COMPOUND-SPECIFIC DNA ADDUCT PROFILING WITH NANOPORE SEQUENCING AND IONSTATS

Yrjö Koski[1,2], **Divyesh Patel**[2,3], **Natalia Kakko von Koch**[4], **Paula Jouhten**[4], **Lauri Aaltonen**[2,5,6], **Kimmo Palin**[2,5,6,*], **Biswajyoti Sahu**[2,3,7,*], and **Esa Pitkänen**[1,2,6,*]

[1]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

[2]Applied Tumor Genomics Research Program, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[3]The Norwegian Centre for Molecular Biosciences and Medicine (NCMBM), University of Oslo, Oslo, Norway

[4]Department of Bioproducts and Biosystems, School of Chemical Engineering, Aalto University, Espoo, Finland

[5]Department of Medical and Clinical Genetics, University of Helsinki, Helsinki, Finland

[6]iCAN Digital Precision Cancer Medicine Flagship, Helsinki, Finland

[7]Department of Cancer Genetics, Institute for Cancer Research, Oslo, Norway

[*]Corresponding authors. Email: kimmo.palin@helsinki.fi, biswajyoti.sahu@helsinki.fi, esa.pitkanen@helsinki.fi

ABSTRACT

Covalently bound DNA adducts are mutation precursors that contribute to aging and diseases such as cancer. Accurate detection of adducts in the genome will shed light on tumorigenesis. Commonly used detection methods are unable to pinpoint the exact genomic locations of adducts. Long-read nanopore sequencing has the potential to accurately detect multiple types of DNA adducts at single-nucleotide precision. In this study, we developed a novel statistical toolkit, IonStats, to profile DNA adducts in nanopore sequencing data. With IonStats, we investigated the effects of four adduct-inducing genotoxic compounds on nanopore sequencing, and found both shared and compound-specific perturbations in base quality scores, ionic current profiles, and translocation dynamics. Notably, aristolochic acid II and melphalan treatments profoundly altered nanopore readouts and led to substantial sequence-specific read interruptions. Our study shows that nanopore sequencing can be effectively employed to detect and characterize DNA adducts, paving the way for high-resolution, high-throughput profiling of DNA damage and the exposome.

Keywords Nanopore sequencing · DNA adducts · DNA adduct detection · genotoxicity · IonStats

1 Introduction

DNA adducts are covalently bound lesions caused by exogenous and endogenous genotoxic compounds that react with DNA directly or after metabolic activation [1]. Enzymatic DNA repair mechanisms remove most adducts, but a small fraction may escape the repair [2]. Adducts can interfere with DNA transcription and replication, and give rise to mutations that may lead to cancer [3]. Several compounds have been implicated in cancer initiation, including aristolochic acids (AAs), polycyclic aromatic hydrocarbons (PAHs), and aflatoxin B_1 (AFB_1), linked to urological, lung, and liver cancers, respectively [4, 5, 6, 7]. Beyond cancer, DNA adducts serve as molecular imprints of environmental exposures, contributing to accelerated biological aging and increased mortality [8]. Detecting and characterizing DNA adducts is thus crucial for understanding how exposure to DNA damage can lead to mutations and cellular dysfunction underlying both carcinogenesis and the progressive decline associated with aging.

There are two distinct classes of DNA adducts: monoadducts and crosslinks [9]. Monoadducts are formed through chemical modifications by small molecules on individual nucleotides. The primary targets of monoadduct binding depend on the reactive species, nucleophilicity of DNA sites, steric factors, and DNA sequence context. For small alkylating agents, the primary binding sites are the ring nitrogens N3 and N7, which are most nucleophilic in purine bases. For bulky aromatic compounds, such as AAs, the primary binding sites are exocyclic amino groups [10]. Molecules with multiple reacting groups can form DNA-DNA crosslinks, where two nucleotides are bound by the adduct [11]. Crosslinks can be either inter- or intrastrand, where the bound nucleotides are on different strands or the same strand, respectively [12, 13].

The most prominent methods for DNA adduct detection include mass spectrometry- and amplicon-based methods [14]. High-resolution mass spectrometry enables comprehensive profiling of adduct exposure, detecting most DNA monoadducts that are present in the sample [15, 16]. However, these methods require the DNA to be broken down into individual nucleotides, losing the positional information of the adducts in the genome [14]. Amplicon-based methods solve this issue by recognizing the adducts chemically or enzymatically by using DNA repair enzymes to mark and cleave DNA adducts [17, 18]. However, amplicon-based methods are unable to distinguish between different types of adducts [14].

Nanopore sequencing is an emerging technology with the potential to differentiate between adduct types at single-nucleotide resolution while preserving site-specificity [19, 20]. Nanopore sequencing measures the ionic current as a single-stranded DNA fragment translocates through a nanopore, which is embedded in an electrically insulating membrane. The nucleobases that reside inside the nanopore at any given time alter the ionic current based on their physical structure. Hence, analysis of the ionic current can potentially reveal both the DNA sequence and modifications such as adducts [21, 22].

Previous studies have focused on adducts only in specific known sequence contexts [19, 20]. In this study, we show that nanopore sequencing can be used to characterize adducts in arbitrary sequence contexts for multiple adduct-inducing molecules. We implemented a novel software package called IonStats for distilling nanopore sequencing data into

multiple statistics revealing deviations from the unadducted state. We show that these statistics are informative for adduct formation by creating and analyzing a novel nanopore sequencing dataset of whole genome amplified *Saccharomyces cerevisiae* DNA with four adduct-inducing genotoxic compounds: aristolochic acid II (AAII), cisplatin, melphalan, and mitomycin C (MMC). These compounds vary in size and chemistry, and bind to various nucleotides in DNA. We demonstrate distinct treatment effects by comparing the sequencing statistic distributions between treated and untreated samples. We also observed that some treatments increase read interruptions, where DNA fragments are not fully sequenced, likely due to bulky adducts or crosslinks bound at specific sequence contexts. IonStats is available open source at GitHub (`https://github.com/ykoski/ionstats`).

2 Results

2.1 IonStats: a comprehensive toolkit for capturing adduct effects in nanopore sequencing data

Nanopore sequencing generates large volumes of ionic current signal data, posing substantial challenges for downstream analyses (Fig. 1A,B). To distill informative features for adduct detection, we developed a computational workflow called IonStats (Methods). IonStats collects and summarizes diverse statistics from nanopore signal and basecalled sequencing data by k-mers or reference genome positions (Methods; Fig. 1C). Comparison of the statistics of treated samples with those of untreated samples allows the characterization of treatment effects. In addition, IonStats collects read-level statistics including read counts and lengths, and base quality scores. In our study, we used $k = 9$, or the number of nucleotides that reside within the nanopore (ONT pore version R10.4.1) at any given time [23]. This choice also ensured high k-mer coverage in our experiments.

For each k-mer or reference position, IonStats collects nanopore signal means and standard deviations, base quality scores, and dwell times. The latter indicates the number of measurements obtained for each nucleotide, correlating with the time the nucleotide resided in the nanopore sensing region (Fig. 1A,C). In addition, IonStats computes motor-protein adjusted dwell time (ADT). This statistic reflects the duration that each nucleotide resides at the motor protein, rather than within the sensing region. ADT lets us study the effects that adducts might have when interacting with the motor protein, in addition to in-pore effects. IonStats also provides a novel windowed error score called Average Expected Absolute Deviation (AEAD; Methods). AEAD captures how much the signal means or standard deviations differ from expected values within a k-mer window. Finally, adducts may cause failed translocations, which IonStats can detect by identifying interrupted reads through missing rear barcode sequences.

IonStats provides distributional tests to detect differences between sample groups, typically between treated and control samples. The available tests include the Kolmogorov-Smirnov (KS) test and the Cramér von Mises (CvM) criterion. Compared to the more common KS test, CvM is more sensitive to differences in tails of distributions. Together, these complementary tests enable robust detection of treatment-induced perturbations in nanopore data.

