nature communications

Article

https://doi.org/10.1038/s41467-025-60391-3

Direct detection of 8-oxo-dG using nanopore sequencing

Received: 5 July 2024

Accepted: 21 May 2025

Published online: 05 June 2025

Check for updates

Marc Pagès-Gallego[1,2], Daan M. K. van Soest[1], Nicolle J. M. Besselink[1,2,5], Roy Straver[1,2], Janneke P. Keijer[1], Carlo Vermeulen[1,2], Alessio Marcozzi[3], Markus J. van Roosmalen[2,4], Ruben van Boxtel[2,4], Boudewijn M. T. Burgering[1,2] ✉, Tobias B. Dansen[1] ✉ & Jeroen de Ridder[1,2] ✉

Genomic DNA is under constant oxidative damage, with 8-oxo-7,8-dihydro-2′-deoxyguanosine (8-oxo-dG) being the prominent lesion linked to mutagenesis, epigenetics, and gene regulation. Existing methods to detect 8-oxo-dG rely on indirect approaches, while nanopore sequencing enables direct detection of base modifications. A model for 8-oxo-dG detection is currently missing due to the lack of training data. Here, we develop a strategy using synthetic oligos to generate long, 8-oxo-dG context-variable DNA molecules for deep learning and nanopore sequencing. Our training approach addresses the rarity of 8-oxo-dG relative to guanine, enabling specific detection. Applied to a tissue culture model of oxidative damage, our method reveals uneven genomic 8-oxo-dG distribution, dissimilar context pattern to C>A mutations, and local 5-mC depletion. This dual measurement of 5-mC and 8-oxo-dG at single-molecule resolution uncovers new insights into their interplay. Our approach also provides a general framework for detecting other rare DNA modifications using synthetic DNA and nanopore sequencing.

Genomic DNA is under constant assault from various damaging agents, leading to breaks and chemical modifications such as oxidation. Among the oxidized base adducts that have been identified[1,2], 8-oxo-7,8-dihydro-2′-deoxyguanosine (8-oxo-dG), is the most abundant, since guanine, out of the four bases, has the lowest redox potential[3]. Oxidation of G to 8-oxo-dG can occur both directly in the DNA or in the free nucleotide pool by several processes, including the formation of hydroxyl radicals derived from endogenous reactive oxygen species[4], as well as exogenous sources like ionizing radiation[5], and incorporated into the DNA by several polymerases[6]. Its most pivotal characteristic is its ability to both pair with cytosine (forming a regular Watson-Crick base pair) and adenine (forming a Hoogsteen base pair). 8-oxo-dG, when paired with cytosine, is proactively excised from the DNA in humans by the DNA glycosylases OGG1[7], and adenine, when paired with 8-oxo-dG, is excised by MUTYH[8], but also via preemptive sanitization of the nucleotide pool by MTH1 (also known as NUDT1)[9]. However, upon failure to repair a 8-oxo-dG:C pair prior to replication, it can lead to an 8-oxo-dG:A pair, which upon a second round of replication would become T:A, leading to a C>A transversion[10,11]. If 8-oxo-dG is incorporated from the nucleotide pool opposite of adenine, then a T>G transversion can also occur after replication[12]. Mutations downstream of 8-oxo-dG have significant implications in the development of cancer[13–15]. 8-oxo-dG is the proposed mechanism underlying COSMIC[16] signatures 18 and 36[17–19], and has been recognized as a potential disease biomarker in the field[20,21]. 8-oxo-dG has also been linked to transcriptional and epigenetic regulation[22], in particular in the context of DNA (de)methylation at cytosine (5-methylcytosine, 5-mC): 8-oxo-dG passively inhibits methyltransferases[23], and OGG1 recruits TET enzymes which convert 5-mC to 5-hydroxymethylcytosine (5-hmC) as the first step in the demethylation process[24].

[1]Center for Molecular Medicine, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands. [2]Oncode Institute, Utrecht, The Netherlands. [3]Cyclomics, Utrecht, The Netherlands. [4]Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands. [5]Present address: Princess Máxima Center for Pediatric Oncology, Utrecht, The Netherlands. ✉e-mail: b.m.t.burgering@umcutrecht.nl; t.b.dansen@umcutrecht.nl; j.deridder-4@umcutrecht.nl

Due to its potential as a disease biomarker, several efforts have been made to quantify 8-oxo-dG in urine[25,26], blood[27], genomic material[28], and several other tissues[21]. The absolute quantification of 8-oxo-dG has predominantly relied on highly sensitive methods such as liquid or gas chromatography followed by mass spectrometry or electrochemical detection, or enzyme-linked immunosorbent assays[21]. However, its accuracy has been subject to debate due to large discrepancies between reported levels[21,29–31], and high variability between laboratories[32]. Furthermore, these methods fail to provide insights into the genomic location of 8-oxo-dG, which precludes unveiling the mechanisms underlying heterogeneous mutation rates, repair mechanisms, and the role of 8-oxo-dG in epigenetic regulation. For this reason, recently several innovative genomics-based methods have attempted to investigate 8-oxo-dG in a genome-wide manner. These approaches include the detection of apurinic sites created by 8-oxo-dG repair enzymes (Click-code-seq[33] and OGG1-AP-seq[34]), ChIP-seq techniques employing an 8-oxo-dG antibody (OxiDIP-seq[35]), pull-down of cross-linked biotin tags attached to 8-oxo-dG (OG-seq[36] and CLAPS-seq[37]), and pull-down of 8-oxo-dG repair enzymes (enTRAP-seq[38]). Collectively, these methods have revealed that 8-oxo-dG, and its repair, is not uniformly distributed throughout the genome, although with some contradictory results between used methods[22,39]. Current genomic approaches have three main downsides: first, they lack single-nucleotide resolution (with the exception of Click-code-seq), which hampers the study of the relationship between 8-oxo-dG and its associated mutational signatures; secondly, these methods rely on short-read sequencing methods and therefore cannot properly investigate genomic repetitive regions; and lastly, indirect detection is associated frequently with false positive (FP) signals due to suboptimal antibody specificity, enzymatic or chemical reactivity[39]. The latter is especially important given the reported rarity of 8-oxo-dG (1-100 8-oxo-dG per 1 million G[21]). This means that, even with a usually considered low false positive rate (FPR) of 1%, the ratio between false and true positives would be approximately 10:1, which would quickly obfuscate any real signal and preclude meaningful conclusions[40].

Nanopore sequencing operates by threading a DNA (or RNA) molecule through a membrane-embedded pore while measuring fluctuations in the electrical current over the membrane, and is currently commercialized by Oxford Nanopore Technologies (ONT)[41]. Changes in the electrical current are indicative of the molecule's chemical properties, which holds the potential to detect base modifications. This enabled the detection of 5-mC using α-Hemolysin pores[42], and later using ONT devices[43,44]. Since then, base modification detection models have been developed both by ONT and the community for detecting both naturally occurring and synthetic modifications[45]. But so far, there is no model available for the detection of 8-oxo-dG by nanopore sequencing. Developing modification detection models presents substantial challenges, especially for rare modifications. Firstly, obtaining sequencing data where the precise location of a rare modification is known is not trivial, as these modifications cannot be verified with other technologies to establish a ground truth. Moreover, the context of the modification must exhibit sufficient diversity to prevent the model from learning sequence biases. The effect of the modification passing through the pore on the electrical current must be pronounced enough to distinguish it from inherent noise and other bases. And finally, the model must be extremely precise to accommodate detection of rare modifications, e.g. while for 5-mC it is acceptable to have a 1% FPR, it would be prohibitively high for rare modifications such as 8-oxo-dG due to its low abundance[40].

Keeping these considerations in mind, we devised an approach to generate a library of long synthetic DNA molecules that each contain 8-oxo-dG in a known, specific, but variable sequence context. Utilizing this ground truth dataset, we demonstrate that 8-oxo-dG has a discernible impact on the nanopore raw signal, leading to systematic errors using the ONT provided basecaller. We trained a deep-learning model capable of detecting 8-oxo-dG from the raw signal with single-nucleotide resolution, high specificity, and employed it in a cell line with inducible, localized oxidative stress in the nucleus. Our experiments show genome wide variability in 8-oxo-dG levels, with increased levels in complex and repetitive regions which were uncharted by previous short-read based methods. For the first time, we are able to simultaneously measure 5-mC and 8-oxo-dG at single-molecule resolution, revealing a strong 5-mC depletion within a 2-kilobase window surrounding 8-oxo-dG. Collectively, our approach showcases a methodology widely applicable to any synthesizable DNA base modification.

Results

8-oxo-dG is detectable using nanopore sequencing

To generate a ground truth dataset, we designed a set of 110 short synthetic oligos (46 base pairs each) that contained 8-oxo-dG in different genomic contexts. The oligo design (Fig. 1a) consists of three barcodes (7 base pairs each) to enable multiplexing. In addition, these barcodes have been optimized to facilitate signal segmentation. The oligos were designed to contain complementary overhangs (10 base pairs) enabling concatenation via hybridization and ligation to create long reads that can efficiently be sequenced on the nanopore platform (*Methods, oligo design*). The 8-oxo-dG base is surrounded by two predefined bases on either side (K_1, K_2, K_3, K_4), and five additional random bases (2 bases on the 5′ end (N_1, N_2), and 3 bases on the 3′ end (N_3, N_4, N_5)) to ensure context diversity. We decided to use predefined bases next to the 8-oxo-dG base, instead of random bases, because the effect of the modification would make it impossible to accurately determine which true bases corresponded to each oligo (Fig. 1b). Knowing the true sequence of an oligo is necessary to train a basecaller. Finally the oligos include an unmodified guanine base surrounded by the same bases as the 8-oxo-dG base to serve as a built-in control (Fig. 1a, Supplementary Data 1).

Using the concatenated synthetic oligos, we first established whether 8-oxo-dG has an effect on the basecalling performance of ONT's standard base calling model. We found that there was a substantial increase in basecalling errors specifically at the site of 8-oxo-dG (not basecalling 8-oxo-dG as G) and its neighboring bases (Fig. 1b). This increased error rate is also observable on R10 pore chemistry (Supplementary note 1). The basecaller's most likely mistakes, regarding 8-oxo-dG, are deletions (16.4%) or incorrect calls as cytosine (13.5%) or adenine (12.5%) (Supplementary Fig. 1). Moreover, the basecalling error rate is not uniform across different contexts, varying from 1.4% to 91% (Supplementary Data 2). This variability suggests that the signal alterations are 5-mer context dependent. To establish that the increased error rate was not an artifact of our oligo generated data, we also assessed the error rate on the reverse strand, which is devoid of any modified bases. As expected, here we did observe a consistently low error rate, which was much more uniformly distributed across the different bases and contexts (Supplementary Fig. 2). Interestingly, the cytosine opposite to the 8-oxo-dG base exhibited a slightly higher error rate (median 12.3%) compared to the other bases (median 5.0%) (Supplementary Data 3). Note that the DNA is unwound when it enters the pore, and hence this cytosine is no longer bound to 8-oxo-dG when it is being sequenced. We hypothesized that the unwinding of the DNA by the helicase might be affected by 8-oxo-dG. We therefore compared the speed (based on raw to expected signal alignment) between cytosines, and neighboring bases, paired to 8-oxo-dG and cytosines paired to G in the same 5-mer contexts. We observed that on average cytosines paired to 8-oxo-dG had fewer measurements, suggesting faster translocation of the DNA, than cytosines paired to G (77% of evaluated 5-mers) (Supplementary Data 5). This suggests that, despite only sequencing one DNA strand, the opposite strand may have an impact on translocation speed via structural effects that impact the proteins in the pore, thus affecting base calling accuracy.

