scientific reports

Check for updates

OPEN

Quantitative analysis of dose dependent DNA fragmentation in dry pBR322 plasmid using long read sequencing and Monte Carlo simulations

Pierre Beaudier[1,2,3], Sara A. Zein[1,3], Konstantinos Chatzipapas[1], Hoang Ngoc Tran[1], Guillaume Devès[1], Laurent Plawinski[1], Rémy Liénard[1], Denis Dupuy[2], Philippe Barberet[1], Sébastien Incerti[1], Franck Gobet[1] & Hervé Seznec[1]✉

Exposure to ionizing radiation can induce genetic aberrations via unrepaired DNA strand breaks. To investigate quantitatively the dose–effect relationship at the molecular level, we irradiated dry pBR322 plasmid DNA with 3 MeV protons and assessed fragmentation yields at different radiation doses using long-read sequencing from Oxford Nanopore Technologies. This technology applied to a reference DNA model revealed dose-dependent fragmentation, as evidenced by read length distributions, showing no discernible radiation sensitivity in specific genetic sequences. In addition, we propose a method for directly measuring the single-strand break (SSB) yield. Furthermore, through a comparative study with a collection of previous works on dry DNA irradiation, we show that the irradiation protocol leads to biases in the definition of ionizing sources. We support this scenario by discussing the size distributions of nanopore sequencing reads in the light of Geant4 and Geant4-DNA simulation toolkit predictions. We show that integrating long-read sequencing technologies with advanced Monte Carlo simulations paves a promising path toward advancing our comprehension and prediction of radiation-induced DNA fragmentation.

Keywords DNA damages, Ionizing radiation, Nanopore sequencing, Monte Carlo simulation

DNA strand breaks are a common form of biological damage that can occur naturally due to endogenous factors such as cell division (mitosis, meiosis), the presence of free radicals, as well as exogenous sources including ionizing radiation (IR), chemicals, and UV exposure[1]. While the general population is regularly exposed to low doses of IR from natural sources, certain professionals (astronauts, aircraft pilots, nuclear workers, etc.) may experience higher levels of exposure[2]. These individuals require ongoing monitoring to mitigate the risks of long-term DNA damage, which is a significant factor in the development of tumors[3]. IR can directly induce DNA damage through ionization by energy deposit on the DNA molecules[4–6] or indirectly through the interaction of reactive oxygen species generated by the radiolysis of water surrounding the DNA[6]. Understanding these damaging effects is pivotal in estimating radiation-induced health risks. This demands both experimental and numerical investigations to decipher, quantify, classify, and predict radiation-induced damage at the molecular level within the DNA macromolecules.

Accurately measuring DNA fragmentation yields raises many challenges. Factors such as the protection provided by a nucleus or cell membrane, potential activation of repair systems soon after irradiation, DNA extraction procedures that can cause damage to the molecules, irradiation protocols with well-defined irradiation sources at the DNA scale and the lack of reliable and precise fragmentation measurement methods contribute to the complexity. Agarose electrophoresis analysis (AGE) is the most common method, allowing simple and routine comparison of various irradiation conditions by quantifying changes in plasmid DNA conformations (supercoiled, open circular, linear)[7–9]. Nevertheless, AGE provides only a qualitative assessment of the fragmentation

[1]CNRS, LP2iB, UMR 5797, Univ. Bordeaux, 33170 Gradignan, France. [2]CNRS, INSERM, ARNA, UMR5320, U1212, Univ. Bordeaux, 33000 Bordeaux, France. [3]These authors contributed equally: Pierre Beaudier and Sara A. Zein. ✉email: herve.seznec@cnrs.fr

degree in the sample. It gives indirect access to the yield of single-strand break (SSB) through the commonly accepted but questionable assumption that an SSB causes a change in DNA conformation from a supercoiled to a relaxed form[10]. AGE also fails to localize the position of the damage site in DNA macromolecules. Furthermore, fragmented molecules by double-strand break (DSB), which are typically not visible or present as a smear, are not directly studied using this method. The gel measurement technique is also constrained by the maximum DNA size of approximately 20 kilobase pairs (kbp), which greatly limits the choice of DNA models that are compatible with this method. Although electrophoretic migration methods that surpass this size limit do exist, they are more expensive and more intricate, without necessarily offering better precision in analyzing fragmented molecules[11]. Therefore, employing alternative method for detecting damage sites and measuring size distributions is crucial for a better understanding of radiation-induced damage. For example, Xu et al.[12] used atomic force microscopy to assess plasmid fragmentation, enabling the direct observation of damage clustering in irradiated DNA.