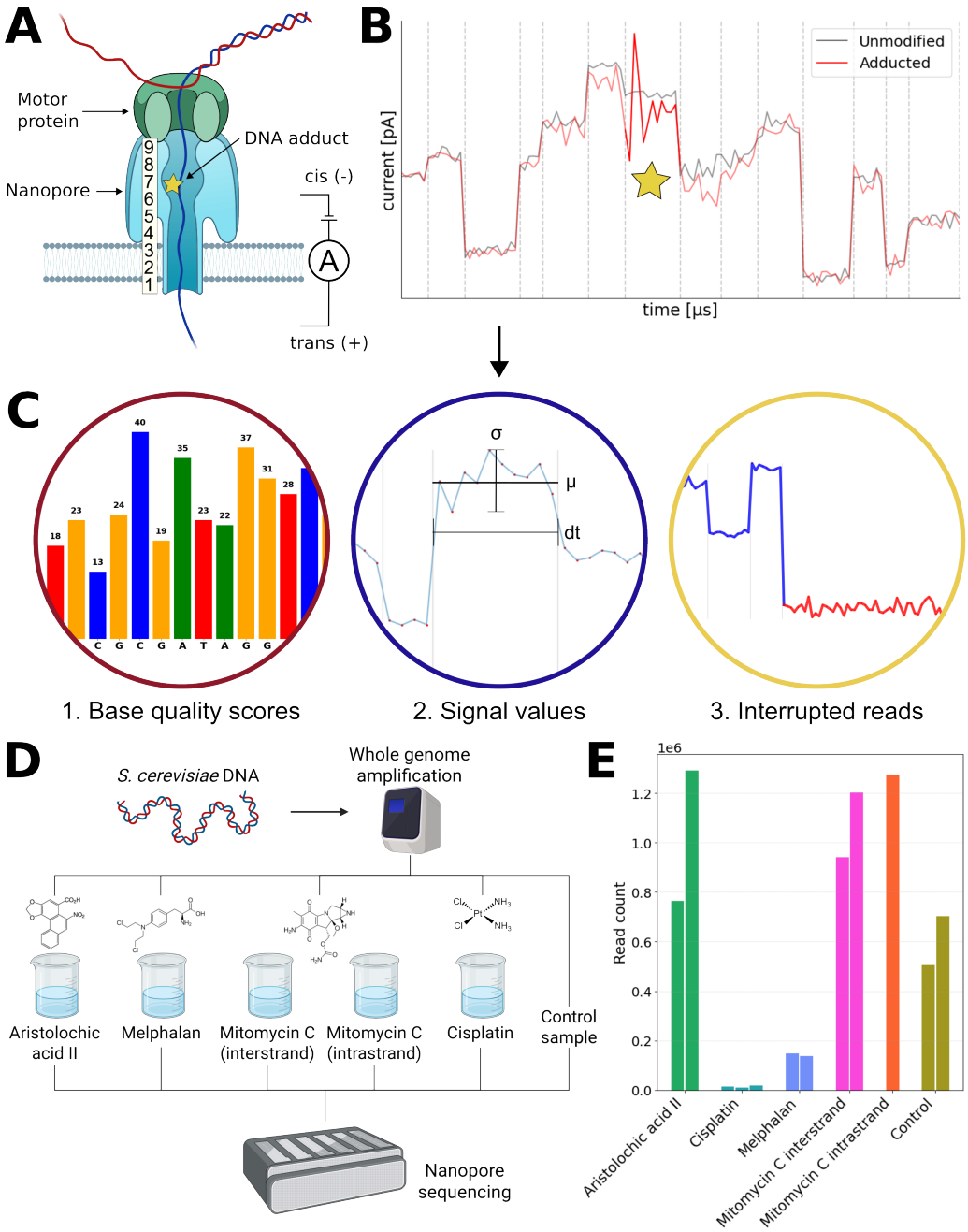


Figure 1: Key concepts and study design. **A:** Visualization of the nanopore sequencing process. Positions of the nanopore are numbered. **B:** An illustration of the effect that a DNA adduct can have on the ionic current. **C:** Overview of the data analysis approaches that IonStats uses to distill nanopore sequencing data into multiple statistics to compare treated and untreated control samples. **D:** Visualization of the experimental setup of this study. **E:** Number of reads per sample replicate obtained in the experiment.

2.2 Adduct-induced DNA damage reduces nanopore read length and quality

We first examined whether the effects of adduct treatments can be observed in read-level nanopore sequencing statistics after treating *S. cerevisiae* DNA with four genotoxic, adduct-inducing compounds (Fig. 1D; Methods). These statistics

included the number of reads sequenced (Fig. 1E), read lengths and mean base quality scores (Fig. 2), and rear barcode quality scores (Suppl. Fig. 1, Suppl. Table 1). AAII, melphalan, and cisplatin samples showed lower mean quality scores (16.6, 17.8, and 5.23, respectively; Suppl. Table 2), whereas in mitomycin samples, scores were higher (MMC interstrand 19.9, MMC intrastrand 19.9) than in untreated controls (19.6, $p < 10^{-16}$ for all compounds) (Fig. 2). Both melphalan and cisplatin samples had shorter reads (median lengths: 1833 and 3658, respectively) compared to other samples (AAII 6744, MMC interstrand 7430, MMC intrastrand 7331, control 7061; Suppl. Table 2). The replicates generally correlated very well, except for AAII, where the two replicates differed by mean quality score (means 15.8 and 17.1; $p < 10^{-16}$).

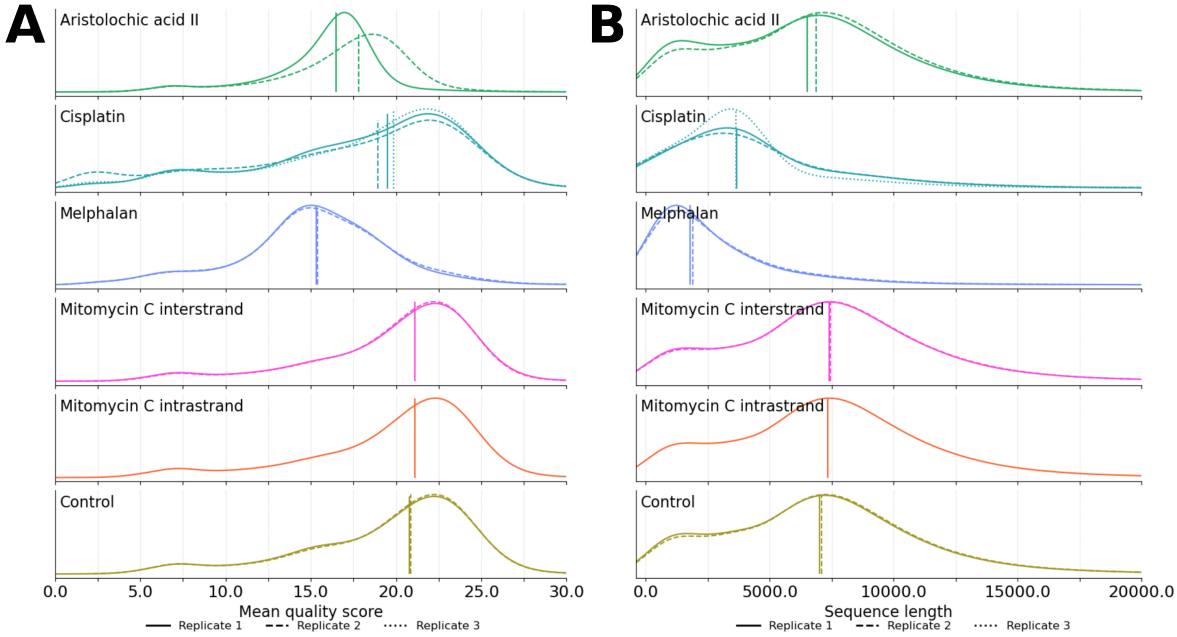


Figure 2: Overview of the sequencing statistics in the nanopore sequencing dataset. **A:** Read mean quality scores by sample. **B:** Read lengths by sample. The vertical line shows the median value of each replicate.

2.3 Treatment effects are observed in tails of nanopore signal-level statistics

We then focused on the signal-level statistics obtained with IonStats. Overall, for most treatments, the distributions of these k-mer-level statistics were consistent with those observed in controls (Suppl. Table 3 & 4, Suppl. Fig. 2). However, for AAII, melphalan, and cisplatin, the variability of many statistics differed substantially from controls (Suppl. Table 3 & 5, Suppl. Fig. 3). This led us to believe that the treatment effects are primarily observed at the tails of k-mer value distributions. Consistent with this hypothesis, the extreme quantiles (1st and 99th percentiles) of signal statistics revealed pronounced shifts (Fig. 3, Suppl. Table 5), while their means showed only minor differences (Suppl. Table 3, 4). This supported the notion that adduct formation affects only a subset of DNA nucleotides, producing a mixture of modified and unmodified k-mers (Fig. 3 A, B). Due to insufficient k-mer coverage, cisplatin samples were excluded from these analyses (Suppl. Table 1).

