo was used [50,51]

DNA Shape predictions

DNA structures can be described in terms of base-pair and base-step parameters that consist of three translational and rotational movements between the bases or the base pairs, respectively. At the base-pair step level, DNA deformability along these six directions has been described by the associated stiffness matrix [52]. From the ensemble of molecular dynamics (MD) simulations considering the tetramer environment using the newly refined parmbsc1 force field, we retrieved the 6x6 matrix describing the deformability of the helical parameters for each possible DNA tetramer. Pure stiffness constants corresponding to the six base-pair step parameters (shift, slide, rise, tilt, roll and twist) were extracted from the diagonal of the matrix, and the total stiffness (*K*_tot) was obtained as a product of these six constants and used as an estimate of the flexibility of each base pair step in a tetramer. For predictions of minor groove width (MGW), propeller twist (ProT), electrostatic potential (EP), helical twist (HelT) and roll (Roll) the getShape function from 'DNAshapeR' package was used [53]. Input FASTA files, containing sequences in close proximity to identified DSB (±5 nt), were extracted with a custom Python script. Central values of DNA shape parameters and EP were derived as the mean values of nucleotide at positions 3 and 4 of the motifs for MGW,ProT and EP; for stiffness, HelT and Roll the mean of the base-pair values is calculated for steps 2

610 to 4 in each hexamer. For the multivariate analysis of those parameters we used all values
611 along the full 6nt of hexanucleotide sequences predicted in the same way. The interaction
612 potential (electrostatic and van der Waals) of Na^+ probes with DNA duplexes was determined
613 using a linear approximation to the Poisson-Boltzmann equation and dielectric constant for
614 the DNA as implemented in the CMIP program [54].

615

616 **Model and Molecular dynamics set up**

617 The 3D structure and protonation state of colibactin were built starting from the smile
618 (https://pubchem.ncbi.nlm.nih.gov/compound/138805674#section=InChI) using
619 MarvinSketch (MarvinSketch, version 6.2.2, calculation module developed by ChemAxon,
620 http://www.chemaxon.com/products/marvin/marvinsketch/). The geometry of the model
621 and the partial atomic charges were assigned to the structure with General Amber Force
622 Field (GAFF) [55]. Parameters and topology files were prepared with Acpype [56]. The colibactin
623 was then simulated in explicit solvent at 298 K (see below for details) for 250 ns, and along
624 the simulation the distance between the cyclopropanes was monitored (see Extended Data
625 Fig. 3B), to study their orientation and the overall length of the free colibactin. Using
626 HADDOCK 2.4 [57], we then built the DNA-colibactin complex. For the docking, we selected a
627 representative structure of free colibactin along the MD simulation, with an average distance
628 among the cyclopropanes (red line, Extended Data Fig. 3B) and an equilibrated structure of
629 the DNA (sequence CGAAATTTCG). After the initial docking, which positioned the molecule
630 correctly along the minor groove of the DNA, we then manually rotated the molecule slightly
631 to improve the orientation of the cyclopropanes towards the N3 of the closest adenine using
632 PYMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC (2018)). To check the
633 stability of this complex and to equilibrate its structure, the model was simulated (see details
634 MD simulation below) and minimized in solution with positional restraints on the solute
635 using our well-established multi-step protocol [58,59]. The minimized structure was thermalized
636 to 298 K at NVT, and then simulated first applying harmonic restraints of 5 kcal/mol·Å2 on
637 the DNA structure and distance constraints between the cyclopropane and the N3 of the
638 adenine (respectively 4 and 5 bases apart), each represented by a harmonic restraint of 2.5
639 kcal/mol·$Å^2$. To further check the stability of the complex, we then slowly removed the
640 constraints and ran 10 replicas of MD simulation of the complex for a total of 500 ns by
641 means of Molecular Dynamics simulations at NPT (P = 1 atm; T= 298K) and one MD

simulation 350 ns long. The first 10 ns of the simulations were considered as an equilibration step and were discarded for further analysis.

In each MD simulation (DNA, free colibactin and their complex, respectively), we placed the solute in the center of a truncated octahedral box of TIP3P water molecules [60], neutralized by K+ ions. In each simulation, the Berendsen algorithm [61] was used to control the temperature and the pressure, with a coupling constant of 5 ps; and the SHAKE algorithm was utilized to equilibrium the length of hydrogen atoms involved in the covalent bonds [62]. Long-range electrostatic interactions were accounted for by using the Particle Mesh Ewald method (14) with standard defaults, and a real-space cut-off of 10 Å. For the DNA we used the newly revised force field parmBSC1 [63]. All simulations were carried out using AMBER 18 [64], analyzed with CPPTRAJ [65] and visualized using VMD 1.9.4 [66].

Cancer somatic mutation data and mutation enrichment at colibactin-specific patterns

We obtained somatic variant data from the TCGA Unified Ensemble "MC3" Call Set [67] ("TCGA pan-cancer dataset") and from the supplementary data of Giannakis et al. [68]. To test for enrichment of mutations at any motif, we first identified positions of all hexanucleotide motifs in the exonic portion of the genome. Somatic variants occurring at A or T bases were grouped in one of 6 classes (quartile 1-4, outlier or *POLE* mutated sample) depending on the total SNV number and *POLE* mutation status of the corresponding tumor sample. We included all cases with an exonic mutation with predicted impact on the coded protein in *POLE* and also those with similar mutations in the related gene *POLD1* in the class of *POLE*-mutated cases. We then computed the mutation rate for each hexanucleotide motif with respect to the number of genomic bases covered in exonic regions for the same motif. As a baseline, we established the mutation rates of all WWWWWW motifs and subtracted their mean from the mutation rate of all other hexanucleotide motifs. We then tested for significance of the mutation rate at colibactin associated AAWWTT motifs (i.e., AAATTT and AAAATT/AATTTT) compared to the remaining WWWWWW motifs using Mann-Whitney-U tests and computed the false discovery rate (FDR) using the method of Benjamini-Hochberg [69]. We applied a similar procedure for pentanucleotide motifs AAATT/AATTT and AAAAT/ATTTT. Reads from WGS of colorectal cancers [70] (EGA database accession code EGAS00001003010) were aligned to GRCh38 with BWA-MEM [45] and called using Mutect2 [71].

All single nucleotide variant calls (PASSed by Mutect2) were used to determine the number of mutations overlapping WWWWW pentanucleotides and WWWWWW hexanucleotides and further analyzed in a similar way as for exome sequencing data on an individual sample basis. For correlation of colibactin-specific mutation load with SBSA and clinical parameters, we computed a colibactin mutation proportion. We considered colibactin mutations as all A/T > C/G mutations at 2nd and 5th position in the AAATTT, AATATT, AAAATT and AATTTTT motifs. That number was divided by all other A/T > C/G mutations outside those motifs. Somatic mutation and motif analyses with WGS CRC samples were performed with BasePlayer[72]. Allelic proportions of somatic variants from rectal MSS WGS CRC classified as colibactin associated (A/T > C/G mutations at 2nd and 5th position in the AAATTT, AATATT, AAAATT and AATTTTT) were compared to those of all other somatic variants or age-related signature SBS1 associated variants (C/G>T/A at CpG) of the same tumors using

Correlation with known trinucleotide signatures

Known trinucleotide SBS signatures were obtained from COSMIC Mutational Signatures v3 [73] Signatures from Lee-Six et al. [74] were generated using R scripts and trinucleotide frequencies provided with this publication. We used the trinucleotide context of the three possible sequences of the AAWWTT motif (AAATTT, AATTTT and AATATT) to determine trinucleotide frequencies in those motifs. We used the reverse complement of the trinucleotide if the middle base was an A and extended the sequence by N (i.e., all possible nucleotides) for trinucleotides overlapping the end of the motif by one base. For each possible trinucleotide, we established the total number of occurrences in AAWWTT. We assumed a uniform distribution of all possible mutations T>A/C/G nucleotide changes across all positions of the motif, and then generated a trinucleotide change signature by computing the proportion of each combination of trinucleotide and nucleotide change in AAWWTT motifs. The cosine correlation was computed between this AAWWTT signature and all known signatures in R. Contributions to known trinucleotide signatures in cancer samples were estimated using the R package deconstructSigs [75].

Driver gene analysis

Entity-specific driver genes were obtained from the Catalog of Cancer Genes (Cancer Genome Interpreter, https://www.cancergenomeinterpreter.org/genes)[76]. Pan-cancer driver genes were obtained from a published analysis based on the TCGA cohort [77]. Somatic variants from this database were classified as protein changing if the impact class was HIGH or it was MODERATE, and the protein change was classified by PolyPhen or SIFT as probably damaging or deleterious, respectively. Somatic variants from samples of genome-wide screens other than TCGA in COSMIC[78] were classified as protein changing if it had been classified as pathogenic by FATHMM or if its consequence was a frameshift deletion or a nonsense substitution.

Analysis of biases in AAWWTT mutation enrichment due to other signatures

We obtained frequencies of pentanucleotide changes, trinucleotide changes and trinucleotide SBS signature contributions for PCAWG samples from the PCAWG project[79]. We derived a theoretical AAWWTT pentanucleotide signature similar to the AAWWTT trinucleotide signature above. We projected trinucleotide changes of known cancer-associated SBS signatures to trinucleotides, computing the weight of each pentanucleotide change as the weight of the corresponding central trinucleotide change in the SBS signature followed by normalization to a sum of one across all pentanucleotide changes. We computed the contribution of SBSA in all PCAWG samples using trinucleotide counts provided by PCAWG and used the contributions of all other SBS signatures as provided by PCAWG.

Clinical data

Limited clinical data on tumor patients was obtained from a study by Katainen et al [70] that was reviewed and approved by the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS). Signed informed consent or authorization from the Finnish National Supervisory Authority for Welfare and Health has been obtained for all the sample materials used.

Break end analysis

DSB positions were derived from sBLISS estimated breakpoints in a strand-specific manner as the middle of the positions of the first base in the mapped genomic fragment and the last base in 5' direction (i.e. the direction of the linker). Genomic positions of oligonucleotide patterns of interest were identified in GRCh37. Breakpoints falling with the pattern +/- 2bp flanking regions were kept and the relative position of the breakpoint in the 5'-3' motif were computed. Each strand-specific breakpoint was classified as mapping to the strand containing the motif in forward or backward direction. Counts for each relative pattern position and motif direction were computed for each pattern and normalized by the total number of breaks mapping to this pattern for each sample. Differences between normalized counts were computed for each experiment between *pks+ E. coli* and *pks- E. coli* conditions.