https://doi.org/10.1038/s41467-025-60391-3

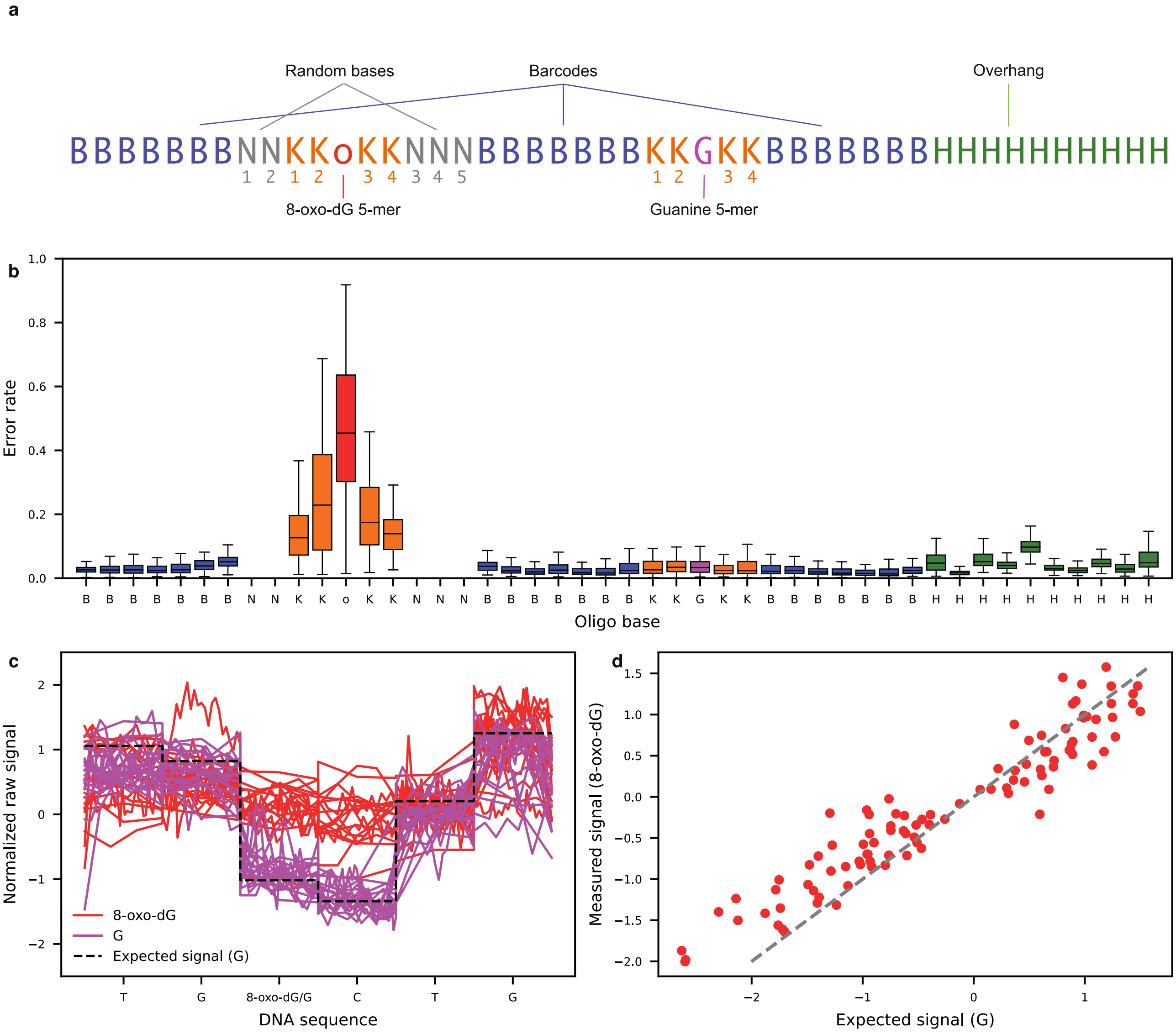


Fig. 1 | 8-oxo-dG has a detectable effect on the nanopore raw signal. a Schematic of the design of the 8-oxo-dG containing oligos. **b** Error rate per oligo base across all sequenced 8-oxo-dG containing repeats. Random bases are excluded from the analysis as we do not know their true reference. Horizontal bar represents the median, boxplots minimum and maximum bounds represent the 25th and 75th percentiles, respectively, and whiskers extend to 1.5 times the interquartile range. Data derived from ten thousand randomly sampled oligos. **c** Example of 8-oxo-dG (red) and G (purple) signal in the TG(8-oxo-dG/G)CTG context. Dashed black line indicates the expected signal value based on the G containing sequence. **d** Average measured normalized G signal and 8-oxo-dG signal per measured 5-mer as segmented using Tombo. Identity line indicated as the dashed gray line. Source data are provided as a Source Data file.

We then evaluated what specific effect 8-oxo-dG has on the raw signal. To this end, we aligned the measured raw signal to the expected signal (obtained from ONT's k-mer model) based on the known underlying sequence (*Methods, Raw data alignment*, Supplementary note 2). We observed a significant difference between the expected (which assumes an unmodified guanine) and measured signals. For example, in Fig. 1c it can be seen that the measured signal is higher than expected when 8-oxo-dG is present. This effect is also clear when evaluating other sequence contexts (Figure 1d), indicating that 8-oxo-dG has a clear effect on the measured signal. The measured signal is also significantly different from the other 3 bases (A, C and T) (Supplementary Fig. 6, Supplementary Fig. 7) and is most dissimilar to the pyrimidines. Considering these observations, we concluded that 8-oxo-dG is discernible from the other four canonical bases, suggesting that training a machine learning algorithm for its detection would be feasible.

Bonito basecaller fine-tuning

The nanopore signal originating from 8-oxo-dG containing sequences appears to have several context dependent distinctive features. We therefore sought to train a neural network model that could distinguish 8-oxo-dG from G. We opted for a two step approach, similar to ONT's *Bonito* + *Remora* combination (Fig. 2a). First, *Bonito* performs regular basecalling (A, C, G, T); and afterwards, *Remora* classifies the base of interest (e.g. C vs 5-mC, or in our case G vs 8-oxo-dG) as modified or not, using a small data window (e.g. 100 data points).

Since we already observed that the basecaller provided by ONT is not 8-oxo-dG aware, (Fig. 2b), we first fine-tuned a *Bonito* model to basecall 8-oxo-dG as G. To that end, we used a publicly available pre-trained version of *Bonito* from ONT, and fine-tuned it by training it with oligo concatemers containing 8-oxo-dG and publicly available human data from the telomere-to-telomere (T2T) reference dataset (*Methods, Genomic DNA: Telomere-to-telomere*). We also fine-tuned a model with

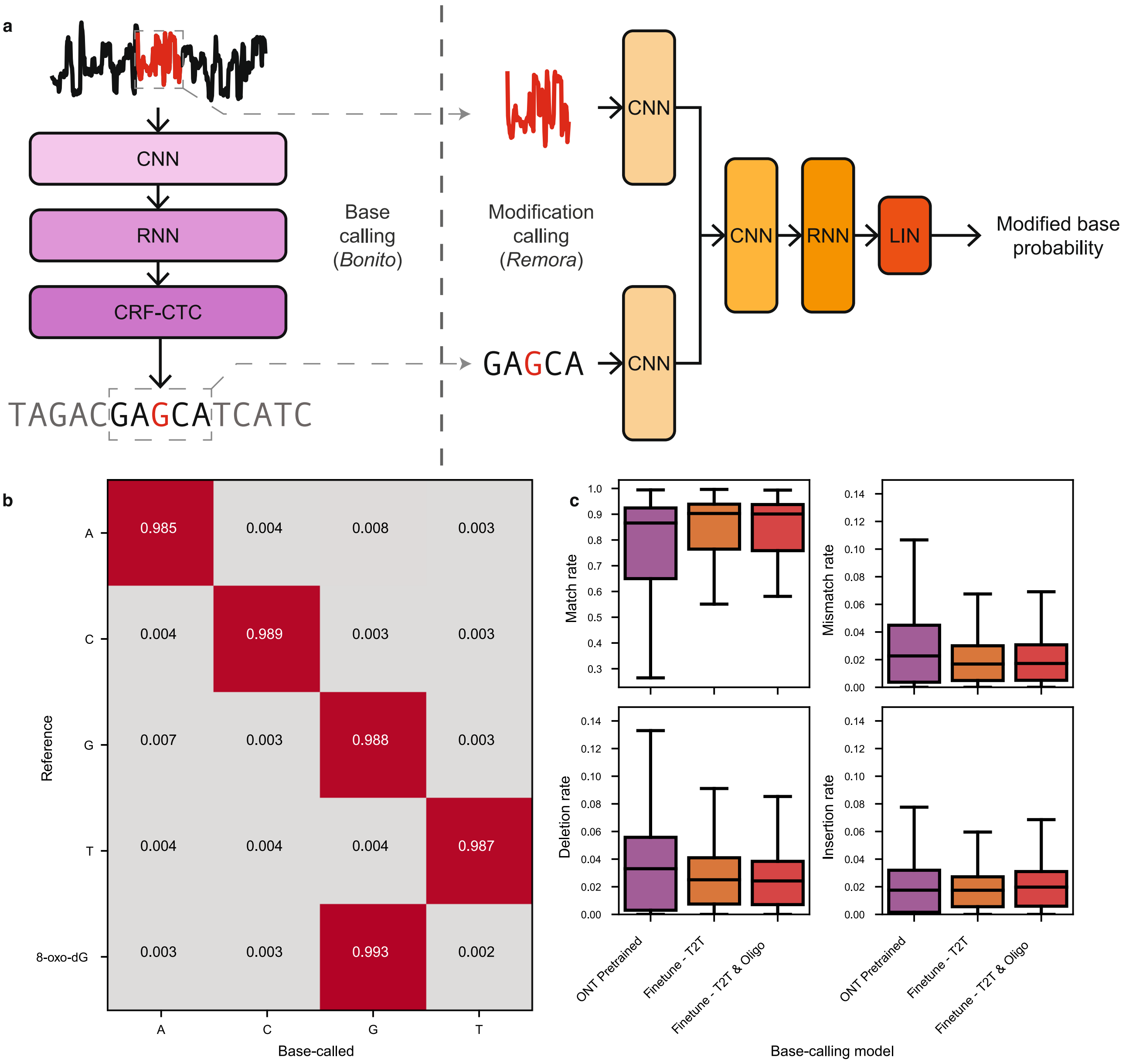


Fig. 2 | Fine-tuning of a *Bonito* model to basecall 8-oxo-dG. a Schematic of a two step approach to call modified bases in nanopore sequencing. First, the raw signal is basecalled to a DNA sequence. Upon basecalling the (potentially modified) base of interest, a small window of raw signal that corresponds to that particular base is cut; and together with a portion of the basecalled sequence, is given as input to a second model that predicts whether the base is modified or not. **b** Confusion matrix of the fine-tuned *Bonito* base caller. Values indicate the fraction of outcomes for each ground truth base. **c**, Match, mismatch, deletion and insertion rates of the pre-trained and different fine-tuned models using different datasets on the T2T human reference genome nanopore data. Horizontal bar represents the median, boxplots minimum and maximum bounds represent the 25th and 75th percentiles, respectively, and whiskers extend to 1.5 times the interquartile range. Data derived from the 32 thousand reads in the test set.

only data from the T2T dataset to assess the effect of fine-tuning itself on basecalling performance. We then evaluated the basecalling performance of these models on the T2T dataset, and their capacity to basecall 8-oxo-dG as G, in a cross-validated manner (*Methods, Base calling cross-validation*). We observed that after fine-tuning the models with oligo data, 8-oxo-dG was basecalled as G (~3% median error rate), and immediate neighboring bases are basecalled correctly at error rates similar to the rest of the bases (Fig. 2b, Supplementary Fig. 8). To ensure that the model did not overfit to the oligo data, we also evaluated the basecalling performance on the T2T dataset. We observed that basecalling was slightly more accurate on the fine-tuned model than on the pre-trained model (median error rate decreased from 14% to 10%); and that the addition of oligo data in the training did not have a negative effect on regular genomic basecalling (Fig. 2c). Our *Bonito* fine-tuned model can now accurately basecall 8-oxo-dG as G, and is also capable of regular genomic basecalling.