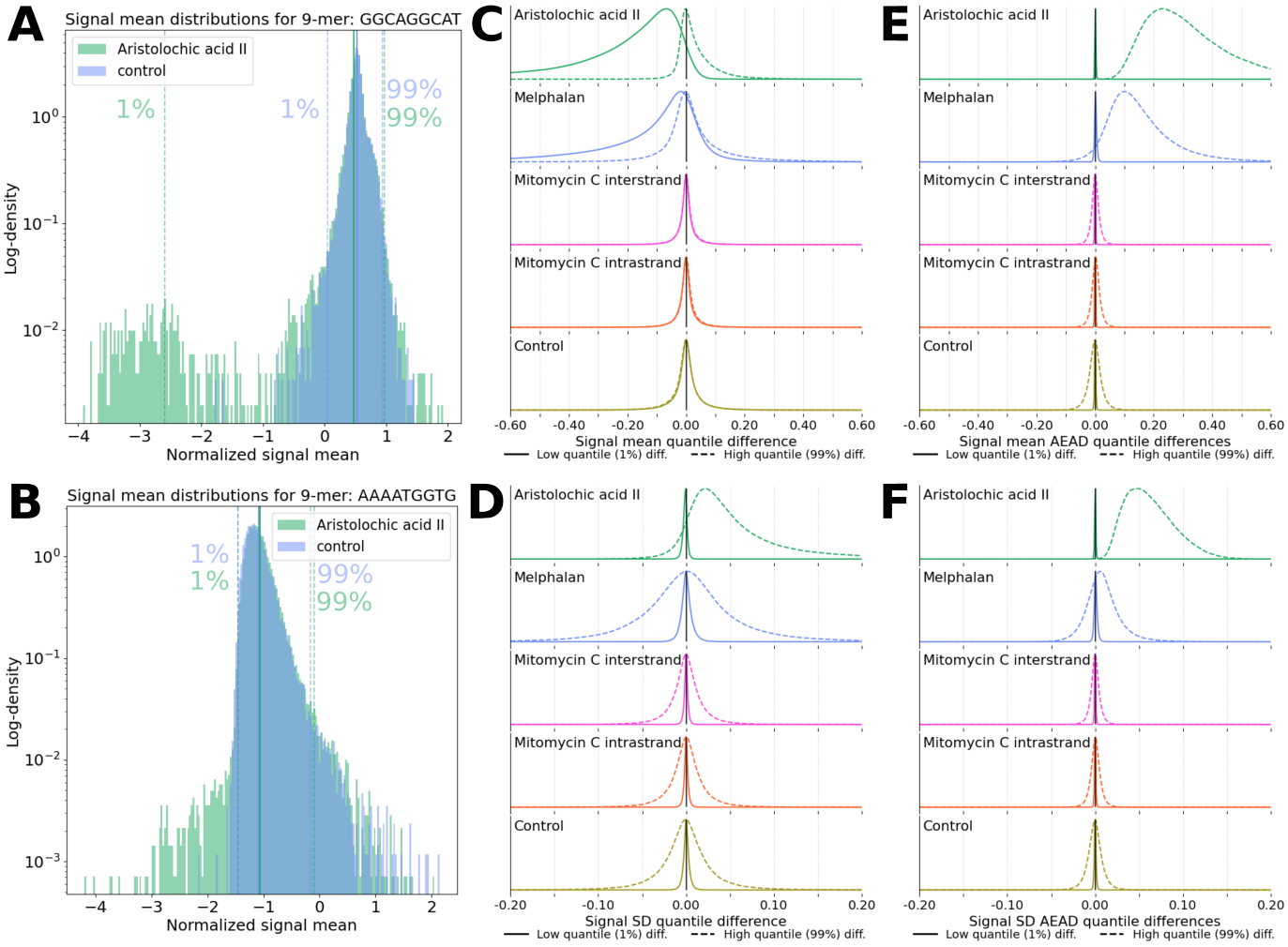


Figure 3: **A**: Nanopore signal mean distribution for the 9-mer GGCAGGCAT, where 1% quantile outliers distinguish the AAII-treated sample from controls. Dashed lines and annotations indicate the 1% and 99% quantiles; solid lines indicate the distribution means. **B**: An example of a 9-mer (AAAATGGTG) where the signal mean distribution is non-normal but highly similar between AAII-treated and control samples, with no substantial differences in the extreme quantiles. **C-F**: Quantile differences for signal mean (**C**), standard deviation (**D**), mean AEAD (**E**), and standard deviation AEAD (**F**). For treatments, differences for each k-mer are shown between treatment and controls; for controls, between-replicate differences are shown.

Among all treatments, we observed the strongest effects in AAII-treated samples, with melphalan samples showing similar but weaker effects (Fig. 3C-F; Suppl. Table 4-7). Specifically, the lowest quantile signal means for most k-mers in AAII samples were lower than those in controls (Fig. 3C). This is consistent with bulky DNA adducts impeding ion flow through the nanopore, thus lowering the measured current. We also found that many k-mers displayed markedly higher variability of signal levels, particularly in AAII samples (Fig. 3D). Furthermore, aggregating signal over multiple consecutive k-mers, our AEAD score displayed more pronounced alterations in AAII and melphalan samples than the other statistics (Fig. 3E,F). While the effects of adduct treatments were evident in both AAII and melphalan samples (Suppl. Table 4), signal-level statistics did not differentiate mitomycin C samples from controls.

2.4 Motif discovery reveals sequences associating with treatment effects

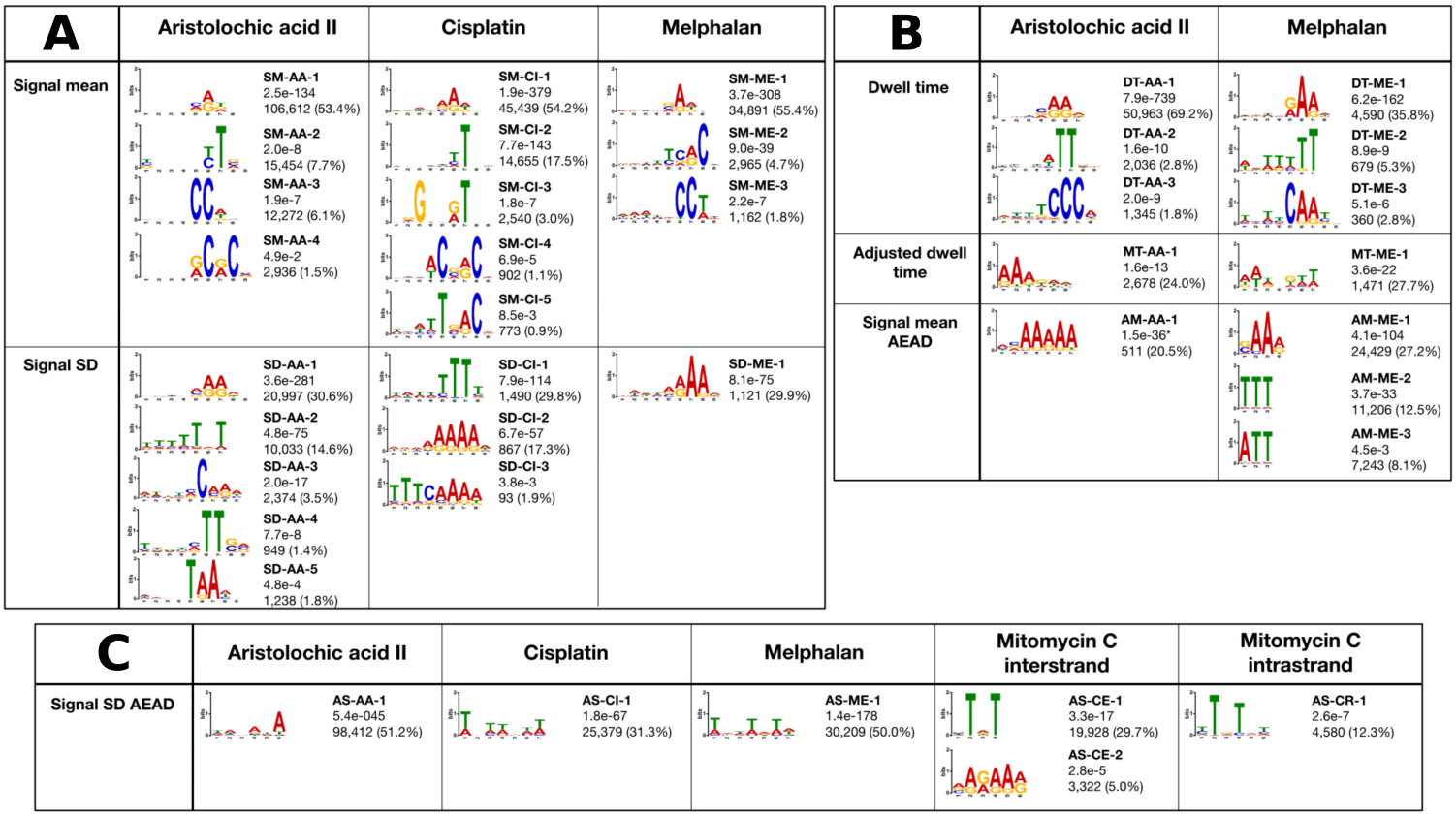


Figure 4: All significant, high-quality motifs found with the STREME analysis. A motif had to occur on both strands and have an E-value smaller than 0.05 to be included. For each motif, its identifier (top), E-value (middle), and the number and fraction of motif matches in the significant 9-mers (bottom) are shown. Significant motifs shown for **A**: signal mean and standard deviation, **B**: dwell times and signal mean AEAD, and **C**: signal standard deviation AEAD. In **B**, * refers to a reduced k-mer set (Methods).