Pattern enrichment around nucleosomes

Genomic positions of all hexanucleotide patterns with 6 A or T were identified in GRCh37. We used nucleosome positions identified in Gaffney at al [80] and determined the distance of each pattern to the nearest nucleosome center, keeping only those pattern sites within +/- 1500bp around the center. Patterns were classified into three groups: a) AAWWTT motifs b) AAAAAA, ATATAT and TATATA which showed a distribution distinct from all other patterns around nucleosome centers and c) all remaining patterns.

Data analysis and visualization

All visualizations and statistical analyses were produced using R v3.4 [81]. For all boxplots, the top, center and bottom of the box correspond to the 75th, 50th (median) and 25th percentile of all data; the whiskers extend to the most extreme data points, which are no more than 1.5 times the length of the box (i.e., the inter-quartile range) beyond the box. All data points further away are shown as outliers.

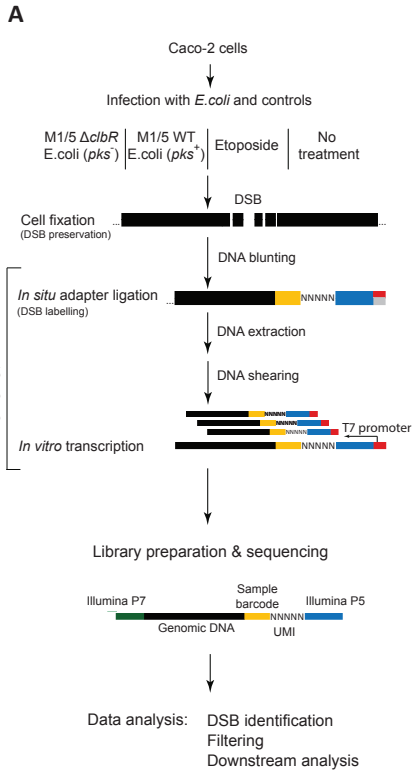
Reporting Summary

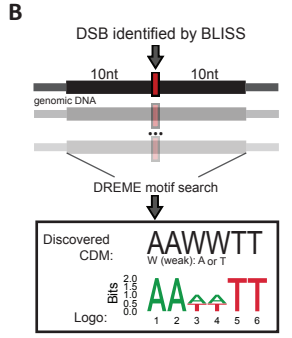
Further information is available from the Life Sciences Reporting Summary linked to this document.

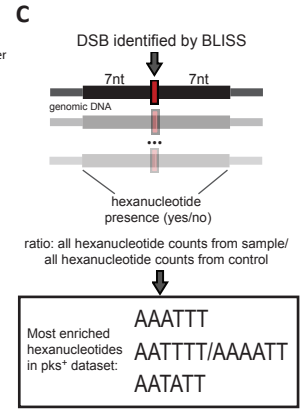
References for Materials and Methods

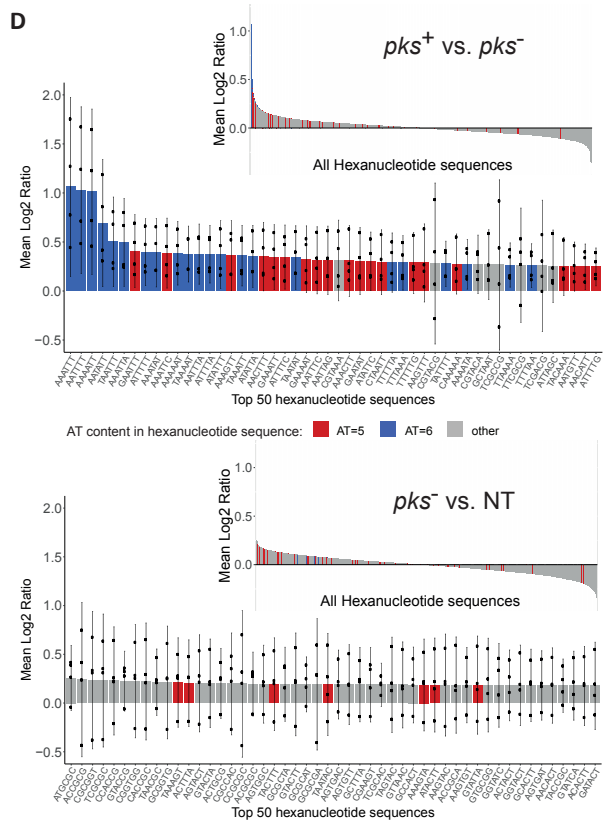
41. Gothe, H.J*., et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol Cell* **75**, 267-283.e212 (2019).
42. Yan, W.X*., et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat Commun* **8**, 15058 (2017).
43. Zhang, F*., et al.* Breaks Labeling in situ and sequencing (BLISS). *Protocol Exchange* **DOI: 10.1038/protex.2017.018**(2017).
44. Li, H*., et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
46. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
47. Speir, M.L*., et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res* **44**, D717-725 (2016).
48. Iannelli, F*., et al.* A damaged genome's transcriptional landscape through multilayered expression profiling around in situ-mapped DNA double-strand breaks. *Nat Commun* **8**, 15656 (2017).
49. Bailey, T.L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653-1659 (2011).
50. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097-6100 (1990).
51. Bembom, O. seqLogo: Sequence logos for DNA sequence alignments. R package version 1.44.0. (2017).
52. Drsata, T*., et al.* Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Res* **42**, 7383-7394 (2014).
53. Chiu, T.P*., et al.* DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211-1213 (2016).
54. Gelpi, J.L*., et al.* Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins* **45**, 428-437 (2001).
55. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. & Case, D.A. Development and testing of a general amber force field. *J Comput Chem* **25**, 1157-1174 (2004).
56. Sousa da Silva, A.W. & Vranken, W.F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res Notes* **5**, 367 (2012).
57. van Zundert, G.C.P*., et al.* The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720-725 (2016).
58. Dans, P.D*., et al.* Long-timescale dynamics of the Drew-Dickerson dodecamer. *Nucleic Acids Res* **44**, 4052-4066 (2016).
59. Perez, A., Luque, F.J. & Orozco, M. Dynamics of B-DNA on the microsecond time scale. *J Am Chem Soc* **129**, 14739-14745 (2007).
60. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **79**, 926-935 (1983).
61. Berendsen, H.J., Postma, J.v., van Gunsteren, W.F., DiNola, A. & Haak, J.R. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **81**, 3684-3690 (1984).
62. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H.J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics* **23**, 327-341 (1977).

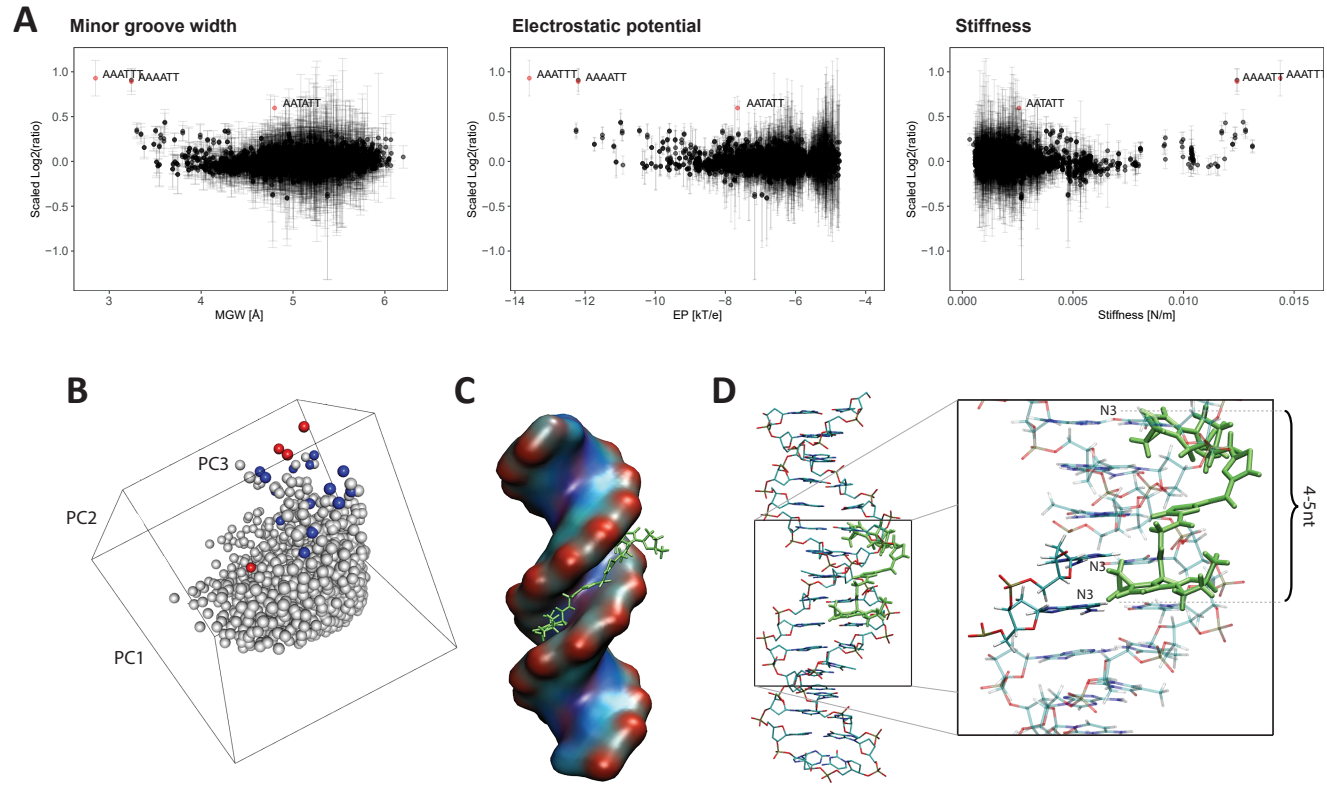
63. Ivani, I., *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat Methods* **13**, 55-58 (2016).
64. Case, D., *et al. AMBER 18. University of California, San Francisco* (2018).
65. Roe, D.R. & Cheatham, T.E., 3rd. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084-3095 (2013).
66. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, 27-38 (1996).
67. Ellrott, K., *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281.e277 (2018).
68. Giannakis, M., *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep* **15**, 857-865 (2016).
69. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289-300 (1995).
70. Katainen, R., *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-821 (2015).
71. Cibulskis, K., *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013).
72. Katainen, R., *et al.* Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer. *Nat Protoc* **13**, 2580-2600 (2018).
73. Alexandrov, L.B., *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv*, 322859 (2018).
74. Lee-Six, H., *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537 (2019).
75. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol* **17**, 31 (2016).
76. Tamborero, D., *et al.* Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med* **10**, 25 (2018).
77. Bailey, M.H., *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318 (2018).
78. Tate, J.G., *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-d947 (2019).
79. Alexandrov, L.B., *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
80. Gaffney, D.J., *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet* **8**, e1003036 (2012).
81. R Core Team. : A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org*.