Detecting 8-oxo-dG with high specificity

We then trained a *Remora*-like model to distinguish 8-oxo-dG from unmodified guanine. We started by training a base model which takes as input both 100 data points of signal and the basecalled 7-mer. Both features were centered around the guanine of interest. The model then outputs a score between zero and one for the likeliness of that particular guanine being 8-oxo-dG (Supplementary Fig. 9). We trained and tested the model on both positive and negative examples of 8-oxo-dG from our synthetic oligo dataset, as well as negative examples of G

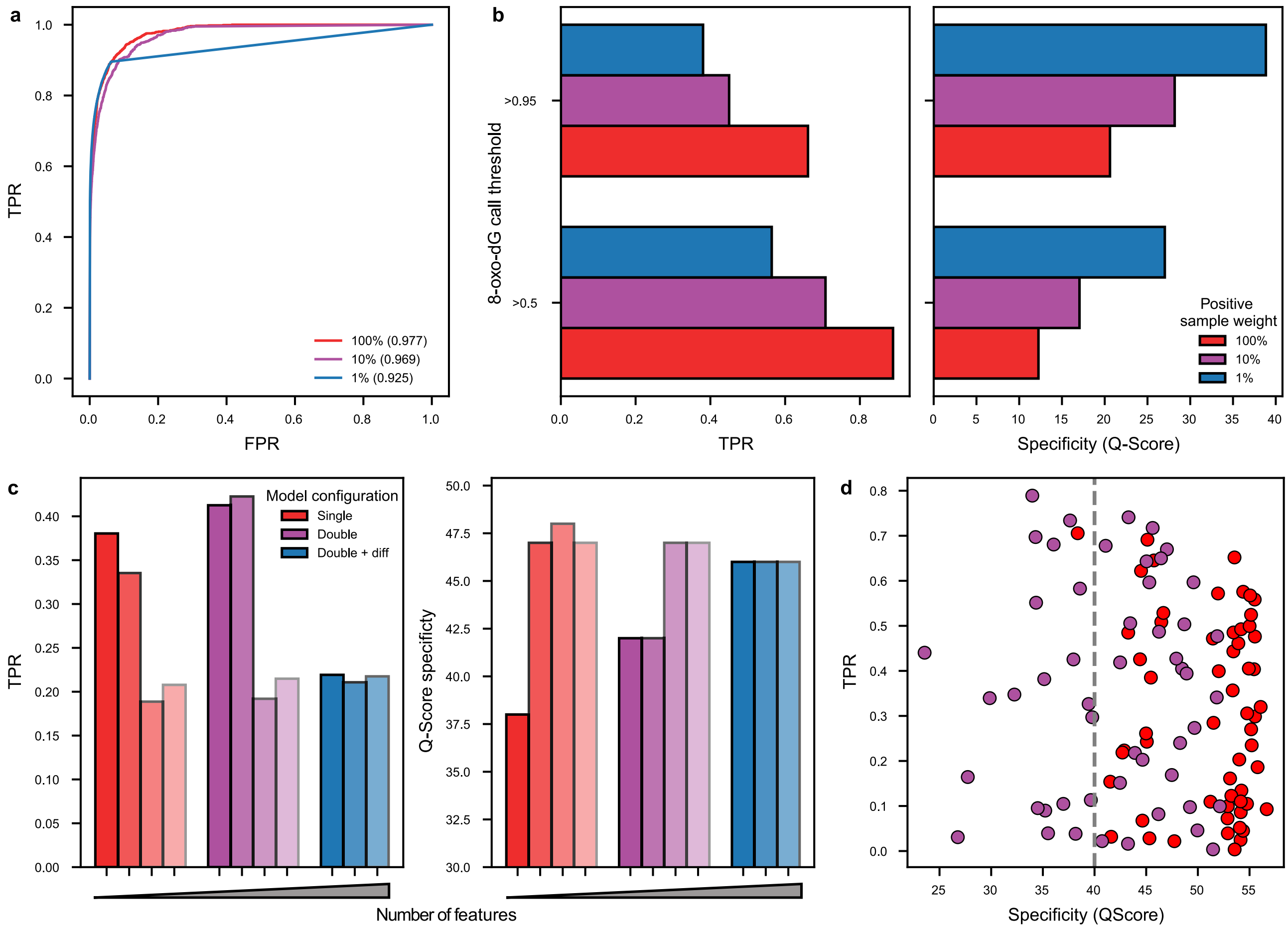


Fig. 3 | Performance of a *Remora* model. a ROC curves on three Remora models with different positive class weights (100%, 10% and 1%), values indicate the AUC. The straight line shape of the receiver operator characteristic (ROC) curve for the 1% weight model was due to a small number of negative samples with a high 8-oxo-dG score, which reduces granularity in the ROC curve thresholds (Supplementary Fig. 10). **b** TPR and Q-Score specificity evaluated on the test fold for two different thresholds (0.5, and 0.95) on the three Remora models with different positive class weights (100%, 10% and 1%). **c** TPR and Q-Score specificity evaluated on the test fold for the experiment in which additional features were added sequentially. Metrics are calculated using a 0.95 threshold. Models include additional features in a cumulative manner, from left to right: basecalls, expected signal, difference between expected and measured signal, and basecall phred quality scores. Red bars include features from the fine-tuned model, purple bars also include features from the Bonito pre-trained model, blue bars include the difference between the features of the Bonito fine-tuned and pre-trained models. **d** Performance of the Remora model with expected signal as feature per 5-mer at the >0.95 score threshold. Red colored dots indicate 5-mers for which there were no false positive calls, for these 5-mers the QScore was annotated as if a single false positive call was made. Source data are provided as a Source Data file.

from the T2T data (*Methods, Modification calling cross-validation, Remora training*). The latter enables performance estimation of 8-oxo-dG in the presence of various modification typically found in DNA. This initial model achieved 93.3% accuracy and 94% specificity at the 0.5 score threshold, with an area under the curve (AUC) of 0.97 (Fig. 3a). This level of performance is comparable to state-of-the-art 5-mC calling[46]. Problematically, 8-oxo-dG is not as prevalent as 5-mC, and requires a model with near perfect specificity to ensure reasonable FPRs (Supplementary Data 4). For example, assuming 1-100 8-oxo-dG per million guanines[21], the base model (specificity of 98.6% at the 0.9 score threshold), would be making 100-10000 false positive (FP) calls per true positive call, thus obscuring any real signal. Ideally, we would have a specificity of at least an order of magnitude lower than the prevalence of 8-oxo-dG (Q50-Q70, i.e. 1 false positive per 0.1–10 million guanine calls; (*Methods, Specificity Q-scoring notation*)). We subsequently explored whether candidate filtering based on basecall errors, or metric learning approaches would improve performance, but we did not observe any major improvements (Supplementary note 3).

To further improve the specificity of our Remora-like model, we explored the use of class weights during training to emphasize the importance of true negatives (TN). For this purpose, we trained two additional models in which the positive class (8-oxo-dG) had a weight of 10% or 1%, relative to the negative class (G). Using a lower weight on the positive class had a negative impact on global performance, as the AUC was reduced to 0.96 and 0.92, respectively (Fig. 3a). However, it increased specificity from Q12 to Q27 when using the 0.5 cutoff; and is further increased from Q20 to Q38 when using the more stringent 0.95 cutoff. However, the true positive rates (TPR) decreased from 88% to 56%, and from 66% to only 38% respectively (Fig. 3b). These results indicate that reducing the weight of the positive class leads to a significant increase in specificity at the cost of an increase in false negatives.

We finally explored if additional signal and sequence features could improve model performance. For example, a signal feature would be the expected signal given the basecalled sequence, and a sequence feature would be the base calling quality scores (*Methods, Remora feature-engineering*). We evaluated these features in a feature expansion experiment in which we sequentially added one additional feature at a time. We trained these models with a 1% weight on positive samples since our objective was to further reduce the FPR. Our results indicate that, after adding the expected signal as a feature, the specificity increased from Q38 to Q47 with only a mild 5% reduction of the TPR. Adding additional features, such as the difference between measured and expected signal, and the phred quality scores further increased specificity (to Q48) but with a TPR reduction from 40% to 20% (Fig. 3c, red). We then added the same features, but based on the pre-trained Bonito base calling model, since its base calls would differ mostly if an 8-oxo-dG was involved. By adding the pre-trained model basecalls and expected signal, the TPR increased by 3%, but the specificity only improved relative to the base model (raw signal + basecalls), not the other models with additional features (Figure 3c, purple). We finally added the difference between the features of the fine-tuned and pre-trained models. These models had similar performance (specificity Q47) and the TPR was at similar levels as the previous models (21%) (Fig. 3c, blue). Models which contained the expected signal achieved significantly worse TPRs, and we observed that these models greatly overfit during training (Supplementary Fig. 11). We hypothesize that some of these features are predictive but easy to overfit to. For example, just by using the difference of expected signal based on the basecalls from the two models, one can achieve an AUC of 0.79 (Supplementary Fig. 14). In conclusion, the largest improvement was derived by adding the expected signal based on the basecalled sequence as it increased specificity the most without compromising the TPR too much, which also holds true for the less stringent threshold of 0.5 (Supplementary Fig. 15).

Our models have demonstrated an FPR within the same order of magnitude as the expected 8-oxo-dG levels. However, we hypothesized that this error may not be distributed uniformly across the different k-mers, and some k-mers might have an increased FPR. We decided to use $k = 5$ since it is the maximum number of bases with a known reference around 8-oxo-dG in our oligo data. Indeed, four 5-mers demonstrated an AUC of approximately 0.5, equivalent to random guessing. In contrast, 73 other 5-mers exhibited an AUC exceeding 0.9, with the remaining 5-mers falling in between (Supplementary Fig. 16., for performance metrics per 5-mer and threshold see Supplementary Data 6). We also observed that there is a large spread in terms of specificity, ranging from Q24 to Q56 (median Q47) (Fig. 3d). Notably, for some 5-mers (58 out of 110, indicated in red in Fig. 3d), we did not detect any false positives, and we calculated their specificity as a single false positive pseudocount. We did not observe that k-mer performance was linked to their coverage in the oligo data (Supplementary Fig. 12; but it was related to their sequence, with k-mers rich in cytosine and guanine performing worse (Supplementary Fig. 17). We also observe that the performance spread is also large in terms of sensitivity (average 0.33 ± 0.22 s.d.), meaning that on average, one out of three 8-oxo-dG molecules will be detected. Because of the low abundance of 8-oxo-dG, we decided to only consider 5-mers that reached a minimum specificity of Q40 (87 out of 110) for subsequent analysis to ensure that the signal to noise ratio is maximized, which corresponds to ~35% of the guanines in the human genome.