Next, we asked whether treatment-driven effects might be linked to recurring DNA sequence contexts. To explore this, we conducted sequence motif discovery on k-mers associated with differences in sequencing statistics between the treated and control samples, as determined by the CvM criterion (Methods, Suppl. Fig. 4). This analysis revealed multiple significant motifs ($E < 0.05$), some specific to individual treatments and others shared across treatments (Fig. 4). No significant motifs were linked to base quality scores in any of the treatments.

We identified motifs in AAII, cisplatin, and melphalan samples to associate with the signal mean level and variability (Fig. 4A). The most prominent motifs (SM-AA-1, SM-CI-1, SM-ME-1; Fig. 4) all showed purine preference at upper pore positions (*i.e.*, closest to the motor protein; Fig. 1A).

We observed pyrimidine-rich motifs at upper pore positions in AAII and cisplatin samples (SM-AA-2, SM-CI-2). Additionally, weaker motifs containing pairs of cytosines, either directly adjacent or separated by one to two nucleotides, were also identified (SM-AA-3, SM-AA-4, SM-CI-4, SM-ME-3). The discovery of pyrimidine-rich motifs was unexpected, since the treatments have been reported to primarily target purines [5, 24, 25, 26]. However, these findings

align with a previous report [23] that the ionic current is strongly influenced by the nucleotide in the seventh pore position in the nanopore we used in our experiments (ONT R10.4.1). For signal variability, motifs of repeated purines showed a strong effect in all treatments (Fig. 4A; SD-AA-1, SD-CI-2, SD-CI-3, SD-ME-1), while cisplatin also yielded a thymine repeat (SD-CI-1), with similar patterns seen in AAII (SD-AA-2, SD-AA-4).

Strong motifs associating with dwell times were observed in AAII and melphalan (DT-AA-1, DT-ME-1; Fig. 4B). These purine-rich motifs resembled those identified in signal SD analyses (SD-AA-1, SD-ME-1). Motifs linked to dwell time differences at the motor protein (ADT) revealed an adenine-rich motif in AAII (MT-AA-1) and a less specific motif in melphalan (MT-ME-1). Such motifs may indicate sequence contexts where adducts influence translocation speed by interacting with the motor protein rather than the sensing region. In general, dwell time motifs contained homopolymeric tracts, suggesting that repetitive sequences are particularly sensitive to adduct-induced perturbations.

AEAD analysis, which sensitively captures effects on longer sequence contexts (Methods), highlighted the role of purine-rich motifs. For the AAII samples, to focus on the most variable set of 9-mers, we used the 9-mers within the highest 1% of CvM test scores. In AAII and melphalan, the signal mean AEAD varied the most at A-rich motifs (AM-AA-1, AM-ME-1). In the melphalan samples, we observed also thymine-rich motifs (AM-ME-2, AM-ME-3). These were the shortest and least position-specific motifs discovered.

Finally, we found gapped sequence motifs associated with signal SD AEAD in every treatment group. In AAII, motif AS-AA-1 is an A-rich motif, while in cisplatin and melphalan, the motifs contain adenines and thymines (AS-CI-1, AS-ME-1). For mitomycin C, we found T-rich motifs (AS-CE-1, AS-CR-1), and for the interstrand mitomycin C sample, an additional purine-rich motif, AS-CE-2. Signal SD AEAD was the only variable associated with significant motifs across all treatments.

2.5 Recurrent sequence motifs display characteristic effects on nanopore signal levels

Quantile analyses revealed that adduct-induced effects were concentrated at the extremes of k-mer distributions, but their directionality and recurrent sequence patterns remained unclear. We thus examined sequence motifs associated with extreme quantiles in more detail (Methods).

Signal levels in both untreated and mitomycin-treated samples correlated closely with controls (control-control $r = 0.997$, MMC interstrand $r = 0.999$, and MMC intrastrand $r = 0.999$; Fig. 5A, Suppl. Fig. 5), indicating the absence of strong effects in these conditions. In contrast, in both AAII and melphalan samples, nearly all k-mers displayed reduced signal levels compared to untreated controls. Moreover, motifs discovered earlier were consistently associated with lower signal levels in AAII, but the association was not as strong in melphalan samples (Fig. 5A). Examining the k-mers which yielded the largest reductions revealed patterns (Fig. 5B) compatible with the previously discovered signal mean and SD motifs (Fig. 4A). In melphalan samples, we observed that motifs associated with reduced signal levels contained more guanines at the seventh pore position and more purines at the preceding positions, which was also observed in AAII. Outliers at intermediate signal levels contained more cytosines at the seventh position and

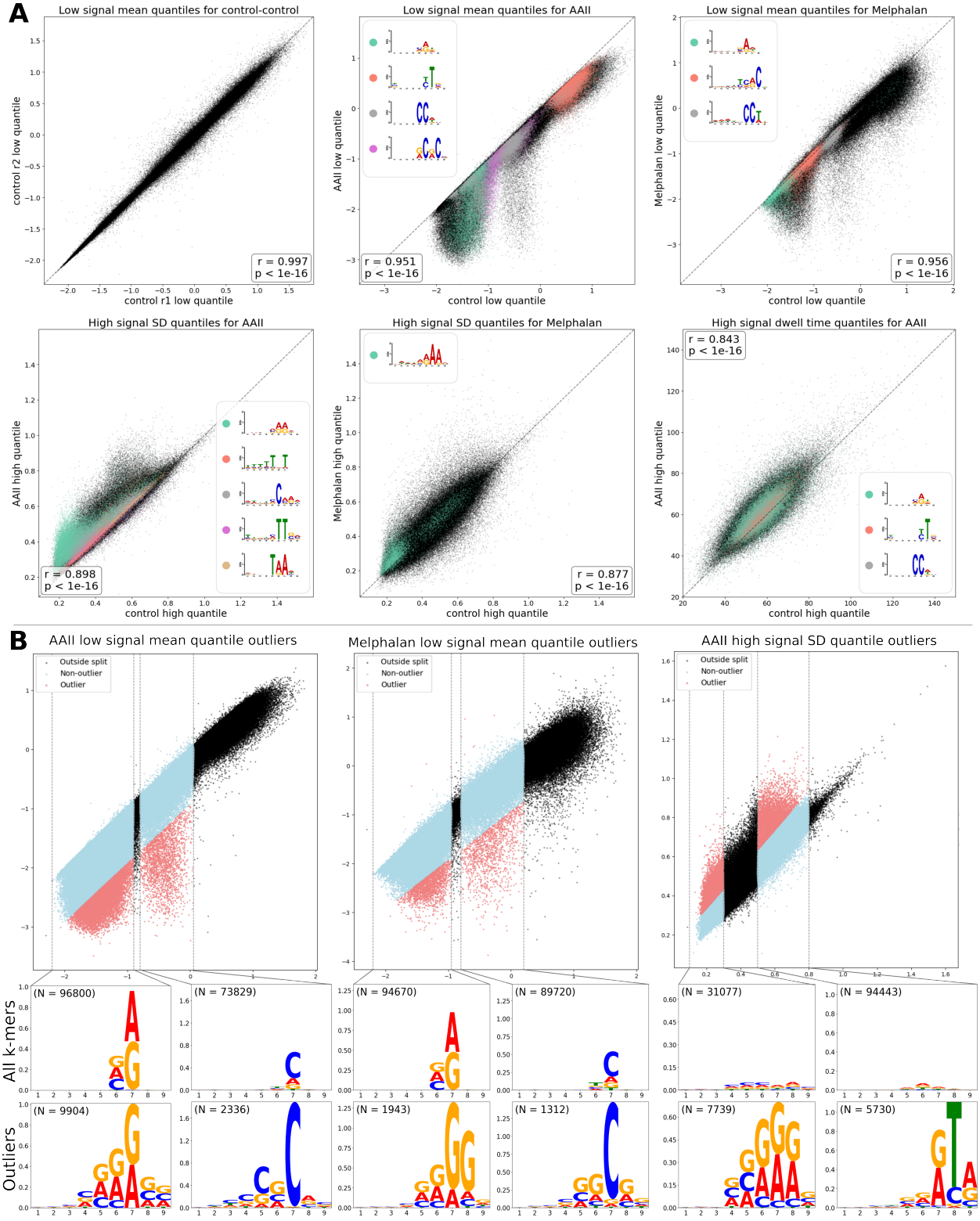


Figure 5: Quantile-based analyses of signal statistics and their association with sequence motifs. **A:** Scatterplots of low quantile signal mean values for all k-mers comparing two control replicates, melphalan, and AAII, and high quantile comparisons for AAII and melphalan signal SD and AAII dwell times. **B:** Sequence logos of outlier values in quantile distributions. The top logo was computed from all k-mers within the range, and the bottom logo was computed with only outlier values, shown in red.

showed increased frequencies of purines or cytosines in the preceding positions in both the AAII- and melphalan-treated samples. Finally, for AAII, SD outliers at low signal values were enriched in purines at pore positions 6–8, and the high signal SD outliers had a [A/G][T/C][A/G] motif at the end of the pore. In contrast, in AAII, the dwell time motifs did not consistently associate with quantile differences, suggesting that this treatment has no consistent effect on nanopore sequencing translocation speed.