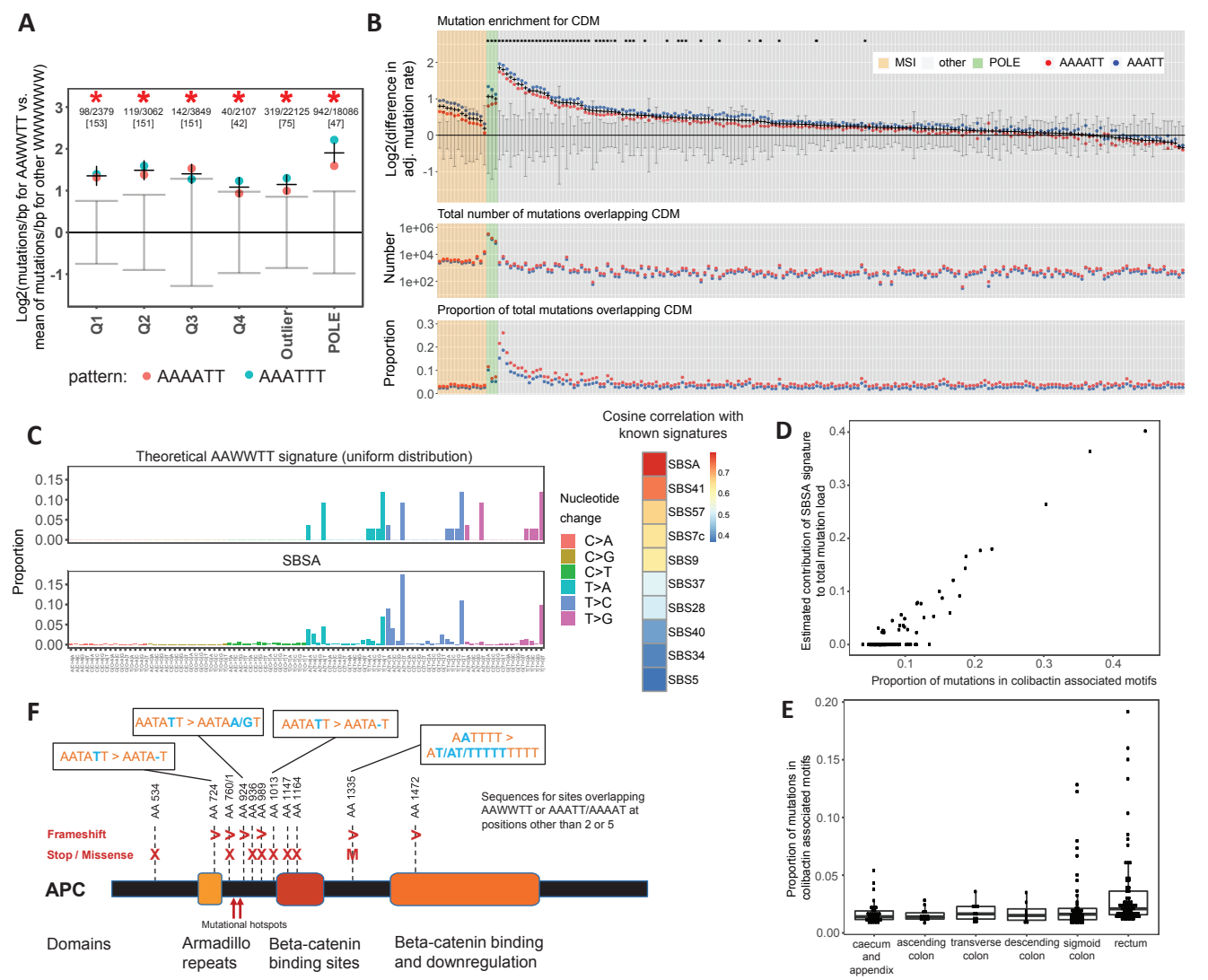

A
Caco-2 cells
Infection with E.coli and controls
M1/5 ΔclbR E.coli (pks-)
M1/5 WT E.coli (pks+)
Etoposide
No treatment
DSB
Cell fixation
(DSB preservation)
DNA blunting
In situ adapter ligation
(DSB labelling)
NNNNN
DNA extraction
DNA shearing
T7 promoter
In vitro transcription
BLISS PROTOCOL
Library preparation & sequencing
Illumina P7
Sample barcode
Illumina P5
Genomic DNA
UMI
Data analysis:
DSB identification
Filtering
Downstream analysis

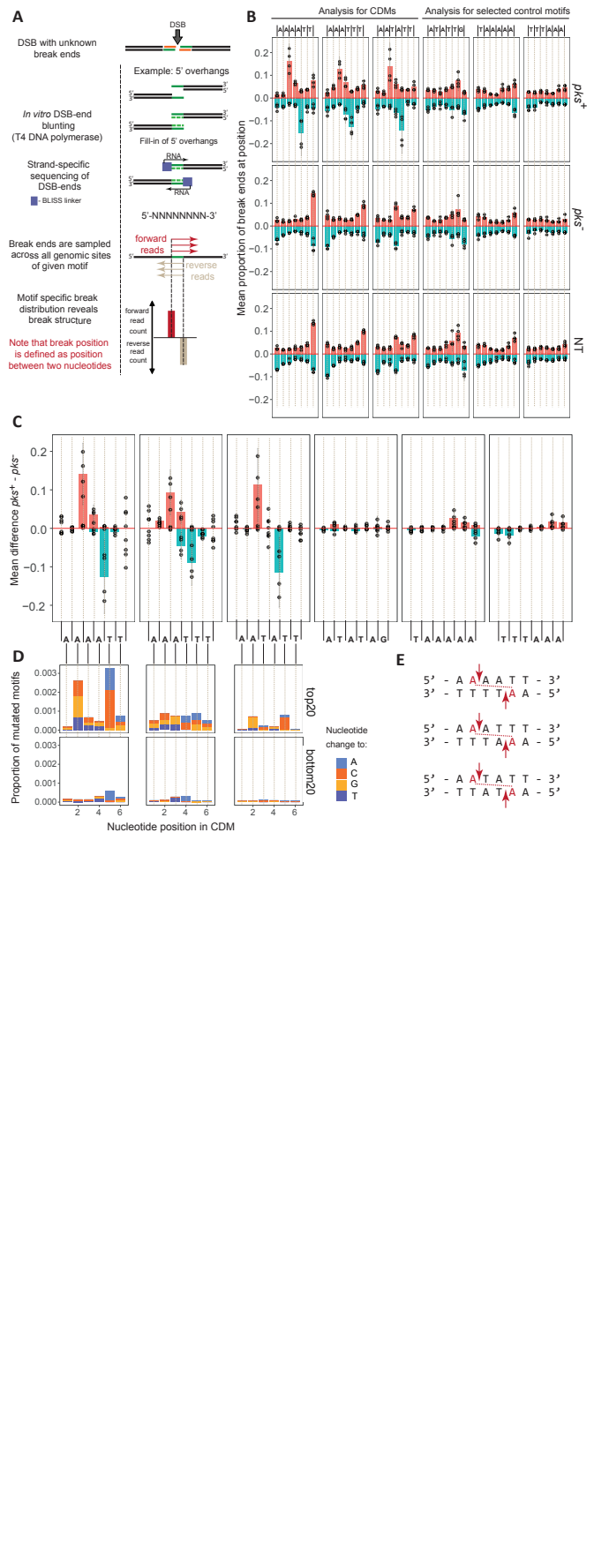

B
DSB identified by BLISS
10nt
10nt
genomic DNA
DREME motif search
Discovered CDM:
AAWWTT
W (weak): A or T
Bits
Logo:


C
DSB identified by BLISS
7nt
7nt
genomic DNA
hexanucleotide presence (yes/no)
ratio: all hexanucleotide counts from sample/ all hexanucleotide counts from control
Most enriched hexanucleotides in pks+ dataset:
AAATTT
AATTTT/AAAATT
AATATT


D
pks+ vs. pks-
Mean Log2 Ratio
All Hexanucleotide sequences
Top 50 hexanucleotide sequences
AT content in hexanucleotide sequence:
AT=5
AT=6
other
pks- vs. NT
Mean Log2 Ratio
All Hexanucleotide sequences
Top 50 hexanucleotide sequences


A
Minor groove width
Electrostatic potential
Stiffness
AAATTT
AAAATT
AATATT
Scaled Log2(ratio)
MGW [Å]
EP [kT/e]
Stiffness [N/m]
B
PC3
PC2
PC1
C
D
N3
N3
N3
4-5nt


A
Log2(mutations/bp for AAWWTT vs. mean of mutations/bp for other WWWWWW)
98/2379 [153]
119/3062 [151]
142/3849 [151]
40/2107 [42]
319/22125 [75]
942/18086 [47]
Q1
Q2
Q3
Q4
Outlier
POLE
pattern: AAAATT AAATTT
B
Mutation enrichment for CDM
Log2(difference in adj. mutation rate)
MSI other POLE AAAATT AAATT
Total number of mutations overlapping CDM
Number
Proportion of total mutations overlapping CDM
Proportion
C
Theoretical AAWWTT signature (uniform distribution)
SBSA
Proportion
Nucleotide change
C>A
C>G
C>T
T>A
T>C
T>G
Cosine correlation with known signatures
SBSA
SBS41
SBS57
SBS7c
SBS9
SBS37
SBS28
SBS40
SBS34
SBS5
D
Estimated contribution of SBSA signature to total mutation load
Proportion of mutations in colibactin associated motifs
E
Proportion of mutations in colibactin associated motifs
caecum and appendix
ascending colon
transverse colon
descending colon
sigmoid colon
rectum
F
AATATT > AATAA/GT
AATATT > AATA-T
AATATT > AATA-T
AATTTT > AT/AT/TTTTTTTTT
AA 534
AA 724
AA 760/1
AA 924
AA 939
AA 989
AA 1013
AA 1147
AA 1164
AA 1335
AA 1472
Sequences for sites overlapping AAWWTT or AAATT/AAAAT at positions other than 2 or 5
Frameshift
Stop / Missense
APC
Mutational hotspots
Domains
Armadillo repeats
Beta-catenin binding sites
Beta-catenin binding and downregulation