H_2O_2 production close to the DNA increases 8-oxo-dG levels

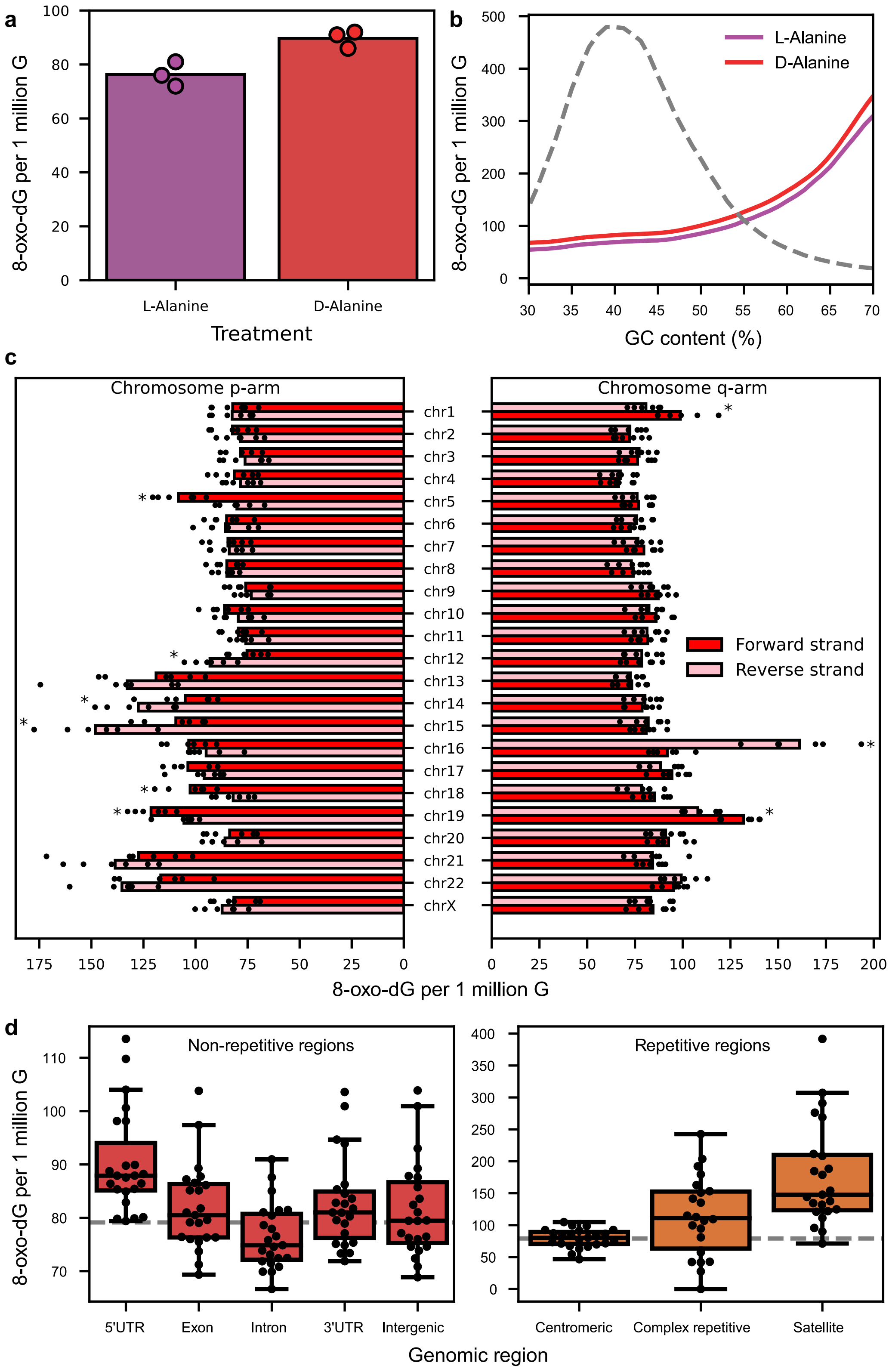
Using our nanopore based 8-oxo-dG modification caller, we sought to explore 8-oxo-dG's locations genome-wide, including repetitive regions previously unexplored by short-read sequencing techniques. To this end we used hTert-immortalized RPE1 cells that express D-amino acid oxidase (DAAO) from *R. gracilis* as a fusion with Histone H2B (RPE1-hTERT-DAAO^{H2B}). Upon administration of D-Alanine (D-Ala), DAAO produces H_2O_2 in the vicinity of DNA, given its fusion to H2B. Exposure to D-Ala has been shown to give rise to C>A mutations in this mode (in a p53 KO background), suggesting that DAAO^{H2B} activation leads to formation of 8-oxo-dG in the genome[47]. RPE1-hTERT-DAAO^{H2B} cells were exposed to a 2 h pulse of 20 mM D-Ala, after which we harvested and sequenced the DNA and analyzed it using our 8-oxo-dG model (*Methods, Cell line sequencing*).

We observed that H_2O_2 production at the chromatin resulted, somewhat unexpectedly, only in a modest increase of 12 additional detected 8-oxo-dG modifications per million guanines (16% increase, p value = 0.016) as compared to control (Fig. 4a). We noted that RPE1-hTERT-DAAO^{H2B} p53$^{-/-}$ cells had overall higher levels of 8-oxo-dG already in control treated cells, compared to RPE1-hTERT-DAAO^{H2B} p53 wild-type, suggesting that the loss of p53 leads to general higher levels of 8-oxo-dG, either by enhanced oxidative stress or slower removal of the modification. After induction of H_2O_2 production, both p53$^{-/-}$ and wild-type cells reach similar total 8-oxo-dG levels (Supplementary Fig. 18). We correlated 8-oxo-dG levels to GC content in 1kb bins (Fig. 4b). We observed that GC content was highly correlated with 8-oxo-dG levels in all conditions (Supplementary Fig. 19). Although this might be expected, we noticed that the rate between 8-oxo-dG and GC content changed at approximately 50%. This indicates that high GC content regions get more easily oxidized, or less effectively repaired, based on the rates measured at lower GC content regions (Supplementary Fig. 20). We next evaluated whether 8-oxo-dG levels were depleted or enriched in specific regions in the genome (Fig. 4c). Overall, we observed differences in 8-oxo-dG levels depending on the DNA strand (Supplementary Fig. 21) and the chromosomal arm (Supplementary Fig. 22). The forward strands of chromosomes 5, 18 and 19 (p-arms) and chromosomes 1 and 19 (q-arms) have a significant (p value < 0.05) 14% or more 8-oxo-dG compared to their reverse strand counterparts, and the reverse strands of chromosomes 12, 14, 15 (p-arms) and chromosome 16 (q-arm) have 9% or more 8-oxo-dG compared to their forward strand counterpart. These particular differences are treatment and p53 status independent (Supplementary Fig. 23, Supplementary Fig. 24, Supplementary Fig. 25, Supplementary Fig. 26). We then analyzed the relative 8-oxo-dG levels across different genomic regions (Fig. 4d). Again, we observed a global increase in 8-oxo-dG levels after H_2O_2 production, irrespective of genomic region (Supplementary Fig. 27). Non-repetitive and centromeric genomic regions showed 8-oxo-dG levels that roughly align with the overall rate of ~80 8-oxo-dG per 1 million G (Fig. 4d, gray dashed line); with the exception of the 5′-UTR, which has the highest median levels, as also previously reported by Ding et al.[36]. On the other hand, complex repetitive and satellite regions contained, relatively, the most 8-oxo-dG with high variance between chromosomes (Fig. 4d). We fitted a linear model to evaluate whether 5-mer content, or the relative abundance of the different genomic regions, were causative for these observations, however none could explain the observed variability (Supplementary note 4).

Finally, we evaluated 8-oxo-dG at telomeric regions since they are guanine rich, and their oxidation has been linked to senescence and telomere shortening[48,49]. To that end, we combined the sequencing data from all experimental conditions, and obtained a total of 487 reads that primarily mapped to the telomeres. We observed a strong bias towards reads mapping on the p-arm reverse strand and q-arm forward strand (Supplementary Fig. 28); which can be explained by blunt end requirement of the ligation mechanism in the library preparation (Supplementary note 5). Notably, we only detected 2 high confidence 8-oxo-dG calls on the q-arm telomeres of chromosomes 16 and 19 (Supplementary Fig. 29), a value ten times lower than expected, given the 50% GC content of the region. This result might indicate that 8-oxo-dG is very efficiently repaired at telomeres as a preventive measure to downstream complications.

The 8-oxo-dG trinucleotide context profile does not match the C>A mutational signatures

Because of its ability to mispair with A, unrepaired 8-oxo-dG present at the time of DNA replication, oxidative stress has been the proposed mechanism behind the COSMIC mutational signatures 18 and 36[17–19]. We therefore wanted to compare C:G>A:T (here mentioned solely as



C>A) mutations to the derived 8-oxo-dG profile to establish if the resulting mutations had the same trinucleotide context as the detected 8-oxo-dG. Using the same RPE1-hTERT-DAAO^{H2B}-p53$^{-/-}$ cell lines, van Soest et al.[47] performed Illumina sequencing to analyze the mutational profile caused by DAAO^{H2B}-derived H_2O_2. This analysis was also performed in a p53 WT background in a parallel experiment not shown in van Soest et al.[47]. We therefore used these mutational profiles and compared them to the 8-oxo-dG profile. It must be noted that the mutational profiles were obtained after 4 rounds of 20mM D-Ala treatment and recovery; while for our 8-oxo-dG analysis, we used a single 20mM D-Ala pulse followed by immediate DNA harvesting. We therefore abstain from drawing any strong conclusions regarding the relationship between number of mutations and 8-oxo-dG due to the treatment difference.

Fig. 4 | 8-oxo-dG distribution across the genome. a Overall 8-oxo-dG molecules per 1 million G molecules per L-Alanine and D-Alanine treated cells. Error bars indicate minimum and maximum calculated values. Dots indicate the values per sequenced condition ($n = 3$). **b** 8-oxo-dG levels across different GC (%) content bins. Blue and red lines indicate values for L-Alanine and D-Alanine treated cells respectively. Gray dashed line indicates the distribution of measured GC content bins. **c** 8-oxo-dG levels per chromosome, chromosome arm and DNA strand. Bars indicate average values. Dots indicate the values for all conditions ($n = 6$). Asterisks indicate a significant p value (<0.05) derived from a two-sided t-test between the values of the forward and reverse strands. Exact p values can be found in the source data. **d** Distribution of 8-oxo-dG levels per genomic region type across all conditions and chromosomes ($n = 23$). The dashed gray horizontal line indicates the overall 8-oxo-dG level across all conditions, irrespective of genomic region. Horizontal bar represents the median, boxplots minimum and maximum bounds represent the 25th and 75th percentiles, respectively, and whiskers extend to 1.5 times the interquartile range. Black dots indicate the underlying data. Source data are provided as a Source Data file.

The mutational profiles were highly similar in terms of COSMIC mutational signatures present (minimum cosine similarity of 0.95) (Supplementary Fig. 30), irrespective of H_2O_2 induction and p53 status. The only major difference between the samples was the amount of mutations[47]. Since our aim is to compare the mutational signatures, we combined the mutational profiles of all the sequenced conditions (Fig. 5a). To make the mutational profile comparable to the 8-oxo-dG profile, we recalculated the mutational profile including only mutations in the 5-mer contexts in which our model performed with at least Q40 specificity, which reduced the total number of C>A variants from 6880 to 2506. We also normalized the relative contribution of each 3-mer based on the number of 5-mers that passed the specificity threshold (Supplementary Fig. 31). We observed that these limitations did not drastically change the originally derived mutational profile (Supplementary Fig. 32).

We detected many more 8-oxo-dG bases than C>A mutations in every analyzed 3-mer context despite exposing the cells to a single rather than four rounds of DAAO^{H2B}-derived H_2O_2 production (Supplementary Fig. 33). The C>A mutational profile (Fig. 5a) and the 8-oxo-dG profile (Fig. 5b) however only moderately agree (cosine similarity of 0.33). Our results do not disprove the hypothesis that 8-oxo-dG is the underlying cause of C>A mutations, but rather that the rate at which 8-oxo-dG leads to a C>A mutation is 3-mer context specific. Notably, the C>A mutational profile has a high similarity (cosine similarity of 0.95) with the 3-mer abundance profile of the human genome (Supplementary Fig. 34) and the observed mutational profile thus seems to more closely relate to 3-mer abundance than to 8-oxo-dG location itself (Fig. 5c). This suggests that C>A mutations are likely driven by, besides 8-oxo-dG, additional mechanisms such as replication timing[50]. Interestingly, we do observe strong 8-oxo-dG enrichment and depletion for certain 3-mers (Fig. 5d), in particular 3-mers that contain a CG or a GG motif. This result coincides with our previous observation of 8-oxo-dG enrichment in high GC content regions, but might hint at other mechanisms such as a relationship with CpG methylation.