The long-read sequencing method by Oxford Nanopore Technologies (ONT) (> 5 kbp contiguous sequence) presents a promising approach for investigating and measuring DNA fragment sizes. The unique feature of such sequencers lies in their ability to theoretically sequence reads of any size, owing to the nanopore technology employed[13–15]. Base-by-base sequencing of a damaged DNA molecule through a nanopore prematurely terminates the sequence on the first damaged base sugar-phosphate group resulting in smaller read sizes compared to native DNA. It is measured with a resolution of one base in position along the DNA strand. A comparison between the reads obtained from a native and a damaged DNA samples allows for the evaluation and quantification of DNA fragmentation yields as well as size distributions of the SSB- and DSB-induced fragments.

Alongside experimental methods, Monte Carlo simulations have emerged as powerful tools for predicting radiation-induced damage to biological systems. Various track structure codes estimate radiation-induced DNA damage, mainly focusing on interactions with liquid water, commonly used as biological material surrogate[16–19].

The Geant4-DNA simulation toolkit is a Monte Carlo track structure code specifically designed to simulate the interactions of IR with biological material, with a particular emphasis on DNA[20–23] and water. Ongoing efforts are oriented towards extending and enhancing the capabilities of this toolkit to simulate interactions at different stages, including physical, physico-chemical, and chemical processes. Furthermore, there is a focus on developing models that can accurately represent the three-dimensional (3D) geometries of DNA molecules as targets for irradiation in simulations. In a recent development of Geant4-DNA, the "molecularDNA" example application proposes three different DNA-scale detailed geometries modeling prototypes of simple linear DNA inside cylinders, an *Escherichia coli* (*E. coli*) and a human genomic DNA content[24]. In this model, the molecular double helix is represented as a continuous sequence of spheres and ellipsoids, representing the four nucleobases and the sugar-phosphate group. By considering the physical and chemical stages of IR interactions with these DNA geometries, early DNA lesions can be estimated by introducing a few numerical parameters such as a damage energy threshold or interaction length ranges that have to be optimized to fit with experimental data. The development of these simulation codes faces a major challenge, given the difficulty of producing experimental validation data, and the ONT sequencing method represents an interesting opportunity in providing such data.

This article details the results obtained by long-read sequencing method following irradiation by 3 MeV proton with doses ranging from 0.5 to 5 kGy of pBR322 plasmids. This plasmid, with a known sequence of 4361 base pairs (bp) occurring in supercoiled and open circular forms, offers simpler geometry than cellular DNA, making them ideal for realistic modeling in Monte Carlo simulations. Integrating these experimental and numerical technologies, paves a promising path toward advancing our comprehension and prediction of radiation-induced fragmentation.

Results

Evaluation of the long-read sequencing quality and yields for quantitative analysis of radiation-induced plasmid DNA fragmentation using Flongle flow cells

We used Flongle flow cells (ONT) to analyze DNA plasmid samples, for its cost-effectiveness and its potential for widespread technological application[25]. Initially, post-irradiation, DNA plasmids undergo linearization using the *BamHI* enzyme to allow for adapter ligation. The library preparation is pivotal for subsequent nanopore sequencing work.

Here DNA fragments are left unrepaired, streamlining the protocol. A DNA-helicase complex forms the connector, which, upon approaching the nanopore, binds to a single-stranded leader at the end of the double-stranded DNA template. Then, the complex unzips the double strand and threads a single strand through the nanopore within an artificial membrane in the sequencing flow cell (Fig. 1a). A controlled electrical current is applied across the membrane, compelling the negatively charged DNA strands to traverse the nanopores. When a DNA fragment enters in a nanopore, it induces a continuous series of changes in the current, which is constantly monitored by an integrated electronics chip within the flow cell. A base-calling algorithm translates these current variations back into the original DNA sequence. As a result, intact DNA strands are entirely sequenced. For molecules with either single strand- or double strand-breaks, the reading process halts at the first damage site (Fig. 1b) and the fragment size between this site and the *BamHI* site is recorded, enabling the visualization of variations in both length and sequence. Thus, we used this technology to evaluate the radiation-induced DNA fragmentation on dry pBR322 plasmids[26]. The same batch of DNA was also analyzed without enzymatic restriction using electrophoresis gel analysis (AGE) (Supplementary Fig. S1a).