A
DSB with unknown break ends
DSB
Example: 5' overhangs
In vitro DSB-end blunting (T4 DNA polymerase)
Fill-in of 5' overhangs
Strand-specific sequencing of DSB-ends
BLISS linker
RNA
5'-NNNNNNNN-3'
Break ends are sampled across all genomic sites of given motif
forward reads
reverse reads
Motif specific break distribution reveals break structure
forward read count
reverse read count
Note that break position is defined as position between two nucleotides
B
Analysis for CDMs
Analysis for selected control motifs
Mean proportion of break ends at position
pks+
pks−
NT
C
Mean difference pks+ - pks−
D
Proportion of mutated motifs
top20
bottom20
Nucleotide position in CDM
Nucleotide change to:
A
C
G
T
E
5' - A A A T T - 3'
3' - T T T A A - 5'

BLISS derived DSB sites in ***pks*⁺ vs. *pks*⁻** *E.coli* infected cells

a)

Colibactin damage motif (CDM)

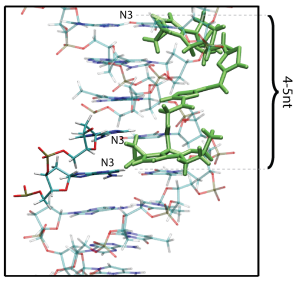
AAWWTT

b)

Structure model

DNA shape

Topological insights

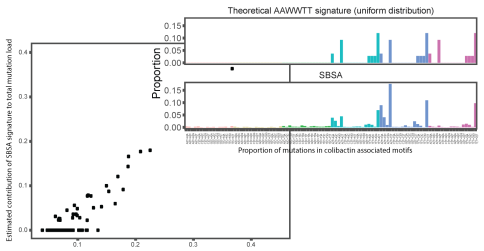


c)

Association of CDM with somatic mutations in cancer genomes

Observed enrichment of CDM associated mutations in distinct cancers

Strong correlation of CDM associated mutations with known SNV mutational signatures

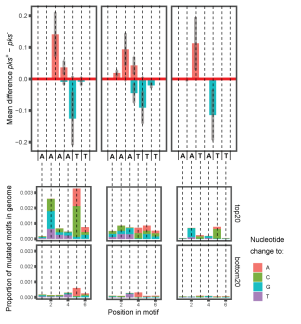


d)

Strand-specific BLISS derived break-ends

Exact break positions determined in CDM

Exact break positions correspond to locations of cancer mutations

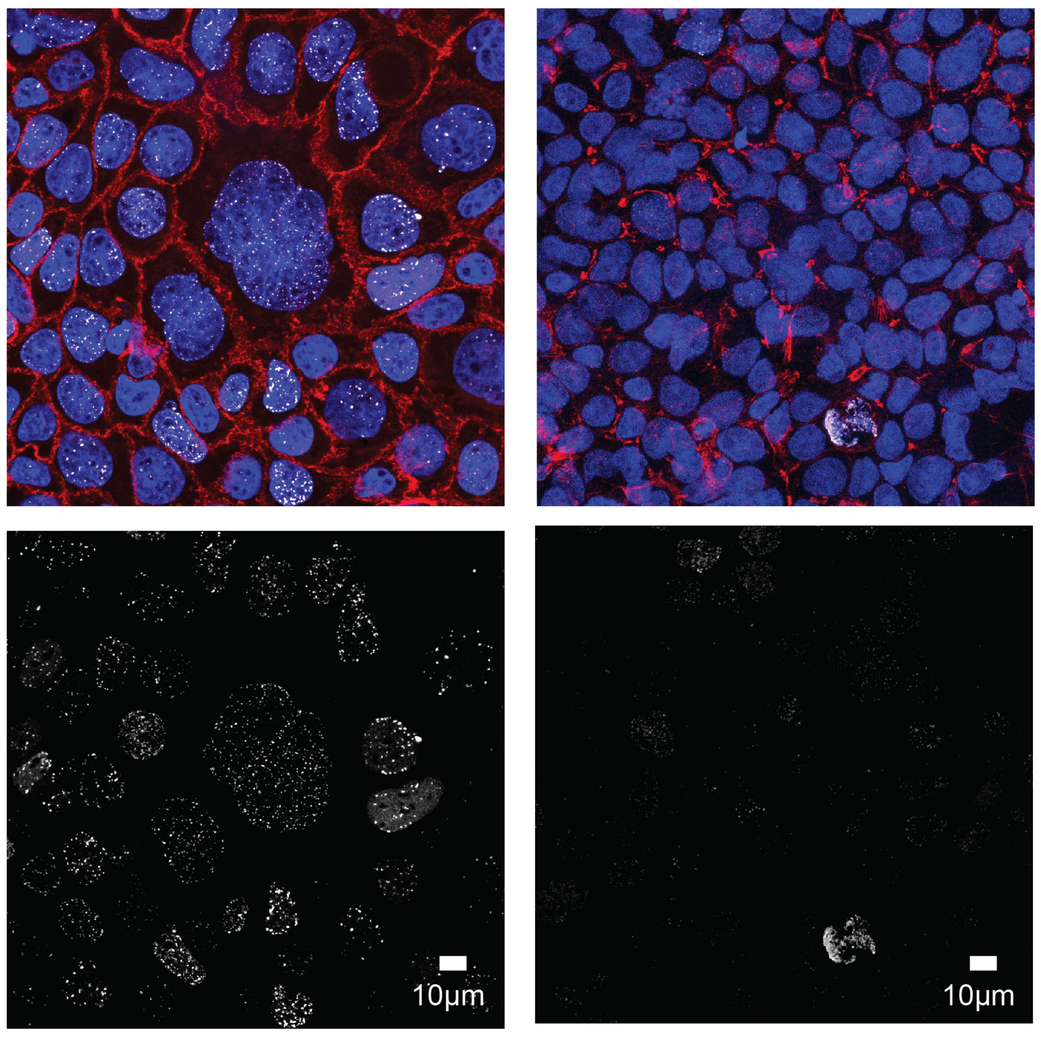


e)

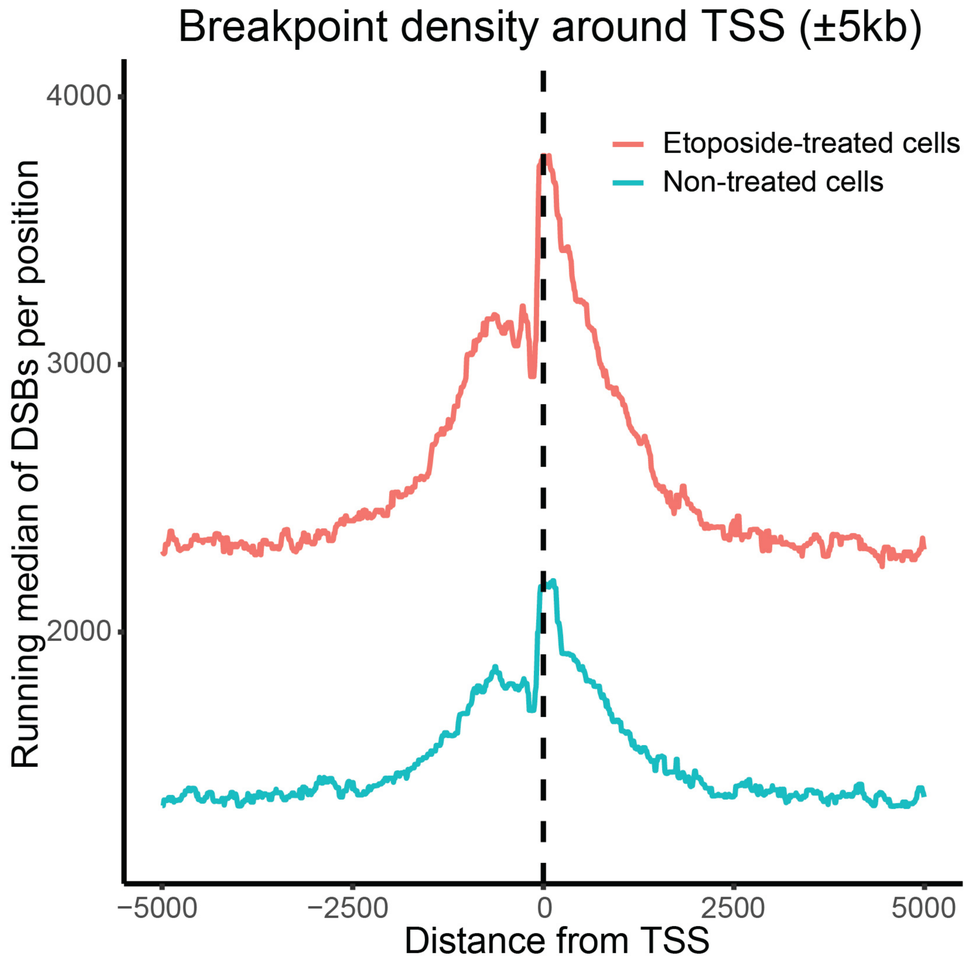
5’ - A A W W T T - 3’
3’ - T T W W A A - 5’

the CDM is the preferred target sequence of colibactin, where DSBs are generated and distinct mutational signatures are found in certain human cancers