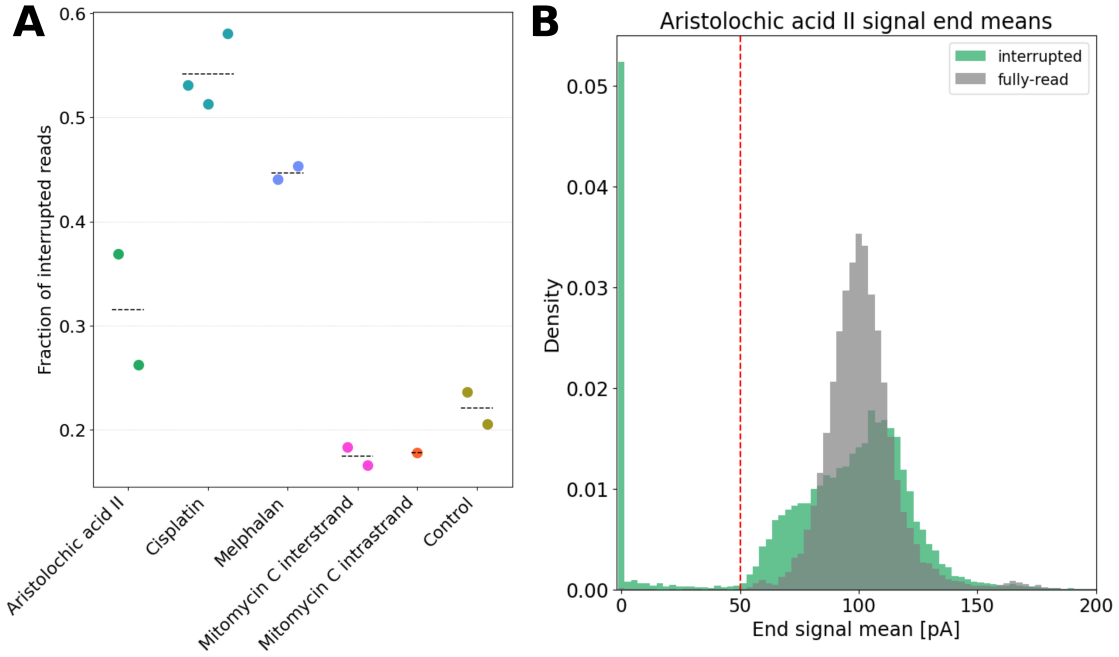
We also tested how different DNA triplets affect extreme values of the sequencing statistics at specific nanopore positions (Suppl. Fig. 6, 7). We found that purine-rich triplets were associated with lower signal levels in AAII and melphalan samples. More specifically, a CAG-triplet in AAII and a GGG-triplet in melphalan had the highest effects. In AAII samples, purine-rich triplets were associated with high signal variation, with AGG and GGG showing the most variability.

In summary, motif discovery revealed sequence logos that were significantly associated with changes in the ionic current. These motifs had purine bases, primarily adenines, at the upper pore positions, in all samples and for all signal statistics, except for SD in cisplatin. In addition, we observed pyrimidine-rich motifs that should not be binding sites of any of the adduct-forming compounds [5, 24, 25, 26]. The discovered motifs correlated with low signal mean quantiles in AAII and melphalan treatments and with high standard deviation quantiles in AAII. Signal SD AEAD was the only variable that had significant sequence motifs associated with every treatment, and the motifs had a unique gapped appearance.

2.6 Read interruptions are associated with upstream guanine repeats at the motor protein

Nanopore sequencing occasionally yields incomplete reads due to interruptions in pore translocation [27]. By examining rear barcode qualities (Methods), we observed that AAII, cisplatin, and melphalan samples produced more reads lacking a rear barcode than control and MMC samples, indicating a higher frequency of interrupted reads (Fig. 6A, Suppl. Fig. 1). To investigate whether these interruptions are associated with particular sequence contexts, we performed motif discovery analysis on interrupted reads. We identified two types of interrupted reads based on the signal level at the read end: passing-like reads where the level was similar to fully-read reads, and blocking reads where the level was lower (Fig. 6B).

We discovered guanine-rich motifs in AAII and melphalan samples associating with read interruptions ($E < 0.05$; Fig. 6C). Passing-like reads in melphalan sample highlighted a simple guanine motif, whereas two blocking read motifs in AAII and melphalan samples had more varied T(G)n and GGC compositions, respectively. Interestingly, these three motifs were all enriched at the motor protein instead of the sensing region (Fig. 6D). This led us to believe that both AAII and melphalan monoadducts may cause interruptions in nanopore sequencing. The GGC-motif in melphalan blocking reads could suggest a binding position for an interstrand crosslink, which would interrupt sequencing by interacting with the motor protein.



C

Sample	Motif (forward strand)	E-value N sites (%)
AAII, blocking read motor protein sequence		1.1e-15 22,726 (40.2%)
Melphalan, blocking read motor protein sequence		3.4e-7 6,492 (46.6%)
Melphalan, passing-like read motor protein sequence		3.6e-18 30,984 (62.2%)

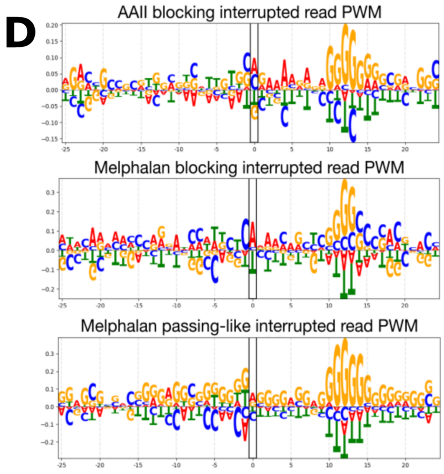


Figure 6: **A:** Fraction of interrupted reads in each sample replicate. **B:** Mean signal levels at read ends for interrupted and fully-read reads in the AAII sample. The red line shows the classification threshold for passing-like and blocking reads. **C:** Sequence motifs detected with STREME associating with read interruptions. **D:** Position weight matrices (PWMs) for the sequences corresponding to motifs centered at the sensing region. The motor protein region is located approximately at position +14 from the sensing region.

3 Discussion

In this study, we investigated how genotoxic, adduct-forming compounds influence nanopore sequencing readouts. To our knowledge, this was the first study to thoroughly examine the effect of DNA adducts in complex DNA contexts (99.96% of all 9-mers), whereas previous studies have analyzed only up to four specific adduct sites [19, 20]. By applying our novel IonStats framework, we profiled adduct-induced changes across multiple dimensions of nanopore data, including base quality, nanopore signal statistics, pore translocation (dwell) times, and read interruptions. This effort revealed a multitude of compound-specific alterations in signal properties and sequence contexts, highlighting

both shared and distinct patterns of how different adducts perturb DNA translocation through the pore. Our observations establish a foundation for linking nanopore-based adduct detection to exposure-associated mutational processes and signatures [28, 3].

Detecting DNA adducts is a crucial step toward understanding how mutational signatures arise. For instance, signature SBS4 found in lung cancers has been linked to tobacco carcinogens [29]. Site-directed mutagenesis studies demonstrate that the mutational signatures of DNA adducts are directly related to DNA polymerase activity, its DNA insertion and extension mechanism, and the DNA sequence context [30]. To properly assign an etiology to a mutational signature, it is essential to consider DNA adduct types and locations as the key intermediate steps that yield somatic mutations. This, however, requires accurate, site-specific methods for adduct detection: we here showed that nanopore sequencing coupled with sophisticated computational tools for ionic current analysis shows substantial promise.

Our software package IonStats can be easily applied to sets of treated and control samples to evaluate sequencing statistics for treatment effects. In contrast to previous DNA and RNA modification detection tools that can be used to characterize DNA adducts [31, 19, 32], IonStats was developed specifically for DNA adduct detection. It introduces several new readouts on signal-level statistics, and translocation dynamics including dwell-times and read interruptions. These readouts enable multifaceted characterization of effects caused by DNA adducts. We hope that this tool will encourage further research on profiling effects of environmental exposures with nanopore sequencing beyond metagenomics [33].

Our treatment-specific findings illustrate both potential and challenges. Aristolochic acid II (AAII) altered sequencing quality and ionic current signals mainly in purine-rich contexts at the nanopore sensing region. In addition, guanine repeats were enriched upstream of the sensing region at the motor protein, associating with translocation interruptions. This finding led us to believe that aristolactam II-dA (AL-dA) adducts are more likely to pass through the nanopore, while AL-dG adducts can cause read interruptions by interacting with the motor protein. This is consistent with the larger size of AL-dG adducts compared to AL-dA adducts [5]. Melphalan produced shorter and poorer quality reads than the other treatments, and displayed widespread effects in adenine-rich contexts and evidence of read interruptions in CG-rich contexts. This could indicate melphalan crosslinks, binding to two guanines on opposing strands [34], to cause translocation failure at the motor protein, similarly to AAII.