8-oxo-dG levels negatively correlate with methylation levels

Previous work has linked 8-oxo-dG with both inhibition of DNA methylation[23,51], as well as active demethylation via TET enzyme recruitment by OGG1[24,52]. Whereas previous methods for genomic 8-oxo-dG detection precluded the simultaneous assessment of 8-oxo-dG and other base modifications, our approach can, for the first time, look at 8-oxo-dG, 5-mC and 5-hmC on the same DNA molecule. Importantly, all base modification callers are trained to detect a base modification in the absence of other base modifications. We expect that these models fail to detect close proximity (~10 bases window) modification combinations. We therefore compared the methylation and hydroxymethylation status of the surrounding CpG sites between G and 8-oxo-dG containing regions in a 10 kilo-base window in the same DNA molecule.

We observed that there is a significant general decrease in methylation levels around oxidized guanines compared to non-oxidized (Fig. 6a, *Methods, 8-oxo-dG related methylation analysis*). The decrease in methylation levels correlates with distance to the oxidized base: it is lowest close to 8-oxo-dG (~45% methylation), and recovers to average genome-wide methylation levels at approximately 2000 bases from 8-oxo-dG (~54% methylation). This effect was observed irrespective of p53 status and DAAO^{H2B}-derived H_2O_2 production, suggesting a DNA damage independent relationship (Supplementary Fig. 35). We then grouped the data by 3-mer analogous to the profile analysis and observed that the decrease in methylation levels correlated with the 3-mer enrichment for 8-oxo-dG (Fig. 6b, c). The decrease in methylation is largest for 3-mers with a CG motif, with the exception of CGT. A decrease in methylation is also observed for 3-mers without a CG motif, this is most apparent for GGA, which has a similar decrease in methylation level as compared to CG containing 3-mers. Interestingly, in 3-mers without a CG motif the methylation levels reached average genome wide levels at approximately 1 thousand base-pairs from 8-oxo-dG, while methylation levels in 3-mers with a CG motif did not return to baseline within 5kb from 8-oxo-dG [Figure 6c]. Strangely, the AGG was the only 3-mer that displayed overall higher methylation levels for 8-oxo-dG containing molecules (Fig. 6c). We observed that 5-hmC levels were slightly higher in 8-oxo-dG vs guanine containing regions (average ~3.5% vs ~3%) (Supplementary Fig. 36). Similarly to 5-mC, we did not observe p53 or H_2O_2 dependent effects (Supplementary Fig. 37). However, upon inspection of the 5-hmC levels per 3-mer, we observed that there is no difference in 5-hmC levels per 3-mer with the exception of two cases (Supplementary Fig. 38). First, 5-hmC levels are highest around the GGA context, reaching an average of 6% in regions containing 8-oxo-dG. And secondly, the opposite effect happens in the CGA context, where 5-hmC levels are highest for regions with G (average ~5,8%) (Supplementary Fig. 38).

Our approach enabled simultaneous assessment of 8-oxo-dG, 5-mC, and 5-hmC on the same DNA molecule for the first time. Together with our analysis of the mutational signatures, our results suggest that 8-oxo-dG abundance has a primary role in epigenetic regulation, and that its mutagenic effect would play a secondary role.

Discussion

Nanopore sequencing holds the potential to detect any base modification, both on DNA and RNA. However, due to technical challenges, accurate models have only been established for a limited number of modifications[45]. Available models focus on highly prevalent modifications, like 5-mC or 6-mA because they can be generated enzymatically, are highly stable, and therefore training data can be easily obtained and verified through alternative sequencing technologies[45]. However, these approaches are not widely applicable to less abundant modified bases[53]. In this study, we leverage synthetic DNA to bridge this gap, enabling the generation of a fully controlled ground truth dataset. We show that our approach can successfully be used to detect 8-oxo-dG without any special sample preparation or antibody pull-down. Furthermore, our methodology has the potential to be expanded to the nanopore-based detection of other base modifications, and to newer pore chemistries (e.g. R10), although this would require the design of the synthetic oligos to be substantially adapted. For R10 pores particularly, due to the double read head which increases the number of bases that influence the signal, the required amount of random bases would need to be considerably larger to mitigate the risk of sequence biases. Future work should also assess whether Bonito fine-tuning and Remora training could be applied as is or if further optimization would

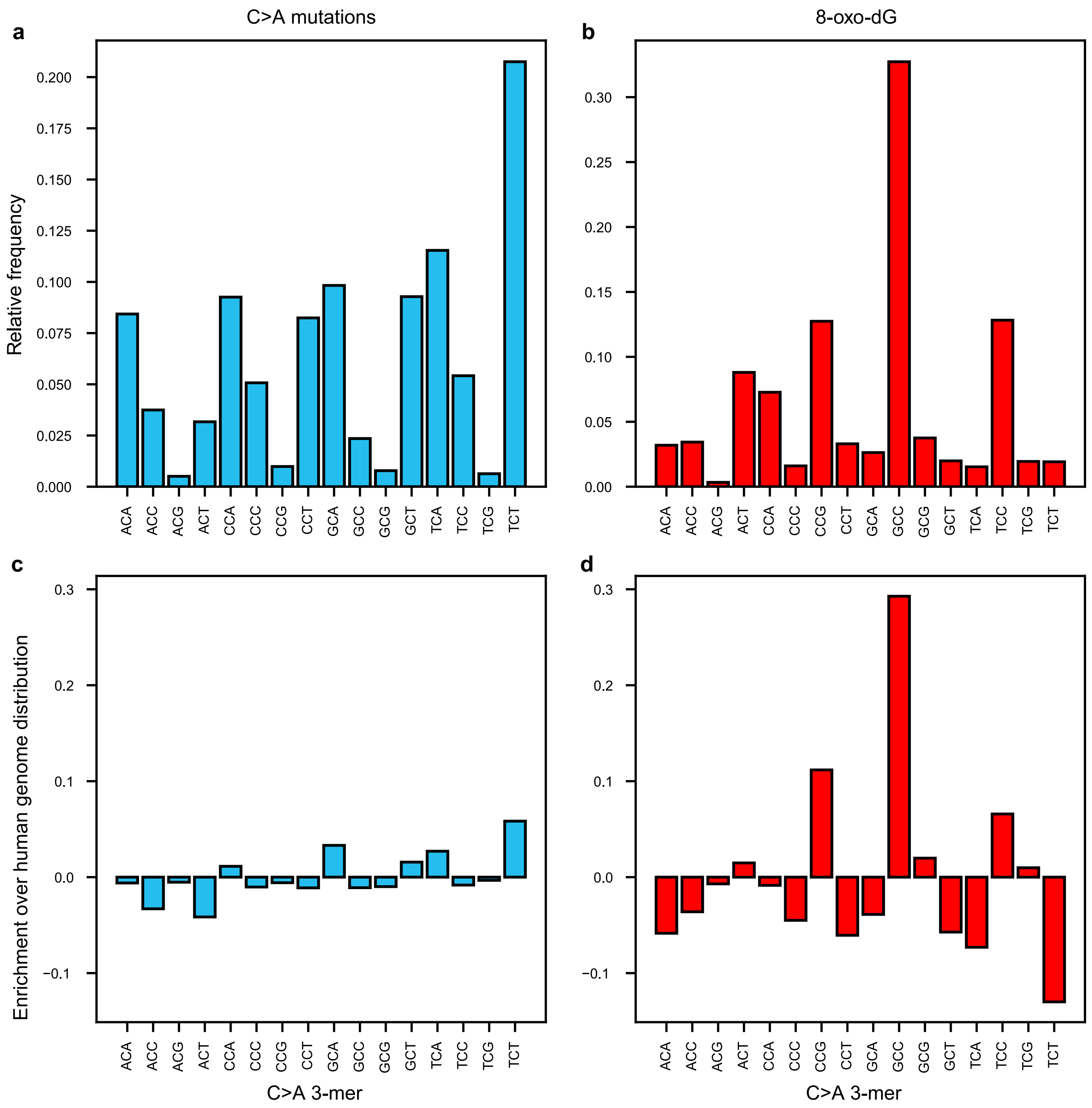


Fig. 5 | Mutational and 8-oxo-dG signatures. a Combined mutational signature (C>A or G>T) from all the cell lines derived from Illumina sequencing. **b** 8-oxo-dG normalized abundance profile for each 3-mer. Note that the 3-mers are annotated as the reverse opposite strand of 8-oxo-dG (e.g. ACT would be equivalent to AXT, where X denotes 8-oxo-dG). **c** Mutation enrichment of each 3-mer normalized to the abundance of each 3-mer in the human genome. **d** 8-oxo-dG enrichment of each 3-mer normalized to the abundance of each 3-mer in the human genome. Source data are provided as a Source Data file.

be needed to support newer pore chemistry with a larger receptive field, or that engineering of other features could aid model convergence. Leveraging R9 pores for detecting base modifications (and other applications) may be therefore preferable, since the small read head greatly simplifies training data generation and model training as shown in our work and a number of other recent papers[54–57].

Our work showcases that developing deep-learning models required for detecting rare modifications from the raw nanopore squiggles is not trivial. False positives drastically decrease the signal-to-noise ratio and, if not low enough, will mask any inherent biological signal[40]. We address this by using label weighting and feature engineering. However, while we achieved a major increase in specificity, this comes at the price of reduced sensitivity of the model. Rare base modifications also pose a challenge in regular base calling. As we have shown, 8-oxo-dG naive basecallers have a significantly increased error rate on molecules that contain an 8-oxo-dG modification. Therefore, until these rare modifications are integrated in the basecaller training pipelines, perfect accuracy for all natively sequenced molecules will not be achievable.

Using our highly specific model, we show differential 8-oxo-dG levels across the different chromosomes and genomic regions of RPE1-hTERT cells. Our observations fit with the variability shown in OGG1 driven apurinic sites[34]; and we also observe a similar distribution of 8-oxo-dG over non-repetitive regions as reported by Ding et al.[36]. For the first time, we can evaluate 8-oxo-dG abundance on highly repetitive sequences, which show high levels of 8-oxo-dG, indicating either