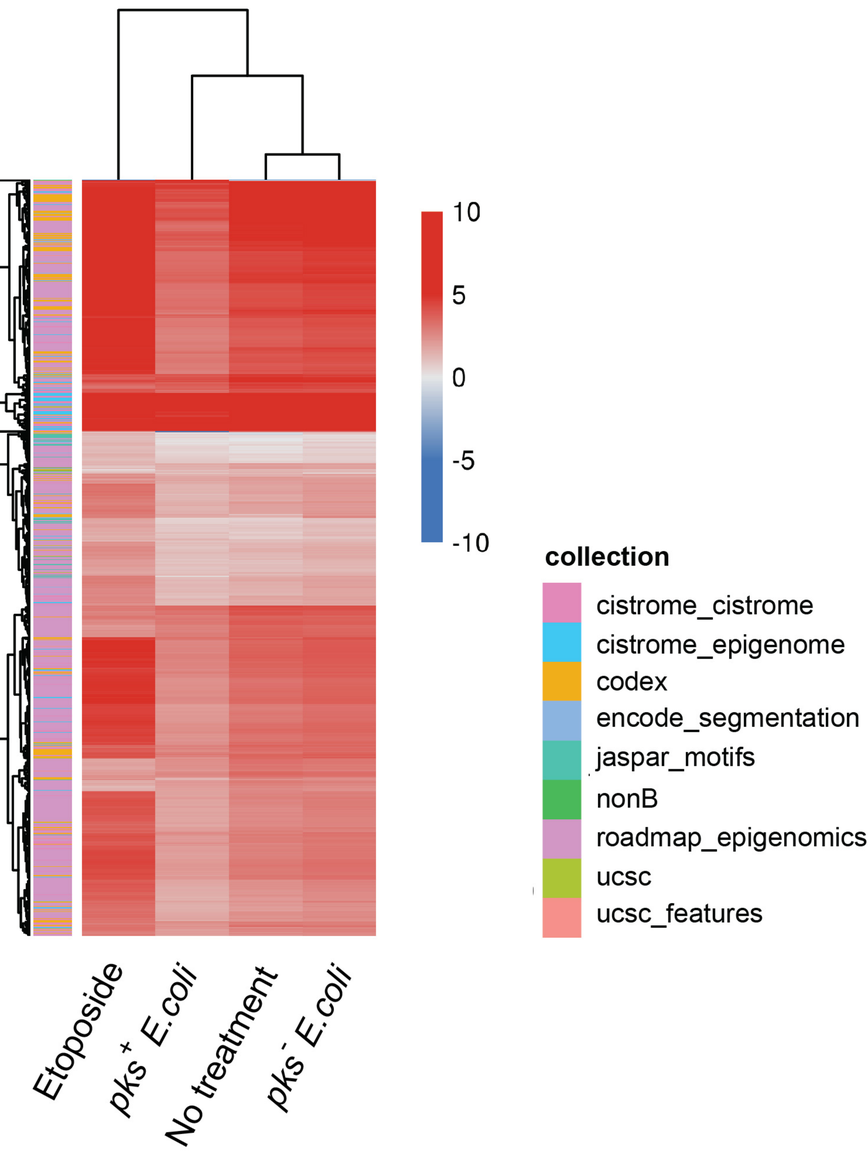
A

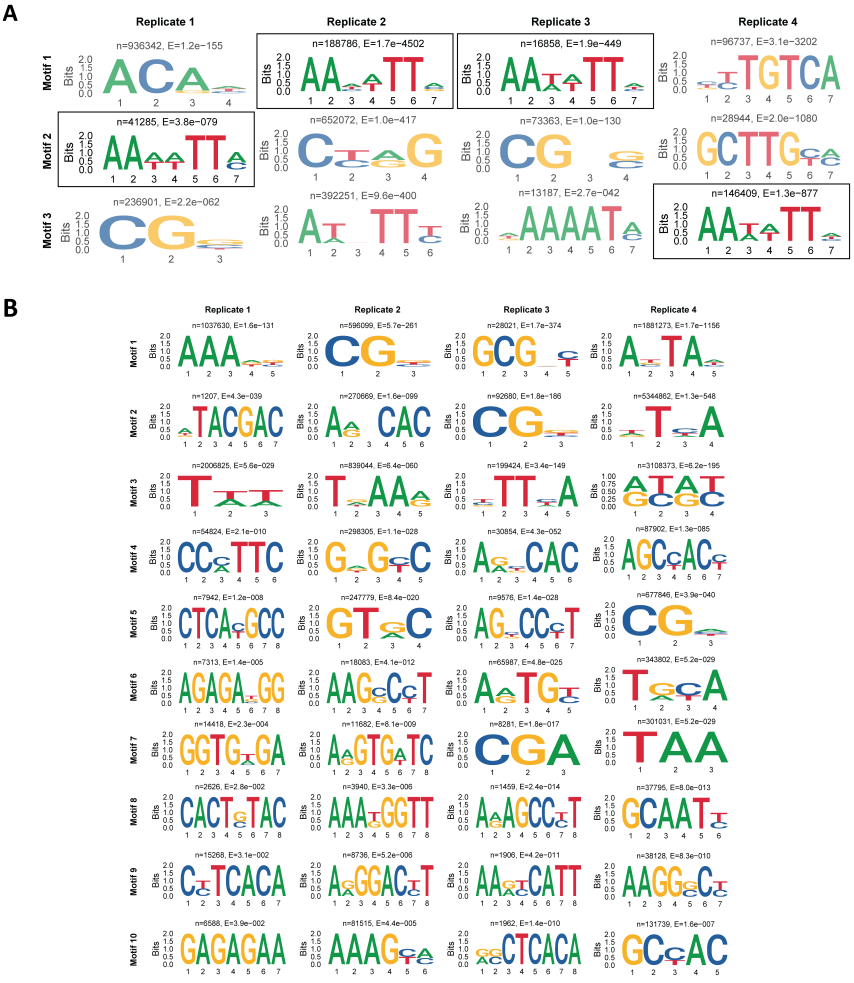

pks+ E.coli infection
ΔclbR E.coli infection
10μm
10μm
White: γH2AX, blue: nuclei, red: phalloidin