To establish a baseline and ensure that any observed changes in subsequent irradiated samples were indeed due to radiation-induced effects, we evaluated the quality and sequencing yield of control (non-irradiated) plasmid batches. The plasmid size distribution of the control samples exhibited a bell-shaped curve centered around the expected whole-genome read size of 4361 bp (Supplementary Fig. S1b). Two factors influenced read size deviation from expected whole-genome size. First, the relatively high error rate of the sequencer used

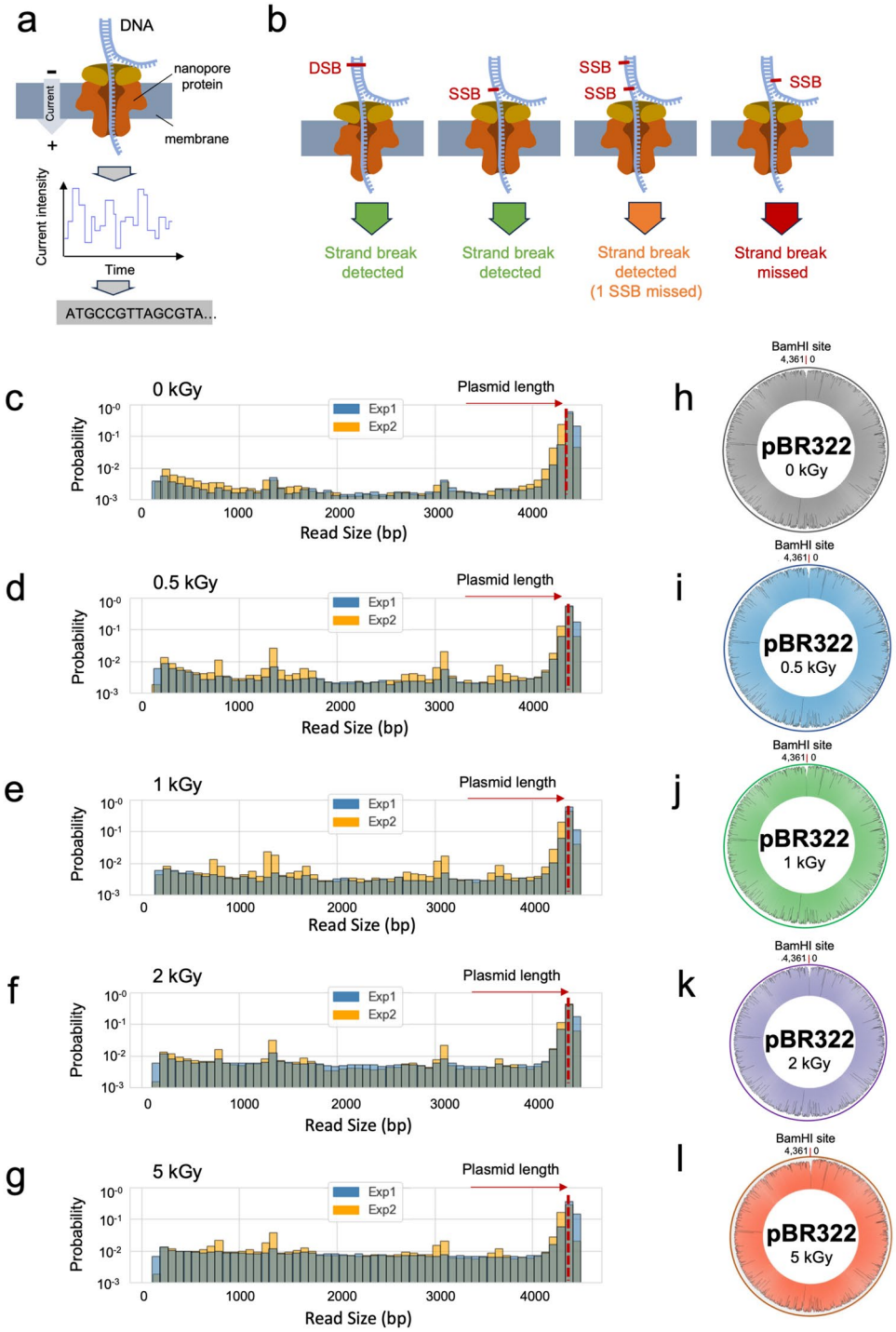

Figure 1. Description of the Oxford Nanopore sequencing technologies. (**a**) Principle of Oxford Nanopore Sequencer. (**b**) Scheme of the different scenarios that can arise after DNA irradiation inducing single strand (SSB) and double strand breaks (DSB). (**c**–**g**) Read length probability distribution of reads from both sequencing runs for all experimental conditions at 0, 0.5, 1, 2 and 5 kGy. Note that y-scale is chosen to be the same as control. The lengths of reads mapping on the pBR322 genome were obtained from alignment files. Minor differences in the fragmentation rate were also observed between the two experimental runs (Exp1 and 2), particularly in the read size distribution. (**h**–**l**) Circular visualization of reads alignment coverage *per* irradiation dose ranging from 0, 0.5, 1, 2 and 5 kGy.

(~ 5% *per* base with the specific library preparation kit and flow cell combination) resulted in the introduction of unintended insertions or deletions of bases in the read sequences compared to the genome sequence, each of these sequencing errors accounting for approximatively 1/3 of the error rate thus a total error rate affecting read length of around 3.3%[27]. On average, deletions occurred more frequently than insertions (Supplementary Fig. S2). Second, soft-clipped sequences, that do not map to the genome, were identified at both ends of the sequenced reads. These soft-clip sequences had an average size of 60 bases. We hypothesized that these soft-clip sequences were the result of attaching free-floating bases to the sticky ends generated by the *BamHI* digestion process (Supplementary Fig. S2). The combination of these two effects led to variations in the read size resulting in an average of an additional 30 bases *per* read, which were method-related rather than reflecting the true fragmentation of the plasmid DNA. This variation explained the presence of reads measuring over 4361 bp and the observed bell-shaped curve in the size distribution (Supplementary Fig. S1b–f). To distinguish between whole-genome reads and fragmented reads, an acceptability threshold based on the size of the reads was established. After analyzing the sequence alignments, a threshold of ± 3% was chosen, retaining the majority of the reads within the bell-shaped curve, and accounting for size differences that could be attributed to alignment variations. Reads larger than 4492 bp were considered "artefact" reads and excluded from further analysis, while reads smaller than 4230 bp were considered fragmented reads.