We observed cisplatin-driven differences in signal statistics primarily in A and T-rich motifs. Although GG-crosslinks are the primary reported cisplatin adduct, our motif discovery analysis was not able to detect these. This may be due to insufficient power resulting from the low sequencing coverage, caused by degradation of the DNA by the cisplatin treatment. We were surprised to find many T-rich motifs in cisplatin and AAII (*e.g.*, SM-AA-2, SM-CI-2, SD-AA-2, SD-AA-4, SD-CI-1), since cisplatin has not been reported to bind to thymines, although GpNpG 1,3-intrastrand crosslink can rarely occur (SM-CI-3) [35]. Furthermore, AL-dT adducts have not been observed in previous studies [36]. The only effects we discovered after MMC treatments were the motifs associated with signal variability over longer sequences captured by the AEAD score, suggesting that AEAD is more sensitive to treatment effects than other statistics studied.

While IonStats enabled comprehensive profiling of treatment effects, it is not able to identify individual reads harboring adducts. Thus, its sensitivity might be limited when adduct frequencies are low. In addition, IonStats has only been applied to *in vitro* samples that are expected to carry relatively large amounts of DNA adducts compared to *in vivo* samples [3]. Further work is required to evaluate the utility of IonStats in *in vivo* settings. Another limitation of our work is that IonStats is not able to quantify the amount of adducts in treated samples directly. A complementary mass spectrometry analysis could be used to estimate the fraction of adducted nucleotides.

Taken together, our work establishes a framework for studying DNA adducts with nanopore sequencing and demonstrates how different compounds leave distinct signatures in sequencing signals. By developing methods for site-specific adduct detection, future studies may bridge the gap between DNA lesions and mutational signatures, clarifying the causal pathways by which exogenous and endogenous exposures contribute to cancer development.

References

[1] Byeong Hwa Yun, Jingshu Guo, Medjda Bellamri, and Robert J. Turesky. DNA adducts: Formation, biological effects, and new biospecimens for mass spectrometric measurements in humans. *Mass Spectrometry Reviews*, 39(1-2):55–82, 2020.

[2] Aziz Sancar, Laura A. Lindsey-Boltz, Keziban Ünsal-Kaçmaz, and Stuart Linn. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual Review of Biochemistry*, 73:39–85, 2004.

[3] Gunnar Boysen, Ludmil B. Alexandrov, Raheleh Rahbari, Intawat Nookaew, Dave Ussery, Mu Rong Chao, Chiung Wen Hu, and Marcus S. Cooke. Investigating the origins of the mutational signatures in cancer. *Nucleic Acids Research*, 53(1), 2025.

[4] Thomas A. Rosenquist and Arthur P. Grollman. Mutational signature of aristolochic acid: Clue to the recognition of a global disease. *DNA Repair*, 44:205–211, 2016.

[5] Marie Stiborová, Volker M. Arlt, and Heinz H. Schmeiser. DNA adducts formed by aristolochic acid are unique biomarkers of exposure and explain the initiation phase of upper urothelial cancer. *International Journal of Molecular Sciences*, 18(10), 2017.

[6] Michael Alavanja, John A. Baron, Ross C. Brownson, Patricia A. Buffler, David M. DeMarini, Mirjana V. Djordjevic, Richard Doll, Elizabeth T.H. Fontham, Yu Tang Gao, Nigel Gray, Prakash C. Gupta, Allan Hackshaw, Stephen S. Hecht, Kirsti Husgafvel-Pursiainen, Elena Matos, Richard Peto, David H. Phillips, Jonathan M. Samet, Gary Stoner, Michael J. Thun, Jean Trédaniel, Paolo Vineis, H. Erich Wichmann, Anna H. Wu, and David Zaridze. Tobacco smoke and involuntary smoking. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, 83:1–1413, 2004.

[7] Thomas W. Kensler, Bill D. Roebuck, Gerald N. Wogan, and John D. Groopman. Aflatoxin: A 50-year Odyssey of mechanistic and translational toxicology. *Toxicological Sciences*, 120(SUPPL.1):28–48, 2011.

[8] M. Austin Argentieri, Najaf Amin, Alejo J. Nevado-Holgado, William Sproviero, Jennifer A. Collister, Sarai M. Keestra, Midas M. Kuilman, Bigina N.R. Ginos, Mohsen Ghanbari, Aiden Doherty, David J. Hunter, Alexandra Alvergne, and Cornelia M. van Duijn. Integrating the environmental and genetic architectures of aging and mortality. *Nature Medicine*, 31(3):1016–1025, 2025.

[9] Mengqiu Cao and Xinyu Zhang. DNA Adductomics: A Narrative Review of Its Development, Applications, and Future. *Biomolecules*, 14(9), 2024.

[10] David K. La and James A. Swenberg. DNA adducts: Biological markers of exposure and potential applications to risk assessment. *Mutation Research - Reviews in Genetic Toxicology*, 365(1-3):129–146, 1996.

[11] Daniel R. Semlow and Johannes C. Walter. Mechanisms of Vertebrate DNA Interstrand Cross-Link Repair. *Annual Review of Biochemistry*, 90:107–135, 2021.

[12] Natalia Y. Tretyakova, Arnold Groehler, and Shaofei Ji. DNA-Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes. *Accounts of Chemical Research*, 48(6):1631–1644, 2015.

[13] Romel P. Dator, Kevin J. Murray, Matthew W. Luedtke, Foster C. Jacobs, Fekadu Kassie, Hai Dang Nguyen, Peter W. Villalta, and Silvia Balbo. Identification of Formaldehyde-Induced DNA-RNA Cross-Links in the A/J Mouse Lung Tumorigenesis Model. *Chemical Research in Toxicology*, 35(11):2025–2036, 2022.

[14] Gunnar Boysen and Intawat Nookaew. Current and Future Methodology for Quantitation and Site-Specific Mapping the Location of DNA Adducts. *Toxics*, 10(2), feb 2022.

[15] Lieselot Y. Hemeryck, Anneleen I. Decloedt, Julie Vanden Bussche, Karen P. Geboes, and Lynn Vanhaecke. High resolution mass spectrometry based profiling of diet-related deoxyribonucleic acid adducts. *Analytica Chimica Acta*, 892:123–131, 2015.

[16] Yuan Jhe Chang, Marcus S. Cooke, Chiung Wen Hu, and Mu Rong Chao. Novel approach to integrated DNA adductomics for the assessment of in vitro and in vivo environmental exposures. *Archives of Toxicology*, 92(8):2665–2680, 2018.

[17] Mikhail F Denissenko, Sundaresan Venkatachalam, Yu Hua Ma, and Altaf A Wani. Site-specific induction and repair of benzo[a]pyrene diol epoxide DNA damage in human H-ras protooncogene as revealed by restriction cleavage inhibition. *Mutation Research - DNA Repair*, 363(1):27–42, 1996.

[18] Bo Cao, Xiaolin Wu, Jieliang Zhou, Hang Wu, Lili Liu, Qinghua Zhang, Michael S Demott, Chen Gu, Lianrong Wang, Delin You, and Peter C Dedon. Nick-seq for single-nucleotide resolution genomic maps of DNA modifications and damage. *Nucleic Acids Research*, 48(12):6715–6725, 2020.

[19] Intawat Nookaew, Gunnar Boysen, Piroon Jenjaroenpun, Hua Du, Pengcheng Wang, Jun Wu, Thidathip Wongsurawat, Sun Hee Moon, En Huang, and Yinsheng Wang. Detection and discrimination of DNA adducts differing in size, regiochemistry, and functional group by nanopore sequencing. *Chemical Research in Toxicology*, 33(12):2944–2952, dec 2020.

[20] Xinjia Zhao, Yuru Liu, Xiaoyu Chen, Zhuang Mi, Wei Li, Pengye Wang, Xinyan Shan, and Xinghua Lu. Detection and Characterization of Single Cisplatin Adducts on DNA by Nanopore Sequencing. *ACS Omega*, 6(26):17027–17034, jul 2021.

[21] Yunhao Wang, Yue Zhao, Audrey Bollas, Yuru Wang, and Kin Fai Au. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, 39(11):1348–1365, nov 2021.

[22] Miten Jain, Hugh E. Olsen, Benedict Paten, and Mark Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1), dec 2016.

[23] Sam Kovaka, Paul W. Hook, Katharine M. Jenike, Vikram Shivakumar, Luke B. Morina, Roham Razaghi, Winston Timp, and Michael C. Schatz. Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment. *Nature Methods*, 22(4):681–691, 2025.

[24] Shaloam Dasari and Paul Bernard. Cisplatin in cancer therapy : Molecular mechanisms of action. *European Journal of Pharmacology*, 740:364–378, 2014.

[25] M R Osborne and P. D. Lawley. Alkylation of DNA by melphalan with special reference to adenine derivatives and adenine-guanine cross-linking. *Chemico-Biological Interactions*, 89(1):49–60, 1993.