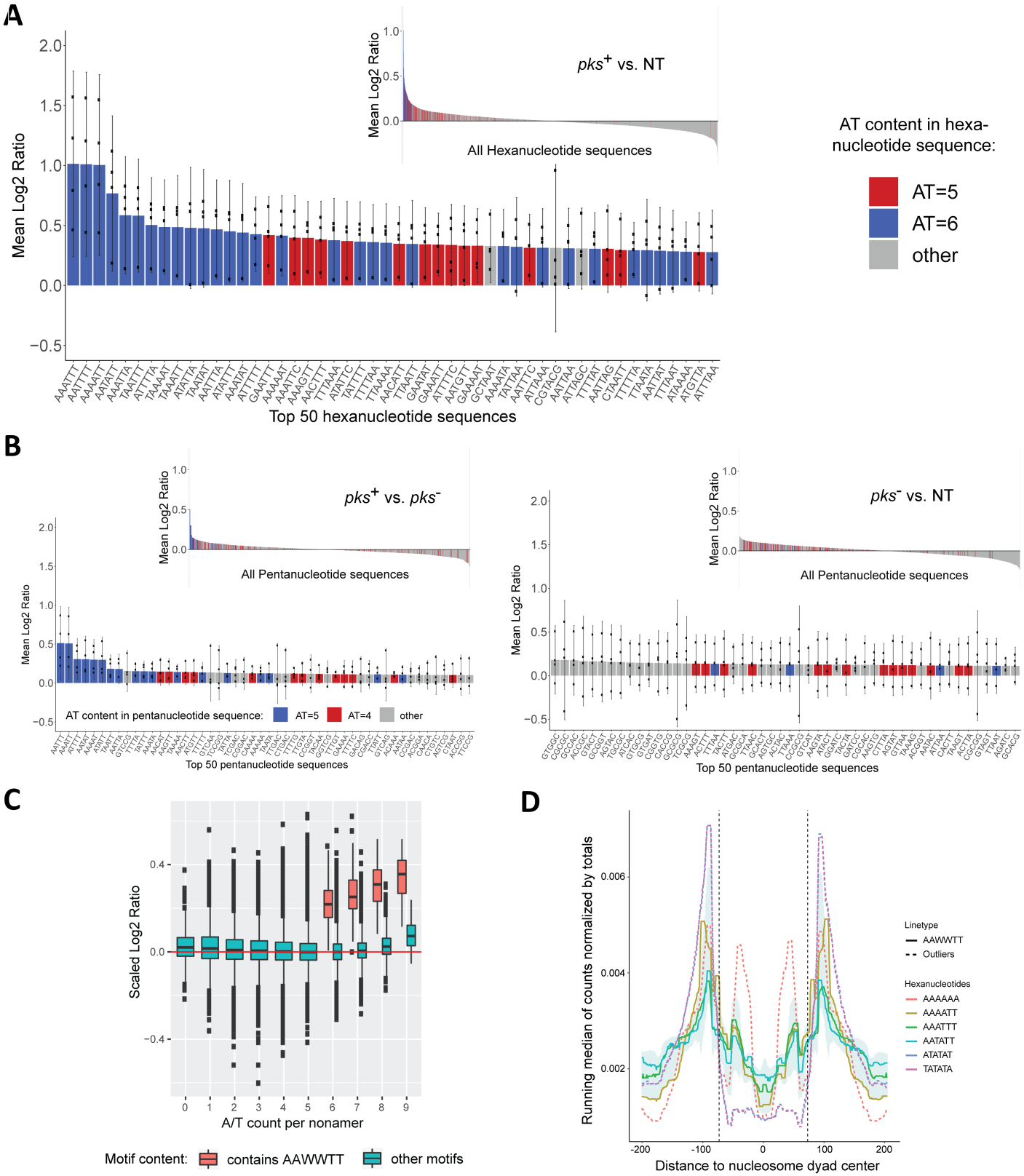
B

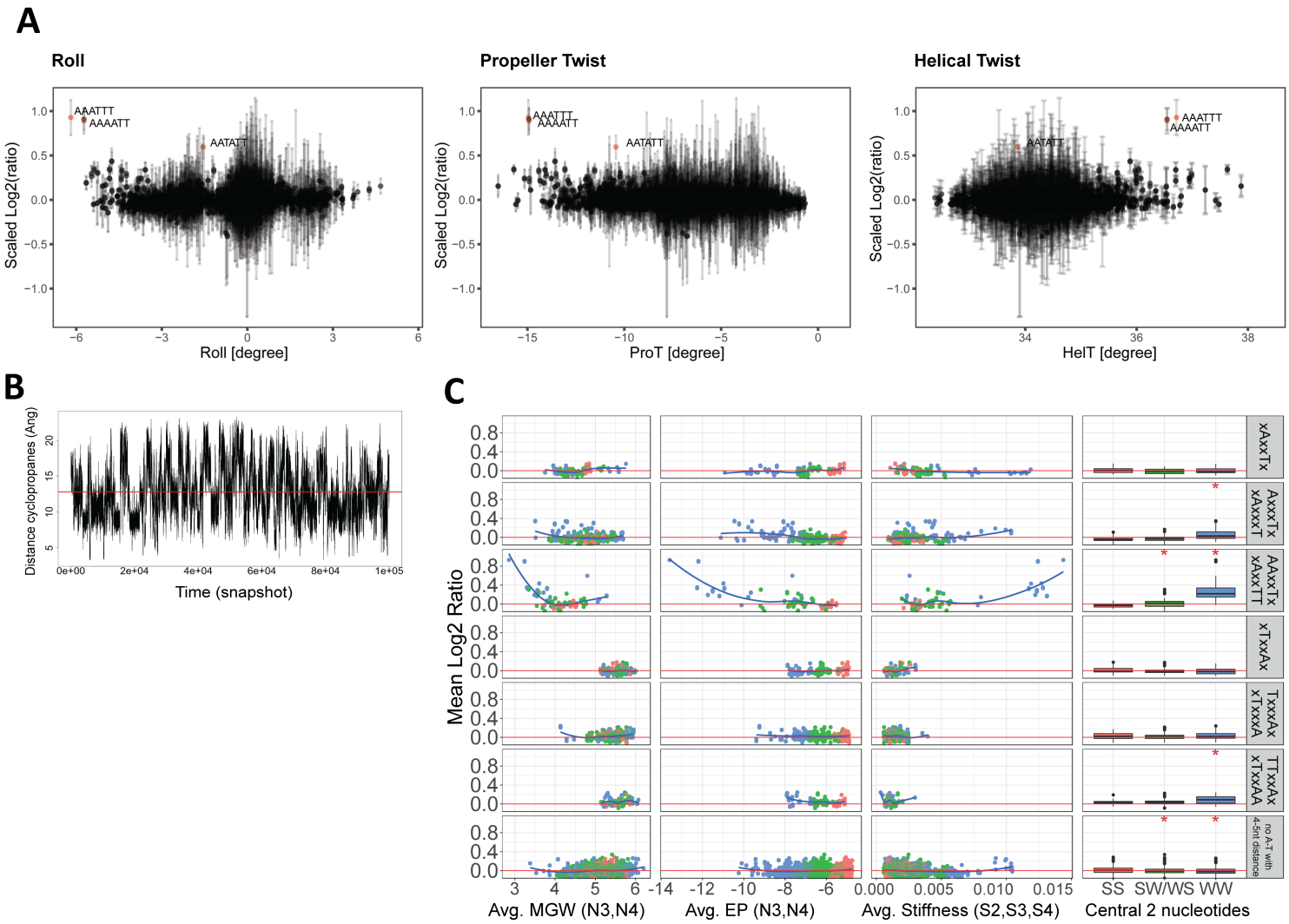

Breakpoint density around TSS (±5kb)
Etoposide-treated cells
Non-treated cells
Running median of DSBs per position
4000
3000
2000
−5000
−2500
0
2500
5000
Distance from TSS

C

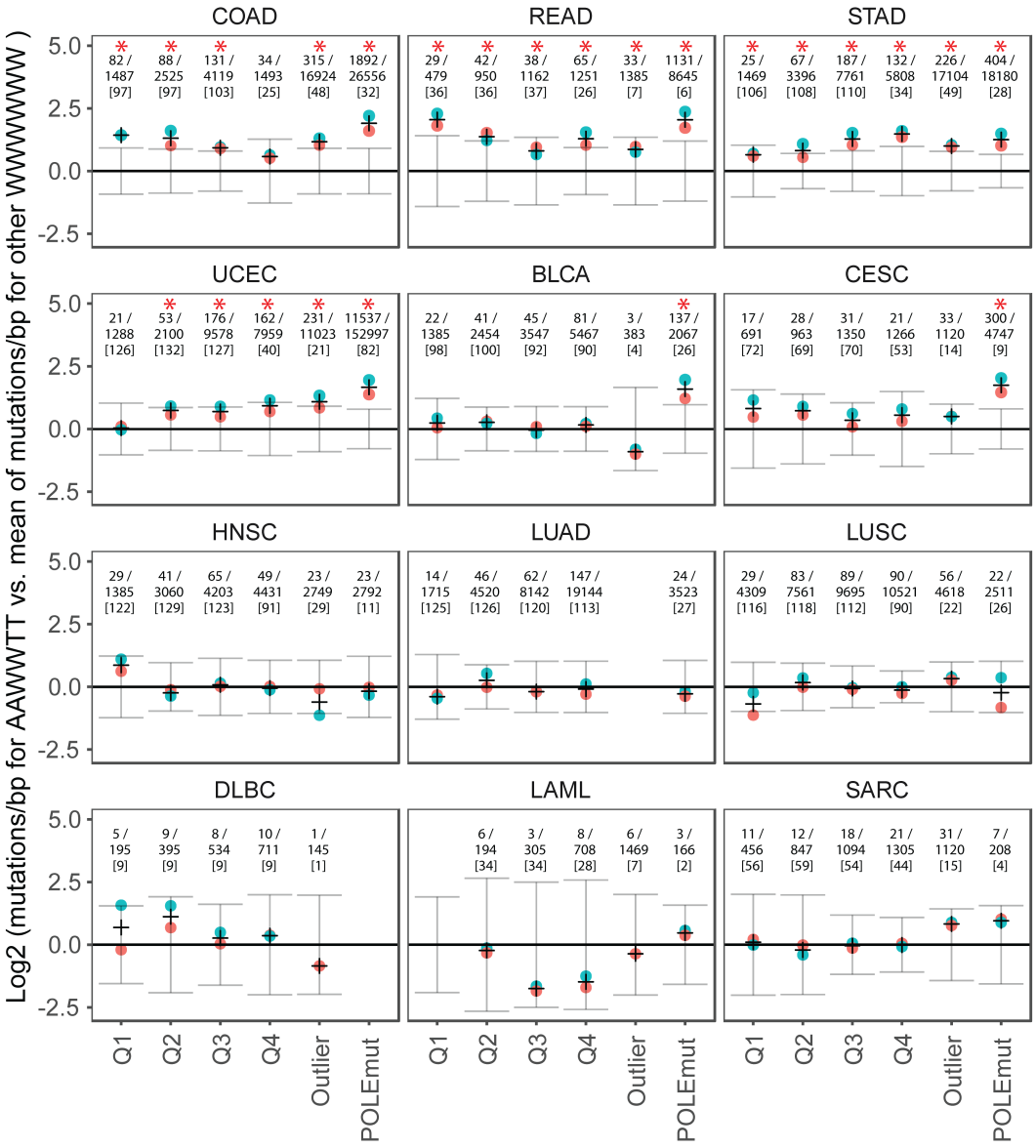

10
5
0
-5
-10
collection
cistrome_cistrome
cistrome_epigenome
codex
encode_segmentation
jaspar_motifs
nonB
roadmap_epigenomics
ucsc
ucsc_features
Etoposide
pks+ E.coli
No treatment
pks- E.coli