Approximately 81% of the reads in the control samples were categorized as whole-genome reads, indicating an inherent fragmentation rate of about 19% due to the sequencing protocol itself. This consistent unwanted fragmentation rate was observed across control samples and previous sequencing runs using the same DNA model and library preparation kit/flow cell combination. In our study, DNA samples were divided equally for sequencing and AGE analysis. AGE showed no fragmentation in control samples, but sequencing revealed significant fragmentation, with only 81% intact fragments. We hypothesize different scenarios: (1) control DNA was already fragmented, but AGE lacks the sensitivity to detect it; (2) DNA fragmentation occurred during the experimental steps required for long-read sequencing; (3) the whole sequencing process itself caused fragmentation. We postulate that it can be expected that the magnitude of this protocol-based fragmentation will be similar among all conditions and thus, we decided to correct our results for the intrinsic fragmentation rate (19%, 81% of fragments being intact) to account for background fragmentation from sample preparation and sequencing. This method of threshold determination was thus applied across all experimental conditions, enabling the comparison between different irradiation doses. In addition, long-read sequencing offers coverage data *per* base across different experimental conditions by aligning sequences. The coverage *per* base can be used to determine the genomic positions corresponding to the 5′ and 3′ ends of the reads as well as the eventual presence of preferential fragmentation sites on the pBR322 sequence, which would be identifiable by significant decreases in coverage on specific bases. In the control sample, we found that almost all reads have either their 5′ and 3′ ends on the *BamHI* site for complete molecules, or only one end for fragmented molecules, as expected given the otherwise lack of blunt end for adapter ligation without the *BamHI* digestion. The visualization of the genomic coverage on the control sample also allows us to observe the presence of a few low coverage bases, despite the absence of induced fragmentation (Fig. 1c–g). The most likely explanation for these bases would be the presence of mutations on our DNA sample compared with the reference sequence used (GenBank J01749.1) since they were consistent among all samples.