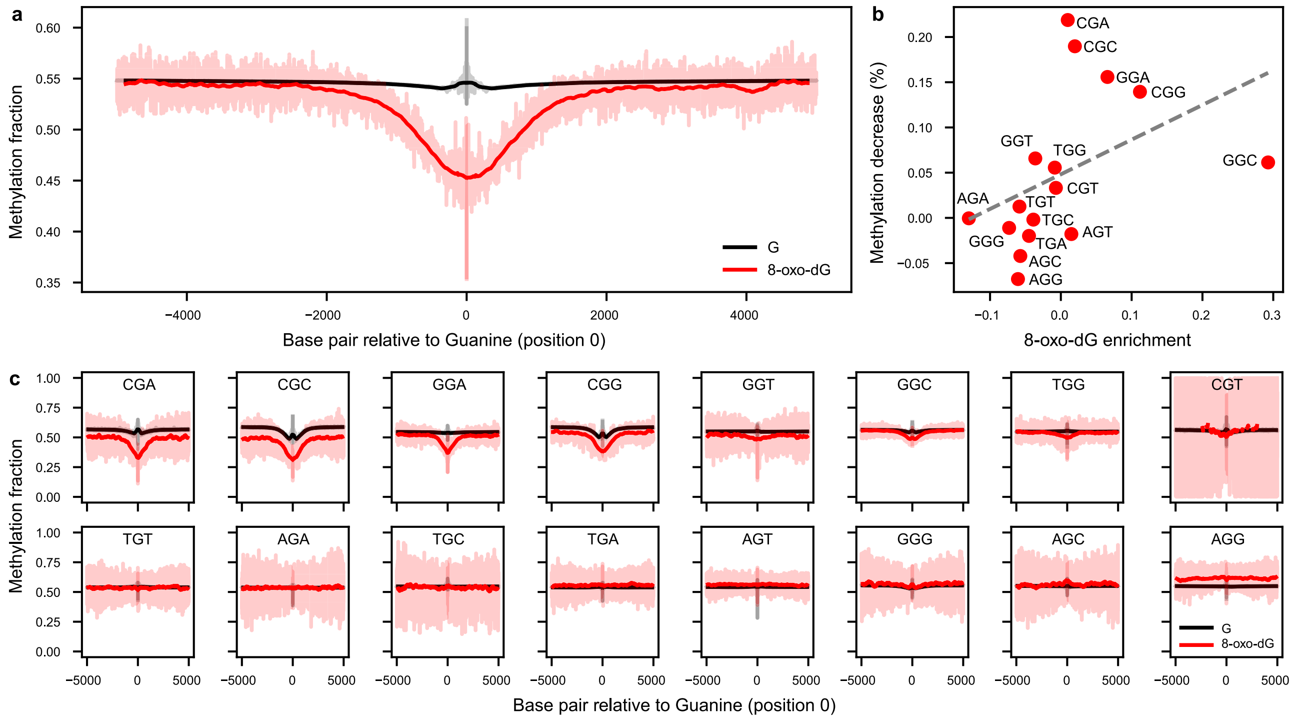


Fig. 6 | Genomic de-methylation surrounds 8-oxo-dG. a Methylation levels for 8-oxo-dG (red) and Guanine (black) containing reads. Reads are centered around (position zero) the 8-oxo-dG or Guanine. Methylation levels are obtained from the same molecule. The transparent red line indicates the underlying data, the dark gray line is the result of an 11 base average convolution. Data from all experimental conditions is included, see (Supplementary Fig. 35) for the per condition analysis. **b** Correlation between 8-oxo-dG enrichment as in Figure 5d and the methylation difference at position zero. **c** Similar to (**a**), but data has been split based on the 3-mer (reverse complement) surrounding the guanine. Source data are provided as a Source Data file.

increased susceptibility to oxidation or less efficient repair. Surprisingly, we observed very low 8-oxo-dG levels at the telomeres despite their oxidation susceptibility[58]. This might suggest that telomeric 8-oxo-dG is very efficiently repaired to avoid telomere shortening and the downstream induction of cellular senescence[48,49]. However, our coverage was limited, and further research using telomere enrichment techniques[59] might advance our understanding of the repair mechanisms of 8-oxo-dG at telomeres[60].

The mutational process behind C>A mutations in COSMIC signatures 18 and 36 has been long attributed to oxidative stress and unrepaired 8-oxo-dG[17–19]. Our results show that, primarily, 8-oxo-dG does not strongly correlate with the same nucleotide context of the signatures. Rather, these signatures strongly follow the human 3-mer genomic content, which indicates that the mutational rate of 8-oxo-dG is 3-mer specific, and could be driven by other mechanisms such as replication timing[50]. Importantly, our results do not discard 8-oxo-dG as the cause of C>A mutations, since the modified base is far more abundant than the resulting mutations. Rather, we envision that these mutations are secondary to the role of 8-oxo-dG in epigenetic regulation.

We observed a distinct decrease in CpG methylation levels around 8-oxo-dG, however it is unclear whether CpG demethylation precedes guanine oxidation, or vice-versa. It has been shown that 8-oxo-dG inhibits methyl-transferase enzymes[23,51,52], which would indicate that oxidation happens in already de-methylated regions. Another possibility would be that active demethylation is promoted via the recruitment of TET enzymes by OGG1[24]. At first sight, this model suggests that, 8-oxo-dG then would have been removed by OGG1, and therefore would not be detected in the first place. However, OGG1 oxidation can block its glycosylase and lyase activity, but not its binding to 8-oxo-dG[61], which would explain why we still measure 8-oxo-dG, and indicates that oxidation happens before de-methylation. Finally, some histone demethylases are known to produce H_2O_2 as part of the histone demethylation reaction (e.g. H3K9[62] and H3K4[63], which could cause the formation of multiple 8-oxo-dGs while oxidatively inhibiting the recruited OGG1. This model would link histone and DNA epigenetics through redox regulation. Perturbation experiments on OGG1, in combination with TET and histone demethylases, would provide valuable insights to further elucidate the underlying regulatory mechanism.

In conclusion, our work showcases a viable methodology for rare modification detection using nanopore sequencing, which could be applicable to detect any synthesizable base modification. These models can then be later used to decipher and further understand epigenetic regulatory mechanisms and the interplay between DNA modifications, as well as potential implications for disease mechanisms and biomarker detection.

Methods

Oligo design

A total of 110 oligos were designed in complementary pairs, wherein each forward oligo contains an 8-oxo-dG base and was paired with its complementary reverse strand oligo devoid of any base modifications. Oligos are 46 base pairs long and contain three barcodes, two 8-oxo-dG/G k-mer regions of 10 and 5 bases respectively, and a 10 base overhang. The barcodes are 7 bases long and were defined to have the lowest basecalling error rate possible (based on the T2T dataset (*Methods, Genomic DNA: telomere-to-telomere*)) and a high sequence entropy (meaning that the same base was never repeated sequentially). The 10 base overhangs are complementary between the forward and reverse strands. The complementary overhangs allows for the creation of long concatemers where one strand is always composed by forward (8-oxo-dG containing) oligos, and the other strand is always composed by reverse (non 8-oxo-dG containing) oligos. After hybridization, the concatemers can be ligated to create long molecules that are more readily sequenced on the nanopore platform. Between the first and second barcode lies an 8-oxo-dG that is immediately surrounded by two known bases at either side, which provides the known sequence context. Similarly, between the second and third barcode a guanine is surrounded by the same four bases. To maximize sequence variability around 8-oxo-dG, five additional random bases were added: two on the 5′ end, and three on the 3′ end.

Oligo concatenation

Oligos were first annealed in a thermocycler by mixing the two complementary oligos in equimolar rates. Oligos were diluted in 1× T4 ligase buffer (NEB REF #B0202S) in a total volume of 10 μL. The mixture was heated to 95 °C for 5 min, and then the temperature was decreased by 0.1 °C per second until 4 °C. After hybridization, oligos were concatenated by ligation. 9 μL of 1× T4 ligase buffer (NEB REF #B0202S) and 1 μL of T4 ligase (NEB REF #M0202S) were added to the solution. Afterwards, the mixture was incubated at 16 °C for 18 h. T4 ligase was then inactivated at 65 °C for 10 min. The resulting DNA was cleaned using the Qiagen PCR & Gel Cleanup Kit (Qiagen REF #28506) according to the manufacturer's instructions, and eluted from the column using Milli-Q water. This process was repeated 3 times to increase the concatemers length (Supplementary Fig. 39). Oligo concatemer concentrations were then quantified via Nanodrop (Thermo Scientific NanoDrop 2000 #ND-2000) and multiplexed prior to library preparation in equimolar rates as indicated in Supplementary Data 1.

Genomic DNA: telomere-to-telomere

We used existing nanopore sequencing data from the reference genome (NA12878/GM12878, Ceph/Utah pedigree) dataset[64]. This T2T dataset contains many different sequencing runs. We arbitrarily chose three experiments so that each different ligation kit (rapid, ligation and ultra) would be included: FAB42828, FAF09968 and FAF04090. We assumed all the sequenced DNA did not contain any 8-oxo-dG as it had been prepared using the standard library preparation protocol, which contains a repair step with Fpg, a DNA glycosylase that removes 8-oxo-dG.

Library preparation

Oligo concatemers and genomic DNA samples were library prepared using the SQK-LSK109 ligation kit according to the manufacturer's instructions (or the SQK-LSK114 for the R10 sequenced batch), with the exception that the FFPE repair step was skipped. The exclusion of this step was deliberate and meant to preserve 8-oxo-dG in our samples since the enzymatic function of Fpg as a DNA glycosylase is responsible for removing 8-oxo-dG from DNA.

Nanopore sequencing

Samples were sequenced using MinION R9.4.1 flow cells for 72h using the GridION device. MinKNOW v22.12.5 or earlier was used to ensure that all our samples were sequenced using 4KHz sampling. Oligo concatemers were multiplexed as indicated in Supplementary Data 1. Each genomic DNA sample was sequenced individually in a single MinION flow cell. K-mer performance per flowcell is represented in Supplementary Fig. 13. One batch of oligo concatemers was sequenced using MinION R10.4.1 flow cells for 72h using the GridION device (Supplementary note 1).

De-multiplexing and reference assignment

Although oligos were concatenated separately, the library preparation protocol contains a ligation step in which all oligo concatemers are pooled together. Therefore, it is possible to get oligo hybrids that contain different reference sequences. For this reason, oligo samples were de-multiplexed using a custom algorithm to detect the individual oligo repeats and barcodes within a read. Given the basecalls of a read, all possible oligo reference sequences within that batch are aligned to the basecalled sequence using a semi-global alignment algorithm (as implemented in the parasail library[65]). The reference sequence with the highest number of matches is considered as the true underlying sequence for that portion of the basecalls. However, only matches to the barcode and guanine 5-mer portions of the reference sequence are considered (max of 26 matches), since the overhang sequence is common for all oligos and the basecalls surrounding 8-oxo-dG cannot be trusted. Then, the aligned portion of the basecalled sequence is masked to avoid further alignment in subsequent iterations. This process is repeated until there are fewer than 15 non-masked bases or more than a set maximum number of iterations is exceeded, which is dependent on the length of the basecalled sequence.

Raw data alignment

Raw data was aligned to the expected signal based on the reference sequence of the read. We used a custom script that featured Tombo's API[66] to align the two sequences. Notably, the expected signal model is based solely on non-modified bases and therefore, we expected the raw signal corresponding to the 8-oxo-dG and surrounding bases to contain mis-alignments. Because our oligos contain random bases, for which we do not know their true reference, we used their aligned basecalls instead.

Base calling cross-validation

The oligo concatemers dataset (total of 2,460,266 reads) was first split into train and test sets. The test set consisted of 100 reads (chosen at random) per different oligo sequence (total of 22,000 reads). The train set consisted of all the reads (total 2,438,266 reads) not in the test set. To assess model performance during training, a portion of the train set was split into the validation set at random, and kept constant throughout the epochs (10% of the total training examples). The T2T dataset was first split into train and test sets. The train set consisted of 40,000 reads, chosen at random, from odd numbered chromosomes. The test set consisted of 10000 reads, chosen at random, from even numbered chromosomes. In the same manner as the oligo data, 10% of the total training examples was used as a validation set to assess model performance during training.

Raw signal normalization

A common approach to normalize the nanopore raw signal is to center and scale its values based on the set of measurements from the whole read ((1)).

$$x = \frac{x - median(x)}{mad(x)} \tag{1}$$

$$mad = median(|x - median(x)|) \tag{2}$$

We observed that both the T2T and oligo data had a consistent shift between expected and measured values (Supplementary Fig. 40, Supplementary Fig. 41). We attributed this larger shift in the oligo data to the consistently repetitive nature of the oligo sequences, which overall, do not have enough sequence variability to guarantee that the standard normalization approach would work as expected. To avoid normalization bias, which could already distinguish genomic from oligo data at the signal level, we performed a second normalization step. In this second normalization, we calculate what would be the optimal median and mad ((2)) values to minimize the distance between measured and expected values after alignment. We do this by first fitting a linear model using least squares regression between the aligned measured and expected values. We then re-scale the initially calculated med and mad values based on the fitted model. Using this second normalization step, we noted that the previously observed signal shift was corrected, and average 6-mer values match between genomic and oligo data (Supplementary Fig. 42). We applied this 2-step normalization approach on all our data both before training and inference.