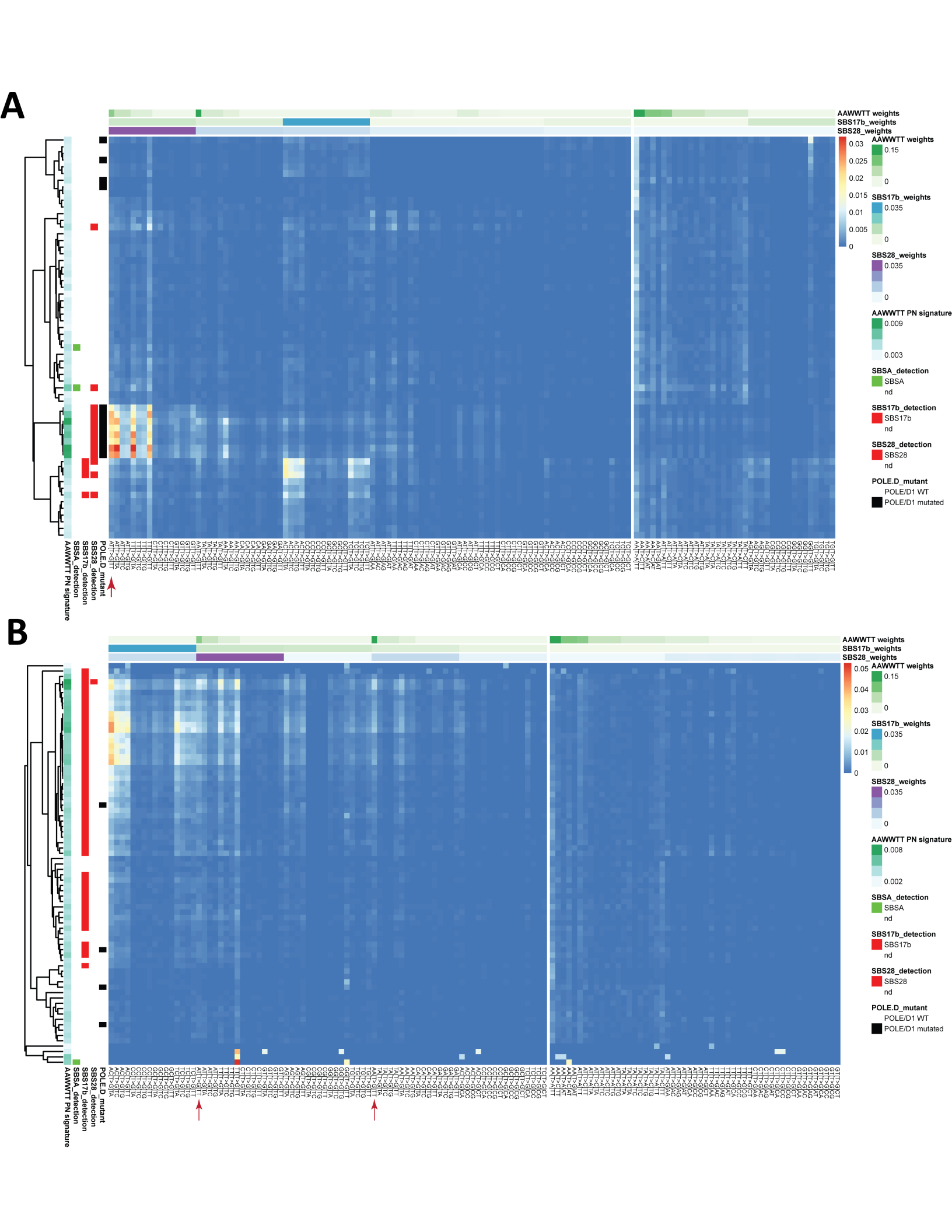

A
Replicate 1
Replicate 2
Replicate 3
Replicate 4
Motif 1
Motif 2
Motif 3
n=936342, E=1.2e−155
n=188786, E=1.7e−4502
n=16858, E=1.9e−449
n=96737, E=3.1e−3202
n=41285, E=3.8e−079
n=652072, E=1.0e−417
n=73363, E=1.0e−130
n=28944, E=2.0e−1080
n=236901, E=2.2e−062
n=392251, E=9.6e−400
n=13187, E=2.7e−042
n=146409, E=1.3e−877
Bits
B
Replicate 1
Replicate 2
Replicate 3
Replicate 4
Motif 1
Motif 2
Motif 3
Motif 4
Motif 5
Motif 6
Motif 7
Motif 8
Motif 9
Motif 10
n=1037630, E=1.6e−131
n=596099, E=5.7e−261
n=28021, E=1.7e−374
n=1881273, E=1.7e−1156
n=1207, E=4.3e−039
n=270669, E=1.6e−099
n=92680, E=1.8e−186
n=5344862, E=1.3e−548
n=2006825, E=5.6e−029
n=839044, E=6.4e−060
n=199424, E=3.4e−149
n=3108373, E=6.2e−195
n=54824, E=2.1e−010
n=298305, E=1.1e−028
n=30854, E=4.3e−052
n=87902, E=1.3e−085
n=7942, E=1.2e−008
n=247779, E=8.4e−020
n=9576, E=1.4e−028
n=677846, E=3.9e−040
n=7313, E=1.4e−005
n=18083, E=4.1e−012
n=65987, E=4.8e−025
n=343802, E=5.2e−029
n=14418, E=2.3e−004
n=11682, E=8.1e−009
n=8281, E=1.8e−017
n=301031, E=5.2e−029
n=2626, E=2.8e−002
n=3940, E=3.3e−006
n=1459, E=2.4e−014
n=37795, E=8.0e−013
n=15268, E=3.1e−002
n=8736, E=5.2e−006
n=1906, E=4.2e−011
n=38128, E=8.3e−010
n=6588, E=3.9e−002
n=81515, E=4.4e−005
n=1962, E=1.4e−010
n=131739, E=1.6e−007
Bits


A
Mean Log2 Ratio
pks+ vs. NT
All Hexanucleotide sequences
Top 50 hexanucleotide sequences
AT content in hexa-nucleotide sequence:
AT=5
AT=6
other
B
pks+ vs. pks-
All Pentanucleotide sequences
AT content in pentanucleotide sequence:
AT=5
AT=4
other
Top 50 pentanucleotide sequences
pks- vs. NT
All Pentanucleotide sequences
Top 50 pentanucleotide sequences
C
Scaled Log2 Ratio
A/T count per nonamer
Motif content:
contains AAWWTT
other motifs
D
Running median of counts normalized by totals
Distance to nucleosome dyad center
Linetype
AAWWTT
Outliers
Hexanucleotides
AAAAAA
AAAATT
AAATTT
AATATT
ATATAT
TATATA

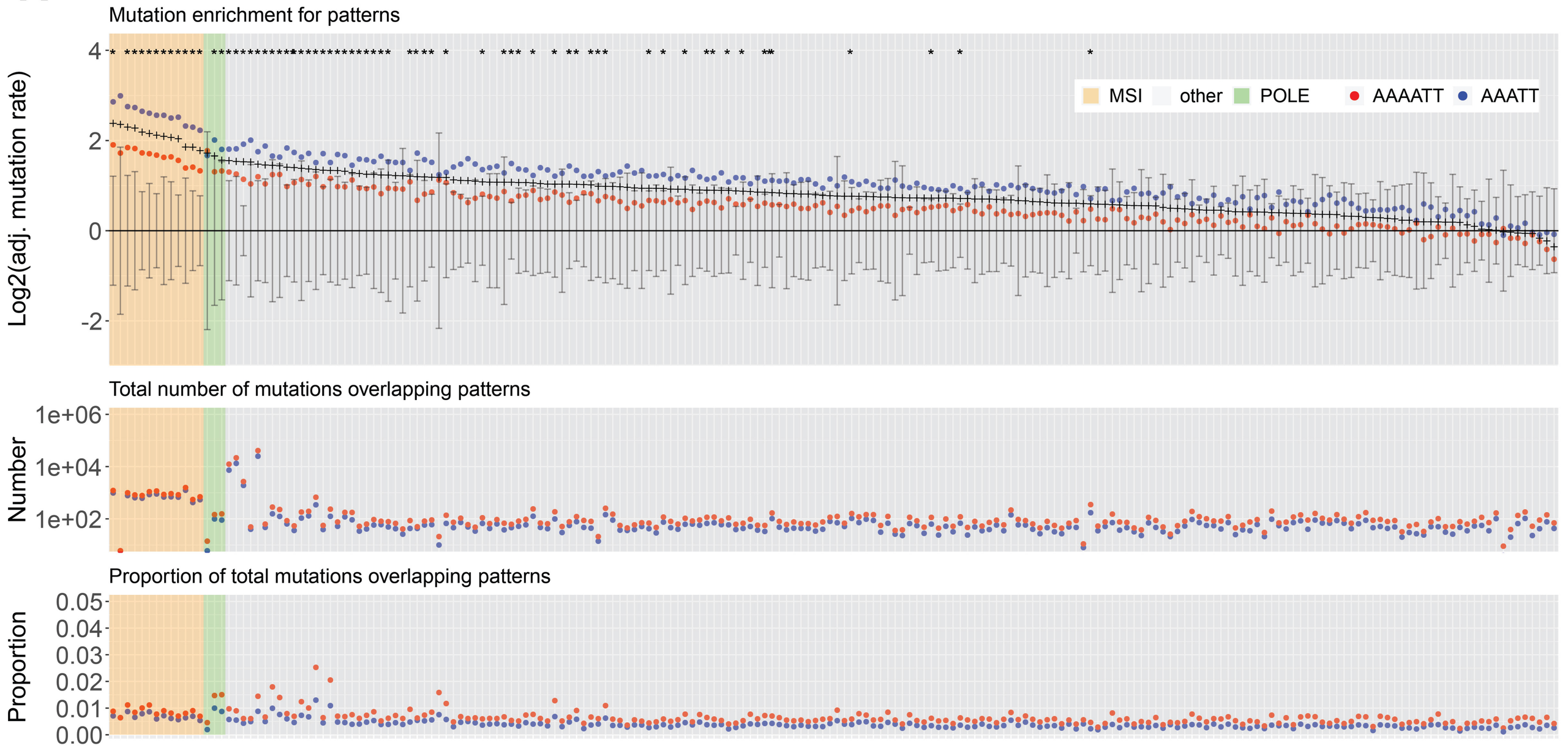

A
Roll
Propeller Twist
Helical Twist
AAATTT
AAAATT
AATATT
Scaled Log2(ratio)
Roll [degree]
ProT [degree]
HelT [degree]
B
Distance cyclopropanes (Ang)
Time (snapshot)
C
Mean Log2 Ratio
Avg. MGW (N3,N4)
Avg. EP (N3,N4)
Avg. Stiffness (S2,S3,S4)
Central 2 nucleotides
SS
SW/WS
WW
xAxxTx
AxxxTx
AAxxTx
xTxxAx
TxxxAx
TTxxAx
no A-T with 4-5nt distance

A

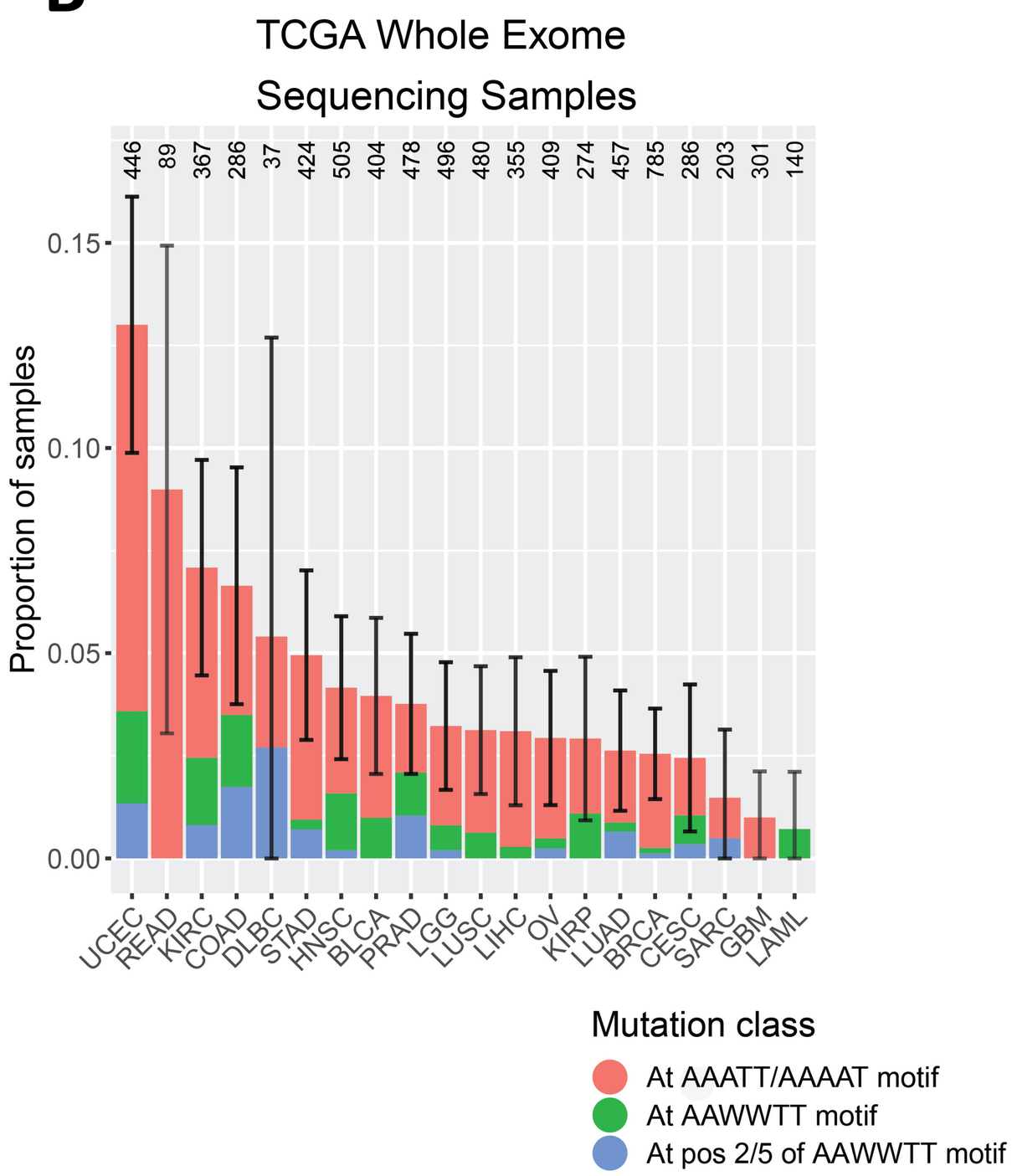

COAD
READ
STAD
UCEC
BLCA
CESC
HNSC
LUAD
LUSC
DLBC
LAML
SARC
Log2 (mutations/bp for AAWWTT vs. mean of mutations/bp for other WWWWWW)
5.0
2.5
0.0
-2.5
Q1
Q2
Q3
Q4
Outlier
POLEmut