[26] Maria Tomasz and Yolanda Palom. The mitomycin bioreductive antitumor agents: Cross-linking and alkylation of DNA as the molecular basis of their activity, 1997.

[27] D. Branton and D.W. Deamer. *Nanopore Sequencing: An Introduction*. World Scientific Publishing Company, 2019.

[28] Lauri J. Sipilä, Riku Katainen, Mervi Aavikko, Janne Ravantti, Iikki Donner, Rainer Lehtonen, Ilmo Leivo, Henrik Wolff, Reetta Holmila, Kirsti Husgafvel-Pursiainen, and Lauri A. Aaltonen. Genome-wide somatic mutation analysis of sinonasal adenocarcinoma with and without wood dust exposure. *Genes and Environment*, 46(1):1–13, 2024.

[29] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A.J.R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinsk, Natalie Jäger, David T.W. Jones, David Jonas, Stian Knappskog, Marcel Koo, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N.J. Tutt, Rafael Valdés-Mas, Marit M. Van Buuren, Laura Van 'T Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M.

Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

[30] Robert P. Fuchs and Shingo Fujii. Translesion DNA synthesis and mutagenesis in prokaryotes. *Cold Spring Harbor Perspectives in Biology*, 5(12):1–22, 2013.

[31] Piroon Jenjaroenpun, Thidathip Wongsurawat, Taylor D Wadley, Trudy M Wassenaar, Jun Liu, Qing Dai, Visanu Wanchai, Nisreen S Akel, Azemat Jamshidi-Parsian, Aime T Franco, Gunnar Boysen, Michael L Jennings, David W Ussery, Chuan He, and Intawat Nookaew. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Research*, 49(2):1–13, 2021.

[32] Jared T. Simpson, Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, 2017.

[33] Mantas Sereika, Rasmus Hansen Kirkegaard, Søren Michael Karst, Thomas Yssing Michaelsen, Emil Aarre Sørensen, Rasmus Dam Wollenberg, and Mads Albertsen. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. *Nature Methods*, 19(7):823–826, jul 2022.

[34] Lawrence F Povirk and David E Shuker. DNA damage and mutagenesis induced by nitrogen mustards, 1994.

[35] Askar Yimit, Ogun Adebali, Aziz Sancar, and Yuchao Jiang. Differential damage and repair of DNA-adducts induced by anti-cancer drug cisplatin across mouse organs. *Nature Communications*, 10(1), 2019.

[36] Wan Chan, Yufang Zheng, and Zongwei Cai. Liquid Chromatography-Tandem Mass Spectrometry Analysis of the DNA Adducts of Aristolochic Acids. *Journal of the American Society for Mass Spectrometry*, 18(4):642–650, 2007.

[37] Volker M Arlt, Heinz H Schmeiser, and Gerd P Pfeifer. Sequence-specific detection of aristolochic acid-DNA adducts in the human p53 gene by terminal transferase-dependent PCR The plant extract aristolochic acid (AA), a mixture consisting. Technical Report 1, 2001.

[38] Fubo Ma, Shuanghong Yan, Jinyue Zhang, Yu Wang, Liying Wang, Yuqin Wang, Shanyu Zhang, Xiaoyu Du, Panke Zhang, Hong-Yuan Chen, and Shuo Huang. Nanopore Sequencing Accurately Identifies the Cisplatin Adduct on DNA. *ACS Sensors*, 2021.

[39] Dalia Mohamed and Michael Linscheid. Separation and identification of trinucleotide-melphalan adducts from enzymatically digested DNA using HPLC-ESI-MS. In *Analytical and Bioanalytical Chemistry*, volume 392, pages 805–817, nov 2008.

[40] Manuel M. Paz, Sweta Ladwa, Elise Champeil, Yanfeng Liu, Sara Rockwell, Ernest K. Boamah, Jill Bargonetti, John Callahan, John Roach, and Maria Tomasz. Mapping DNA adducts of mitomycin C and decarbamoyl mitomycin C in cell lines using liquid chromatography/electrospray tandem mass spectrometry. *Chemical Research in Toxicology*, 21(12):2370–2378, dec 2008.

[41] Heng Li. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.

[42] Jurgen F. Nijkamp, Marcel van den Broek, Erwin Datema, Stefan de Kok, Lizanne Bosman, Marijke A. Luttik, Pascale Daran-Lapujade, Wanwipa Vongsangnak, Jens Nielsen, Wilbert H.M. Heijne, Paul Klaassen, Chris J. Paddon, Darren Platt, Peter Kötter, Roeland C. van Ham, Marcel J.T. Reinders, Jack T. Pronk, Dick de Ridder, and Jean Marc Daran. De novo sequencing, assembly and analysis of the genome of the laboratory strain Saccharomyces cerevisiae CEN.PK113-7D, a model for modern industrial biotechnology. *Microbial Cell Factories*, 11(1):36, 2012.

[43] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

[44] Timothy L. Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, 37(18):2834–2840, 2021.

[45] Ammar Tareen and Justin B Kinney. Logomaker: Beautiful sequence logos in Python. *Bioinformatics*, 36(7):2272–2274, 2020.

[46] Fabian A. Buske, Mikael Bodén, Denis C. Bauer, and Timothy L. Bailey. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, 26(7):860–866, 2010.

Acknowledgments

We would like to thank Veera Erkkilä, Pinja Perkkiö, Iina Vuoristo, and Inga-Lill Åberg for technical support, and Harri Kangas for discussions. We acknowledge CSC – IT Center for Science, Finland, for generous computational resources. Figure 1D was created with BioRender.com. This study was supported by the Research Council of Finland (#328890 to EP) and the University of Helsinki Doctoral Programme in Integrative Life Science.

Data availability statement

All sequencing data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB101972.

4 Methods

4.1 Whole genome DNA amplification

Yeast genomic DNA was extracted from the haploid *Saccharomyces cerevisiae* strain CEN.PK113-7D by using the MasterPure Yeast DNA Purification Kit (MPY80200, LGC). High molecular weight yeast genomic DNA was uniformly amplified by using the REPLI-G mini whole genome amplification kit (150023, Qiagen). In brief, 20 ng of yeast genomic DNA was used as a template and amplified by incubating with REPLI-G DNA Polymerase at 30 °C for 16 hours. Amplified yeast genomic DNA was purified by using 1.8x Ampure XP beads (10136224, Beckman Coulter) and eluted in 10 mM Tris-HCl, pH 8.0, or double-distilled water. Amplified yeast genomic DNA was fragmented and size-selected for 10-15 Kb by using g-tube (520079, Covaris). For Melphalan treatment, DNA was extracted for ultrapure distilled water (10977035, Invitrogen).

4.2 Aristolochic acid II treatment

AAII DNA adducts were formed by using the protocol described in [37]. Amplified yeast genomic DNA (10 µg DNA, one µg per µl) was treated with 0.25 mM AAII solution in 50 µl of 50 mM potassium phosphate buffer, pH 5.8. AAII was activated by 1 mg of zinc dust and incubated at 37 °C for 30 minutes. AAII-treated DNA was purified by using 1.8x Ampure XP beads and eluted in 25 µl 10 mM Tris-HCl, pH 8.0.

4.3 Cisplatin treatment

Cisplatin DNA-adduct was prepared by using the protocol in [38] with modifications. Cisplatin (3 mg, 10 µmol) was activated by incubation with $AgNO_3$ (6 mg, 35 µmol) at 37 °C for 6 hours in 50 µl of water. The activation reaction was conducted in a 1.5 ml Eppendorf tube covered with an aluminium foil to prevent light exposure. Cisplatin and $AgNO_3$ mixture was centrifuged at 13,000 rpm for 10 minutes, and the clear supernatant containing activated cisplatin was used for DNA-adduct formation. Cisplatin-DNA adduct was formed by incubating DNA (10 µg) and activated

cisplatin in 10 mM HEPES buffer, pH 7.0, at 37 °C for 1 hour in the dark. Cisplatin-treated DNA was purified by using Phenol-Chloroform extraction and eluted in 10 mM Tris, pH 8.0.

4.4 Melphalan treatment

Melphalan DNA-adduct was prepared by using the protocol in [39] with a modification. Amplified yeast genomic DNA was eluted in distilled water (10 µg in 10 µl water). Amplified yeast genomic DNA was incubated with 8 µl of melphalan stock solution (1 mg per ml of DMSO) and 12 µl of DMSO at 37 °C for 1 hour. Melphalan-treated DNA was purified by using Phenol-Chloroform extraction and eluted in distilled water.