Bonito fine-tuning

We fine-tuned a *Bonito* model to basecall 8-oxo-dG as G. We started from a pre-trained open source state provided by ONT in their public bonito repository (https://github.com/nanoporetech/bonito). The model was downloaded using the "bonito download –models" command. We fine-tuned the model for a total of 43,000 training steps using a batch size of 256. The model was fine-tuned using an unbalanced combination of oligo (25% of all training examples) and T2T data (75% of all training examples). We used the Adam[67] optimizer with a constant learning rate of 0.0005. Other parameters of the optimizer were: $\beta1 = 0.9$, $\beta2 = 0.999$, $\mu = 1e\text{-}8$ and $\lambda = 0.0$. We used a dropout of 0.1 in between CNN layers, and a dropout of 0.5 in between RNN layers (Supplementary Fig. 43).

Modification calling cross-validation

To avoid any potential data leakage between the fine-tuning of the base calling model and the training of the modification calling model, we used the same (based on read id) train, validation and test data splits in both steps. To train the model, the oligo dataset was used for both positive (8-oxo-dG) and negative (regular G) examples; while the T2T dataset was used only for negative (regular G) examples.

Remora training

We trained a *Remora* model to distinguish 8-oxo-dG from a regular G. We used a neural network architecture that first encodes the signal and sequence features via convolution, concatenates the encoded vectors, forwards the concatenated vector through a convolution layer and two recurrent layers, and a final linear layer for classification (Supplementary Fig. 9). We trained several models with balanced (via upsampling of the positive label samples, or downsampling of the negative label samples) and unbalanced datasets. We also trained models with different positive label weights: 100% 10% or 1%, with different features (*Methods,Remora feature-engineering*), and a model using metric learning (*Methods,Remora metric learning*). All models were trained as described in the following paragraph.

We trained the model as a classical classification problem using cross-entropy as loss function for a total of 235,000 training steps using a batch size of 256. We used the Adam[67] optimizer with a variable learning rate, which started at 0.00005 and increased linearly for 5000 training steps until 0.001, and then decreased using a cosine function until 0.00005. Other parameters of the optimizer were: $\beta1 = 0.9$, $\beta2 = 0.999$, $\mu = 1e\text{-}8$ and $\lambda = 0.0$. We used a dropout of 0.2 between all layers of the model.

Remora feature engineering

Traditional *Remora* models from ONT include the raw signal and base calls, centered around the base of interest as input features for the model. We explored the use of additional features, both at the signal and base call level. At the signal level, we used the following features: expected signal aligned to the measured signal based on the base calls from the fine-tuned *Bonito* model (8-oxo-dG aware), or from the pre-trained *Bonito* model (not 8-oxo-dG aware); the difference between the measured and expected signals (from both models); the difference between the aforementioned differences. At the sequence level, we used the following features: base calls from both the fine-tuned and pre-trained *Bonito* models, phred quality scores from both the fine-tuned and pre-trained *Bonito* models. Sequence-based features, are encoded using the same length dimension as the signal-based features, here 100 data points). Sequence-based features, are mapped into the 100 data points based on the signal-to-sequence mapping between the raw signal and the base calls. To ensure local feature information, features (both signal and sequence level) that were further than 3 bases, on either side, of the target G (based on the basecalls) were masked with zeros.

Remora metric learning

We trained a *Remora* model using Multi-Similarity loss ($\alpha = 2$, $\beta = 50$, base = 0.5) and Multi-Similarity miner ($\mu = 0.1$)[68] as implemented in PyTorch-Metric-learning. Triplets provided by the miner were filtered to only contain samples from the same 5-mer in an effort to force the model to compare the two labels in the same sequence context. Cosine similarity was calculated on the embedding vector output of the last LSTM layer (Supplementary Fig. 9) and used as a distance metric to calculate the loss. Triplet loss and cross-entropy loss were added together with equal weights before backpropagation.

Specificity Q-scoring notation

Due to the low prevalence of 8-oxo-dG, it is necessary to achieve a near-zero false positive rate (10^{-5}–10^{-7}). To make the annotation of these very small values easier, we convert these using the phred quality score (Q)[69].

Cell line sequencing

RPE1-hTERT-DAAO^{H2B} wild type (WT) and p53$^{-/-}$ cells were treated for 2 h with 20 mM L-Alanine or D-Alanine. Afterwards, genomic DNA was harvested using the DNeasy blood and tissue kit (Qiagen REF #69504) according to the manufacturer's protocol. Nanopore sequencing libraries were prepared and sequenced in the same manner as described for the oligo concatemers (*Methods, Library preparation, Nanopore sequencing*). For additional information regarding DAAO constructs and Illumina sequencing of these cell lines see van Soest et al.[47].

Illumina derived mutation profile

Clones were sequenced at 30x base coverage using an Illumina Novaseq 6000 or an Illumina Hiseq X10 sequencing machine. Sequencing reads from all samples were mapped to the human reference GRCh38 genome using the Burrows-Wheeler Aligner v0.7.17. Duplicate sequencing reads were marked using Sambamba MarkDup v0.6.8. Variants in the mapped data were called using GATK Haplotypecaller version 4.1.3.0 using default settings. Variants were filtered using GATK 4.1.3.0 using several filter settings Supplementary Data 5.To filter out mutations induced during sequencing, clonal expansion or library preparation, we filtered genomic variants using an in-house filtering pipeline, SMuRF v2.1.1. Briefly, the variant allele frequency (VAF) was calculated for each variant by pileup of all bases mapped at the mutation position. Variant data derived from cell clones were filtered for the following criteria: VAF ≥ 0.3, base coverage ≥ 10 and an MQ quality ≥ 60. To select only mutations occurring during in-vitro culture, variants present in the clonal parental cell line were removed. Recurrent mapping or sequencing artifacts were removed by filtering against a blacklist containing variants present in healthy bone marrow mesenchymal stromal cells.

8-oxo-dG mutational profile analysis

High confidence 8-oxo-dG calls (score > 0.95) from >Q40 5-mers were grouped based on the 3-mer of the opposite strand to facilitate comparison with the C>A profile as described in the COSMIC signatures database. Relative frequencies were calculated as the fraction of counts of each 3-mer given the total amount of counts. Counts per 3-mer were normalized to the amount of training 5-mers per 3-mer. Enrichment was calculated by subtracting the relative frequency of 8-oxo-dG calls from the relative frequency of 3-mers in the T2T reference genome.

8-oxo-dG genomic regions analysis

Genomic region annotations were downloaded for the CHM13v2 telomere-to-telomere (T2T) reference genome (https://github.com/marbl/CHM13). Base calls were aligned to the T2T reference genome using minimap2[70]. 8-oxo-dG counts (score > 0.95) were normalized per coverage as well as per 5-mer relative abundance per region. 8-oxo-dG counts on 5′-UTR, 3′-UTR, exon and intron regions were only considered if the molecule was found on the same DNA strand as the annotated region; in intergenic, satellite, centromeres and complex repetitive regions 8-oxo-dG counts were considered regardless of the DNA strand. 5'UTR was annotated as the upstream 5000 base pairs before the start of all annotated coding sequences. 3'UTR was annotated as the downstream 5000 base pairs of all annotated coding sequences. Intergenic regions were considered as all intervals that had no annotation.

8-oxo-dG related methylation analysis

All nanopore sequencing data was methylation called using the guppy tool and ONT's model *dna_r9.4.1_e8.1_modbases_5mc_5hmc_cg_hac.cfg*. Methylation calls were mapped to the T2T reference genome assembly, and then extracted using the modkit tool. Analysis was restricted to only 5-mers in the training set. We evaluated the methylation status within a 10 kilobase window centered around any guanine (canonical or 8-oxo-dG). Methylation prediction scores were then averaged per base pair position to calculate the average methylation status. Methylation scores within a 10 base window around 8-oxo-dG were masked out from the analysis because we assume that the methylation calling model is not aware of signal effects that might be caused by a close proximity 8-oxo-dG molecule.

8-oxo-dG telomere analysis

Reads whose mapping was primarily to the annotated T2T telomere regions were used for analysis. Based on the reference T2T genome, annotated telomere regions were further constrained to the last TTAGGG repetitive element.

Hardware and OS

All models where trained on Linux Rocky 8.10, using 8 CPU cores, 128 Gb of RAM, and 1 NVIDIA RTX6000 with 24 Gb of VRAM. The finetuning of a Bonito model took approximately 3 wall-clock hours. The training of a Remora model took approximately 7 wall-clock hours.

Reporting summary

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

Data availability

Nanopore sequencing data for the synthetic oligos and the cell lines has been uploaded to the European Nucleotide Archive under accession code PRJEB76712. Nanopore human data from the Telomere-to-Telomere can be found at: https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md[64]. Data availability details regarding Illumina sequencing data can be found at van Soest et al.[47]. Source data are provided with this paper.

Code availability

Source code for the 8-oxo-dG caller as a Python package can be found at https://github.com/marcpaga/esox. The specific version of the code associated with this publication is archived in Zenodo and is accessible via [https://doi.org/10.5281/zenodo.15268253][71]. The following Python v3.7 packages were used during the development of the 8-oxo-dG caller: fast-ctc-decode (v0.3.2), jupyterlab (v3.6.1), mappy (v2.22), matplotlib (v3.5.3), numba (v0.54.1), numpy (v1.18.5), ont-fast5-api (v4.1.1), ont-tombo (v1.5.1), pandas (v1.3.5), parasail (v1.3.3), polars (v0.18.3), pytorch (v1.12.1), pytorch-metric-learning (v2.3.0), scikit-learn (v1.0.2), seaborn (v0.12.2), tqdm (v4.65.0), logomaker (v0.8). The following tools were used for data processing and analysis: guppy (v6.3.8), minimap2 (v2.25), modkit (v0.2.0). Nextflow pipeline for Illumina raw data alignment alignment can be found at https://github.com/UMCUGenetics/NF-IAP. The pipeline for variant filtering can be found at https://github.com/ToolsVanBox/SMuRF.

References

1. Cadet, J. et al. Hydroxyl radicals and DNA base damage. *Mutat. Res.Fundam. Mol. Mech. Mutagen.* **424**, 9–21 (1999).
2. Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB* **17**, 1195–1214 (2003).
3. Hush, N. S. & Cheung, A. S. Ionization potentials and donor properties of nucleic acid bases and related compounds. *Chem. Phys. Lett.* **34**, 11–13 (1975).
4. Bergeron, F., Auvré, F., Radicella, J. P. & Ravanat, J.-L. HO• radicals induce an unexpected high proportion of tandem base lesions refractory to repair by DNA glycosylases. *Proc. Natl Acad. Sci. USA* **107**, 5528–5533 (2010).
5. Reisz, J. A., Bansal, N., Qian, J., Zhao, W. & Furdui, C. M. Effects of ionizing radiation on biological molecules—mechanisms of damage and emerging methods of detection. *Antioxid. Redox Signal.* **21**, 260–292 (2014).
6. Markkanen, E. Not breathing is not an option: How to deal with oxidative DNA damage. *DNA Repair* **59**, 82–105 (2017).
7. Klungland, A. et al. Accumulation of premutagenic DNA lesions in mice defective in removal of oxidative base damage. *Proc. Natl. Acad. Sci. USA* **96**, 13300–13305 (1999).
8. Slupska, M. M. et al. Cloning and sequencing a human homolog (hMYH) of the Escherichia coli mutY gene whose function is required for the repair of oxidative DNA damage. *J. Bacteriol.* **178**, 3885–3892 (1996).
9. Sekiguchi, M. MutT-related error avoidance mechanism for DNA synthesis. *Genes Cells* **1**, 139–145 (1996).
10. Kuchino, Y. et al. Misreading of DNA templates containing 8-hydroxydeoxyguanosine at the modified base and at adjacent residues. *Nature* **327**, 77–79 (1987).
11. Shibutani, S., Takeshita, M. & Grollman, A. P. Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. *Nature* **349**, 431–434 (1991).
12. Ohno, M. et al. 8-oxoguanine causes spontaneous de novo germline mutations in mice. *Sci. Rep.* **4**, 4689 (2014).
13. Nakabeppu, Y. et al. Mutagenesis and carcinogenesis caused by the oxidation of nucleic acids. *Biol. Chem.* **387**, 373–379 (2006).
14. van den Boogaard, M. L. et al. Defects in 8-oxo-guanine repair pathway cause high frequency of C > A substitutions in neuroblastoma. *Proc. Natl. Acad. Sci. USA* **118**, e2007898118 (2021).
15. Ohno, M. et al. Oxidative stress accelerates intestinal tumorigenesis by enhancing 8-oxoguanine-mediated mutagenesis in MUTYH-deficient mice. *Genome Res.* **34**, 47–56 (2024).
16. Bamford, S. et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* **91**, 355–358 (2004).
17. Pilati, C. et al. Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J. Pathol.* **242**, 10–15 (2017).
18. Viel, A. et al. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**, 39–49 (2017).
19. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
20. Korkmaz, K. S., Butuner, B. D. & Roggenbuck, D. Detection of 8-OHdG as a diagnostic biomarker. *J. Lab. Precis. Med.* **3**, 95 (2018).