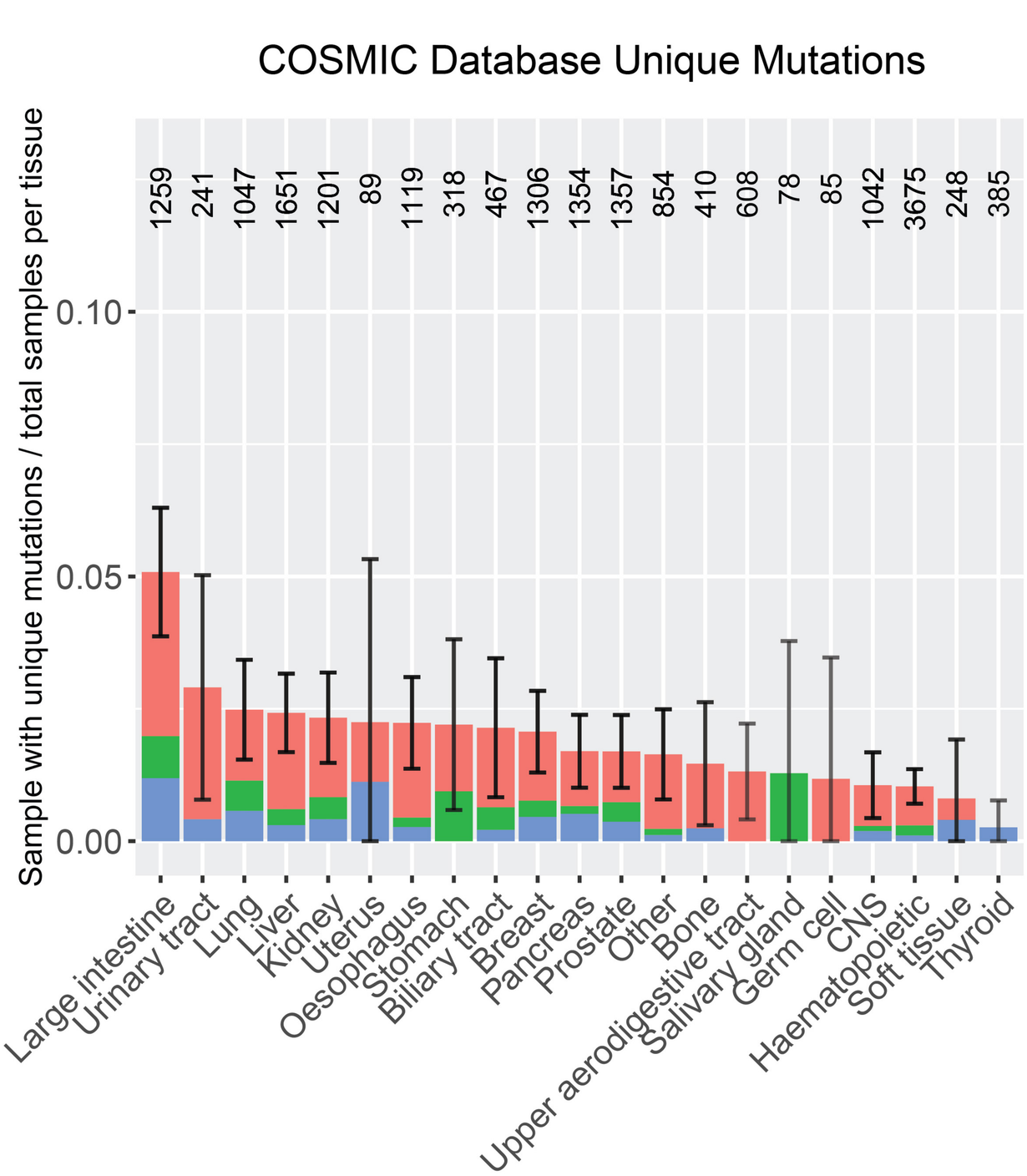

A
AAWWTT weights
SBS17b_weights
SBS28_weights
AAWWTT PN signature
SBSA_detection
SBS17b_detection
SBS28_detection
POLE.D_mutant
B
AAWWTT weights
SBS17b_weights
SBS28_weights
AAWWTT PN signature
SBSA_detection
SBS17b_detection
SBS28_detection
POLE.D_mutant

A


Mutation enrichment for patterns
Log2(adj. mutation rate)
4
2
0
-2
MSI
other
POLE
AAAATT
AAATT
Total number of mutations overlapping patterns
Number
1e+06
1e+04
1e+02
Proportion of total mutations overlapping patterns
Proportion
0.05
0.04
0.03
0.02
0.01
0.00

B


TCGA Whole Exome
Sequencing Samples
446
89
367
286
37
424
505
404
478
496
480
355
409
274
457
785
286
203
301
140
Proportion of samples
0.15
0.10
0.05
0.00
UCEC
READ
KIRC
COAD
DLBC
STAD
HNSC
BLCA
PRAD
LGG
LUSC
LIHC
OV
KIRP
LUAD
BRCA
CESC
SARC
GBM
LAML


COSMIC Database Unique Mutations
Sample with unique mutations / total samples per tissue
0.10
0.05
0.00
1259
241
1047
1651
1201
89
1119
318
467
1306
1354
1357
854
410
608
78
85
1042
3675
248
385
Large intestine
Urinary tract
Lung
Liver
Kidney
Uterus
Oesophagus
Stomach
Biliary tract
Breast
Pancreas
Prostate
Other
Bone
Upper aerodigestive tract
Salivary gland
Germ cell
CNS
Haematopoietic
Soft tissue
Thyroid

Mutation class
At AAATT/AAAAT motif
At AAWWTT motif
At pos 2/5 of AAWWTT motif

A

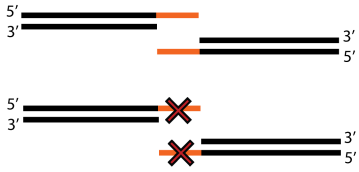
DSB occuring in the living cell
- unknown structure of break-ends
- break ends preservation with PFA

In vitro DSB-end blunting (T4 DNA polymerase)

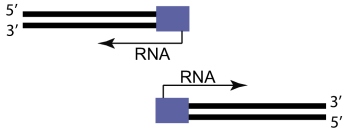
Strand-specific sequencing of DSB-ends

- BLISS linker

[1] 3' overhangs

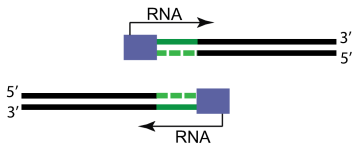


Removal of 3' overhangs

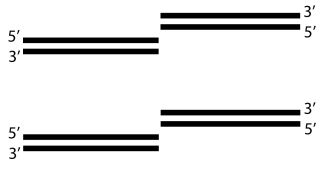


[2] 5' overhangs

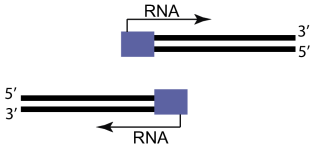
Fill-in of 5' overhangs



[3] blunt-ends



Break is preserved

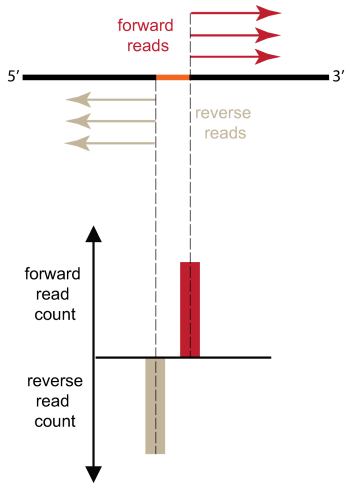


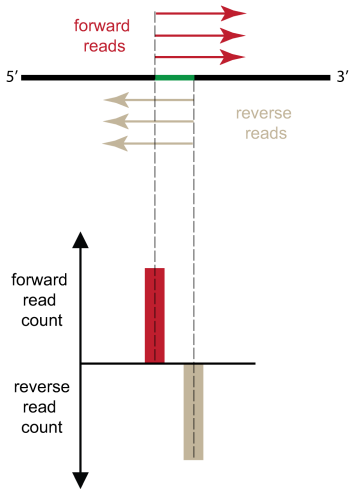
Reads are grouped based on the presence of motif of interest in the break end.
The distribution of break-ends is counted separately for each motif.

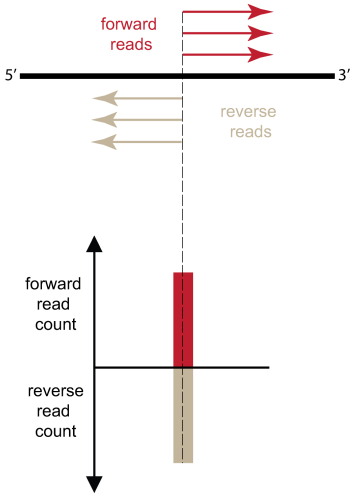
Break ends are sampled across all genomic sites of given motif

Motif specific break distribution reveals break structure

Note that break position is defined as position between two nucleotides







B