4.5 Mitomycin C treatment

4.5.1 Intrastrand crosslink adduct preparation:

Mitomycin C (MMC) intrastrand crosslink DNA-adduct was prepared by using the protocol in [40]. Amplified yeast genomic DNA was diluted to 30 µl of 0.1 M PBS pH 7.4 (P5493, Sigma) (6.54 µg DNA in 30 µl PBS, 60 nmol of DNA). MMC was dissolved in 0.1 M PBS, pH 7.4, to prepare a 100 mM stock solution. 0.24 µmol of mitomycin C (2.4 µl of 0.1 M MMC stock solution was diluted to 27 µl in 0.1 M PBS) and 12 µmol of $Na_2S_20_4$ (3 µl of 40 mM $Na_2S_20_4$ in water) were added to amplified yeast genomic DNA and incubated in a PCR tube in a thermoshaker at 40 °C, 400 rpm with open lid in fume hood. MMC-treated DNA was purified by using 1.8x Ampure XP beads and eluted in 20 µl 10 mM Tris-HCl, pH 8.0.

4.5.2 Interstrand crosslink adduct preparation:

MMC interstrand crosslink DNA adduct was prepared using a protocol similar to intrastrand crosslinking, except the incubation was performed in a closed lid PCR tube, and a total of 60 µmol of $Na_2S_20_4$ was added in five portions of 12 µmol at a five-minute interval. MMC-treated DNA was purified by using 1.8x Ampure XP beads and eluted in 20 µl 10 mM Tris-HCl, pH 8.0.

4.6 DNA sequencing and data collection

Two replicates for each treatment were prepared for sequencing, with an additional cisplatin sample, where the DNA fragments were selected by size. Approximately 400 ng of DNA was used for each sample to prepare the sequencing library with the Native Barcoding Kit 24 V14 (SQK-NBD114.24, Oxford Nanopore Technologies (ONT)) following the manufacturer's protocol. In total, 13 barcodes were used (two for each treatment, two for controls, and one additional for cisplatin size-selected). The constructed libraries were loaded onto the R10.4.1/FLO-PRO114M flow cell on the PromethION 24 device. The sequencing and data collection were performed with the MinKNOW software v24.02.19, which generated .pod5 files containing the raw ionic signal measurements for individual DNA molecules.

The cisplatin treatment heavily damaged the DNA, and we observed a small number of reads for both non-size-selected replicates and the size-selected replicate. For unknown reasons, the yield for the second MMC intrastrand replicate was very low, and we excluded this sample from further analyses.

4.7 Data processing

Basecalling and alignment were performed with ONT's basecaller Dorado (version 0.6.0) with the dna_r10.4.1_e8.2_400bps_sup@v4.3.0 model. The aligner function in Dorado uses minimap2 [41] to align the basecalled reads to a reference genome with the "lr:hq" preset. We used the *S. cerevisiae* strain CEN.PK113-7D assembly from [42] as the reference genome. To map the raw ionic current measurements to positions in the reference genome, a signal-alignment step is required. The ionic current is divided into events, and each event is mapped to a reference position. We applied Uncalled4 [23] to align signals to reference positions for all reads. This allowed us to compute and compare ionic current statistics for different samples.

4.8 Data analysis of sequencing statistics with IonStats

IonStats can extract and compare quality scores, ionic current means, ionic current standard deviations, dwell times, and adjusted dwell times between treated and control samples. We also derived windowed error scores for ionic current mean and standard deviation, referred to here as Average Expected Absolute Deviation (AEAD) scores.

The quality score (Q) is a value given to each base by the basecaller, and it is logarithmically related to the estimated basecalling error probability (P): $Q = -10 \log_{10} P$. The ionic current mean and standard deviation are computed from the raw ionic current measurements that belong to a single event, defined by the signal alignment. The dwell time refers to the time that a k-mer resides inside the nanopore, and we quantified it as the number of measurements that belong to an event. To capture the interaction between putative adducts and nanopore motor protein, we computed the motor-protein shifted dwell times (adjusted dwell times), where the dwell time of a k-mer was associated with a k-mer 14 base pairs downstream in the reference genome [20].

To estimate how much the ionic current varies over consecutive events, we propose the AEAD statistic. For a window of width w, the AEAD for statistic s at position i can be specified as:

$$AEAD_{s,i} = \frac{1}{w} \sum_{j=i-\lfloor \frac{w}{2} \rfloor}^{i+\lfloor \frac{w}{2} \rfloor} |s_j - E[s, k_j]|$$

where s_j is the value of the statistic for event j and $E[s, k_j]$ is the expected value of the statistic s for k-mer centered at position j. We used the 9-mer table provided by Uncalled4 to get the expected values for the signal mean and standard deviation for all 9-mers.

IonStats collects values for selected variables and groups them by the reference k-mer or by reference position for a specific chromosome or contig. It then performs distributional tests on the k-mer/ref_pos distributions to find the

k-mers that show statistically significant differences in the variables of interest. It applies the Cramér–von Mises (CvM) criterion and the Kolmogorov–Smirnov (KS) test to compare the k-mers. For both of these tests, we used the versions that were implemented in the SciPy [43] Python library. These two tests are applied to capture different effects since KS is more sensitive to differences in medians, while CvM is more sensitive to differences in distribution tails. Since both tests assume the data to be continuous, we added a small amount of noise to the discrete variables (quality score, dwell time, and adjusted dwell time) to facilitate test calibration. To control for multiple testing, IonStats applies the Benjamini-Hochberg correction for the p-values of each test with a significance level α. We used $\alpha = 0.05$.

4.9 Motif discovery analysis

To search for patterns that could correspond to recurrent adduct-forming sequence motifs in the DNA, we performed motif discovery analysis with STREME [44] for the significant k-mers associated with each test. In the STREME analyses, we used the set of all k-mers as the control sequences. To process the k-mers only in forward-strand orientation, we used a custom STREME alphabet, which did not consider the complementarity of nucleobases. The motif discovery was performed separately on forward and reverse reads. A motif had to occur on both strands with an E-value smaller than 0.05 to be reported in Fig. 3 and considered significant. All motifs discovered by STREME with $E < 0.05$ are shown in Suppl. Fig. 8-16. We also used the Logomaker [45] Python library to draw sequence logos for the significant k-mers, and for the outlier k-mers in Fig. 5.

For the AAII signal mean AEAD, we restricted the motif discovery analysis only to the k-mers within the top 1% CvM values. This was done to reduce the similarity between the positive set and the control set, which consisted of all k-mers. Initially, by using all significant k-mers, we found no significant motifs, likely due to the high overlap of the positive and control k-mer-sets in the STREME analysis.

4.10 Quantile difference analysis

To study how the extreme values of k-mer distributions compare in treated and control samples, we computed the top and bottom 1%-quantiles for each sequencing variable. We then computed the quantile difference to measure if a k-mer shows more extreme values in the treated samples. The quantile difference for k-mer i is defined simply as: $d_i = q_{\mathrm{t},i} - q_{\mathrm{c},i}$, where c is control and t is treatment, and q can be either the lowest or highest 1%.

4.11 Data analysis for interrupted reads

Nanopore sequencing reads can exhibit low rear barcode qualities, indicating that the read barcode was non-determinable based on the sequence at the end of the read. This led us to believe that interrupted reads, i.e., reads that interrupted the sequencing and did not pass completely through the pore, belong to this subset of reads. We classified reads as interrupted based on the `barcode_rear_score` value, which is recorded for all reads in the sequencing summary file provided by MinKNOW and which ranges between -100 and 100. If `barcode_rear_score` was < 60, we classified

the read as interrupted. We chose this threshold as MinKNOW requires a minimum value of 60 from either the front or rear barcode scores for a read to be assigned to a barcode.

After further investigation, we noticed that some reads classified as interrupted had a low ionic current at the end of the read, while other reads had a similar ionic current level at the end of the read to fully-read reads (Fig. 6A). We divided the interrupted reads into two classes based on their last 20 ionic current measurements: passing-like reads with high end-mean, and blocking reads with low end-mean (< 50 pA). The passing-like reads exhibit a similar signal level at the end of sequencing as fully-read reads, and the blocking reads exhibit a lower signal level. We analyzed the end positions of these reads by extracting 25 bases upstream and downstream (50 in total) of the mapping end position in the reference genome to study whether the reads typically ended at specific sequence motifs, which might be binding sites of adducts. We also performed a motif discovery analysis on subsequences at the mapping end position and on sequences that interact with the motor protein.

4.12 Motif affinity analysis

To estimate how well a k-mer matches a sequence motif, we applied AMA [46] to compute average motif affinity scores. We did this for motifs detected with STREME on all possible k-mers, where $k = 9$, with the `--norc` option to consider motif affinity only in the forward strand direction. In Figure 5A, a 9-mer was colored to match a motif if the AMA score was more than 5. In the case where a 9-mer had high affinity to multiple motifs, we considered only the motif with the highest affinity.

4.13 Analysis of DNA triplets at specific pore positions

We computed the effect of triplet-pore position pairs on quantile differences by iterating all seven triplet positions for a 9-mer by computing the mean quantile difference of all 9-mers that contain a triplet t at position i:

$$mean\_qd_{t,i} = \frac{1}{4^6} \sum_{k \in K[i:i+3]=t} q\_treated_k - q\_control_k$$

where K is the set of all 9-mers.