21. Chiorcea-Paquim, A.-M. 8-oxoguanine and 8-oxodeoxyguanosine biomarkers of oxidative DNA damage: a review on HPLC–ECD determination. *Molecules* **27**, 1620 (2022).
22. Hahm, J. Y., Park, J., Jang, E.-S. & Chi, S. W. 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification. *Exp. Mol. Med.* **54**, 1626–1642 (2022).
23. Weitzman, S. A., Turk, P. W., Milkowski, D. H. & Kozlowski, K. Free radical adducts induce alterations in DNA cytosine methylation. *Proc. Natl. Acad. Sci. USA* **91**, 1261–1264 (1994).
24. Zhou, X. et al. OGG1 is essential in oxidative stress induced DNA demethylation. *Cell. Signal.* **28**, 1163–1171 (2016).
25. Weimann, A., Belling, D. & Poulsen, H. E. Quantification of 8-oxoguanine and guanine as the nucleobase, nucleoside and deoxynucleoside forms in human urine by high-performance liquid chromatography–electrospray tandem mass spectrometry. *Nucleic Acids Res.* **30**, e7 (2002).
26. Roszkowski, K., Jozwicki, W., Blaszczyk, P., Mucha-Malecka, A. & Siomek, A. Oxidative damage DNA: 8-oxoGua and 8-oxodG as molecular markers of cancer. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **17**, CR329–CR333 (2011).
27. Sato, T. et al. Increased plasma levels of 8-hydroxydeoxyguanosine are associated with development of colorectal tumors. *J. Clin. Biochem. Nutr.* **47**, 59–63 (2010).
28. Musarrat, J., Arezina-wilson, J. & Wani, A. A. Prognostic and aetiological relevance of 8-hydroxyguanosine in human breast carcinogenesis. *Eur. J. Cancer* **32**, 1209–1214 (1996).
29. Collins, A., Cadet, J., Epe, B. & Gedik, C. Problems in the measurement of 8-oxoguanine in human DNA. Report of a workshop, DNA oxidation, held in Aberdeen, UK, 19-21 January, 1997. *Carcinogenesis* **18**, 1833–1836 (1997).
30. Garratt, L. W. et al. Interpretation of urinary 8-oxo-7,8-dihydro-2′-deoxyguanosine is adversely affected by methodological inaccuracies when using a commercial ELISA. *Free Radic. Biol. Med.* **48**, 1460–1464 (2010).
31. Cadet, J., Douki, T. & Ravanat, J.-L. Measurement of oxidatively generated base damage in cellular DNA. *Mutat. Res.* **711**, 3–12 (2011).
32. Barregard, L. et al. Human and methodological sources of variability in the measurement of urinary 8-oxo-7,8-dihydro-2′-deoxyguanosine. *Antioxid. Redox Signal.* **18**, 2377–2391 (2013).
33. Wu, J., McKeague, M. & Sturla, S. J. Nucleotide-resolution genome-wide mapping of oxidative DNA damage by click-code-seq. *J. Am. Chem. Soc.* **140**, 9783–9787 (2018).
34. Poetsch, A. R., Boulton, S. J. & Luscombe, N. M. Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis. *Genome Biol.* **19**, 215 (2018).
35. Gorini, F. et al. The genomic landscape of 8-oxodG reveals enrichment at specific inherently fragile promoters. *Nucleic Acids Res.* **48**, 4309–4324 (2020).
36. Ding, Y., Fleming, A. M. & Burrows, C. J. Sequencing the mouse genome for the oxidatively modified base 8-oxo-7,8-dihydroguanine by OG-Seq. *J. Am. Chem. Soc.* **139**, 2569–2572 (2017).
37. An, J. et al. Genome-wide analysis of 8-oxo-7,8-dihydro-2′-deoxyguanosine at single-nucleotide resolution unveils reduced occurrence of oxidative damage at G-quadruplex sites. *Nucleic Acids Res.* **49**, 12252–12267 (2021).
38. Fang, Y. & Zou, P. Genome-wide mapping of oxidative DNA damage via engineering of 8-oxoguanine DNA glycosylase. *Biochemistry* **59**, 85–89 (2020).
39. Poetsch, A. R. The genomics of oxidative DNA damage, repair, and resulting mutagenesis. *Comput. Struct. Biotechnol. J.* **18**, 207–219 (2020).
40. Kong, Y., Mead, E. A. & Fang, G. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat. Rev. Genet.* **24**, 363–381 (2023).
41. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).
42. Clarke, J. et al. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270 (2009).
43. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
44. Rand, A. C. et al. Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).
45. White, L. K. & Hesselberth, J. R. Modification mapping by nanopore sequencing. *Front. Genet.* **13**, https://doi.org/10.3389/fgene.2022.1037134 (2022).
46. Yuen, Z. W.-S. et al. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat. Commun.* **12**, 3438 (2021).
47. van Soest, D. M. K. et al. Mitochondrial H2O2 release does not directly cause damage to chromosomal DNA. *Nat. Commun.* **15**, 2725 (2024).
48. Fouquerel, E. et al. Targeted and persistent 8-oxoguanine base damage at telomeres promotes telomere loss and crisis. *Mol. Cell* **75**, 117–130.e6 (2019).
49. Barnes, R. P. et al. Telomeric 8-oxo-guanine drives rapid premature senescence in the absence of telomere shortening. *Nat. Struct. Mol. Biol.* **29**, 639–652 (2022).
50. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
51. Turk, P. W., Laayoun, A., Smith, S. S. & Weitzman, S. A. DNA adduct 8-hydroxyl-2′-deoxyguanosine (8-hydroxyguanine) affects function of human DNA methyltransferase. *Carcinogenesis* **16**, 1253–1255 (1995).
52. Maltseva, D. V., Baykov, A. A., Jeltsch, A. & Gromova, E. S. Impact of 7,8-dihydro-8-oxoguanine on methylation of the CpG Site by Dnmt3a. *Biochemistry* **48**, 1361–1368 (2009).
53. Sood, A. J., Viner, C. & Hoffman, M. M. DNAmod: the DNA modification database. *J. Cheminform.* **11**, 30 (2019).
54. Liu, C. et al. Decoding the m6a epitranscriptomic landscape for biotechnological applications using a direct rna sequencing approach. *Nat. Commun.* **16**, 798 (2025).
55. Wang, Z. et al. Training data diversity enhances the basecalling of novel rna modification-induced nanopore sequencing readouts. *Nat. Commun.* **16**, 679 (2025).
56. Delgado-Tejedor, A. et al. Native rna nanopore sequencing reveals antibiotic-induced loss of rrna modifications in the a- and p-sites. *Nat. Commun.* **15**, 10054 (2024).
57. Motone, K. et al. Multi-pass, single-molecule nanopore reading of long protein strands. *Nature* **633**, 662–669 (2024).
58. Oikawa, S. & Kawanishi, S. Site-specific DNA damage at GGG sequence by oxidative stress may accelerate telomere shortening. *FEBS Lett.* **453**, 365–368 (1999).
59. Schmidt, T. T. et al. High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nat. Commun.* https://doi.org/10.1101/2023.11.28.569082v1 (2024).
60. Barnes, R. P., Fouquerel, E. & Opresko, P. L. The impact of oxidative DNA damage and stress on telomere homeostasis. *Mech. Ageing Dev.* **177**, 37–45 (2019).
61. Wang, K., Maayah, M., Sweasy, J. B. & Alnajjar, K. S. The role of cysteines in the structure and function of OGG1. *J. Biol. Chem.* **296**, 100093 (2021).
62. Perillo, B. et al. DNA oxidation as triggered by H3K9me2 demethylation drives estrogen-induced gene expression. *Science* **319**, 202–206 (2008).
63. Amente, S. et al. LSD1-mediated demethylation of histone H3 lysine 4 triggers Myc-induced transcription. *Oncogene* **29**, 3691–3702 (2010).

64. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
65. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinform*. **17**, 81 (2016).
66. Stoiber, M. et al. De novo Identification of DNA modifications enabled by genome-guided nanopore signal processing. https://doi.org/10.1101/094672v2 (2017).
67. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *Published as a conference paper at ICLR* (2015).
68. Wang, X., Han, X., Huang, W., Dong, D. & Scott, M. R. Multi-similarity loss with general pair weighting for deep metric learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019).
69. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res.* **8**, 186–194 (1998).
70. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
71. Pagés-Gallego, M. Esox code. https://doi.org/10.5281/zenodo.15268253 (2025).

Acknowledgements

We acknowledge the Utrecht Sequencing Facility (USEQ) for providing sequencing service and data. USEQ is subsidized by the University Medical Center Utrecht and The Netherlands X-omics Initiative (NWO project 184.034.019). We thank Edwin C. A. Stigter and Mehmet C. Gülersönmez for their help on mass-spectrometry validation of the synthetic oligos. This work was funded by Health Holland (No. LSHM19029). B.M.T.B., J.d.R., and R.v.B. are members of the Oncode Institute, which is partly funded by the Dutch Cancer Society (KWF Kankerbestrijding). T.B.D. was funded by the Dutch Cancer Society (KWF Kankerbestrijding, KWF grant 14798).

Author contributions

M.P.-G., T.B.D., B.M.T.B., and J.d.R. conceived the project. M.P.-G. developed the 8-oxo-dG caller. M.P.-G. and N.J.M.B. performed the experiments to ligate the synthetic oligos. D.M.K.v.S performed the cell lines experiments. J.P.K and M.J.v.R performed analysis on Illumina derived mutational signatures. A.M. provided feedback on oligo design. R.S and C.V provided feedback in algorithm implementations. R.v.B provided feedback on mutational analysis. M.P.-G. drafted the first version of the manuscript with guidance from J.d.R., T.B.D., and B.M.T.B. J.d.R, T.B.D., and B.M.T.B contributed to major parts of the manuscript and revised the manuscript. All authors read and approved of the final manuscript.

Competing interests

J.d. R. and A.M. are co-founders and directors of Cyclomics, a genomics company, they declare no competing interests. M.P-G., D.M.K.v.S., N.J.M.B., R.S., J.P.K., C.V., M.J.v.R., R.v.B., B.M.T.B., and T.B.D. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60391-3.

Correspondence and requests for materials should be addressed to Boudewijn M. T. Burgering, Tobias B. Dansen or Jeroen de Ridder.

Peer review information *Nature Communications* thanks Christopher Mason, Michael Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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