Quantitative analysis of radiation-induced plasmid DNA fragmentation revealed SSB and DSB yield discrepancies between long-read sequencing and agarose electrophoresis analysis (AGE)

Quantitative analysis of radiation-induced DNA fragmentation using long-read Sequencing. Two sets of air-dried pBR322 plasmid DNA were irradiated with 3 MeV protons at various doses (0, 0.5, 1, 2, and 5 kGy) using the experimental set-up described in Bourret et al.[28] (Supplementary Fig. S1a). Post-irradiation, DNA samples were incubated with *BamHI* enzymes for 4 h and 18 h called Exp1 and Exp2 in the size distributions reported in Fig. 1c–g, respectively. An average of 10^5 sequencing reads were acquired *per* individual sequencing run (Fig. 1 and Supplementary Fig. S1). Sequencing measurements revealed a progressive rise in fragmented reads with increasing irradiation doses in both experimental runs. The unfragmented reads decreased from approximately 81% in the control to around 60% at 5 kGy (Fig. 1c,g). Visualizing the 5 kGy sample showed an overall decrease in normalized base coverage (Fig. 1g–l), as anticipated due to the smaller size of reads overall in this condition. No preferential fragmentation sites were observed compared to the control; low-coverage bases matched those in the control sample (Fig. 1c–h). The majority of read ends consistently aligned with the *BamHI* digestion site, similar to the control, suggesting a negligible probability of radiation-induced DSBs.

Minor differences in the read size distribution were observed between the two experimental runs (Fig. 1c–h). Strand fragments varied in size from the minimum read size of 200 bp up to 4230 bp, without any privileged value standing out in the run with the lowest duration of *BamHI* enzyme incubation (Exp1). Some specific read sizes are nevertheless observed in the run named Exp2, such as for example 1291, and 3615 bp. A longer digestion time leads to unintended activity of the enzyme on pseudo-restriction sites that differ by one base from the expected *BamHI* site. This resulted in some reads corresponding to the sequence between two digestion sites, explaining the presence of unexpected peaks at some specific read sizes, which were then identified as artifacts related to the sequencing preparation protocol.

Comparison of radiation-induced SSB and DSB yields from long-read sequencing and AGE. The yield of intact strands in the read size distribution (size of 4361 ± 131 bp) is a read out of the probability that no sugar-phosphate group is broken along the chain. This probability is in turn related to the yield of single strand breaks (SSB) and double strand breaks (DSB) in the irradiation, which can then be calculated by fitting the experimental yields of intact strand as a function of the proton dose with the adequate law. To extract these SSB and DSB damage yields, we first corrected the raw probability of intact strands from the response function of the sequencer for

DNA strands at a size of 4361 ± 131 bp. This response is the measured frequency of 81% of intact whole plasmid strand in the control samples. The corrected probability of intact strand is plotted in Fig. 2c as a function of the radiation dose. As expected, the probability of not breaking an intact (i.e., non-fragmented) strand decreases with dose, reaching a value of around 60% at 5 kGy.

Then, we determined the function that best describes the data. We observed that there is no specific preference for irradiation damage in any base pair (Fig. 1h–l). This suggests an equal probability of breaking the DNA strand between consecutive bases, thus the probability of breaking n bonds in a strand is described by a binomial distribution. The number of bases in one strand N = 4361 is very large and we can assume that the probability of breaking a bond in the strand at a given dose is very small.

In this context, the binomial distribution becomes a Poisson distribution and the probability of breaking n bonds in one strand is given by:

$$P(n) = \frac{\overline{n}^{n}}{n!} e^{-\overline{n}} \tag{1}$$

with $\overline{n} = \beta_{damage}{}^{*}D^{*}N$ is the mean number of breaks in one strand of N base pairs, D the dose and $\beta_{damage} = \beta_S + \beta_D$ the total yield of SSB and DSB in one strand *per* unit of dose and base pair. The percentage of full-size strands in an irradiated sample at a given dose is therefore a signature of the probability of no fragmentation. Specifically, the probability of a DNA strand integrity is determined by Eq. (2):

$$P(D,\ n = 0) = e^{-\beta_{damage}\, N\, D} \tag{2}$$

From Eq. (2), we extracted the fit of the experimental data with β_{damage} as a free parameter (curve labelled Poisson 1 in Fig. 2c). This total yield of SSB and DSB is next multiplied by 2, for simplification purposes DSBs are not taken into consideration due to their low probabilities of occurrence, to account for the fact that only one DNA strand is sequenced and to compare with the yields extracted from AGE data, which are expressed *per* double stranded DNA. We achieved strong alignment between the analytical law and experimental measurements with

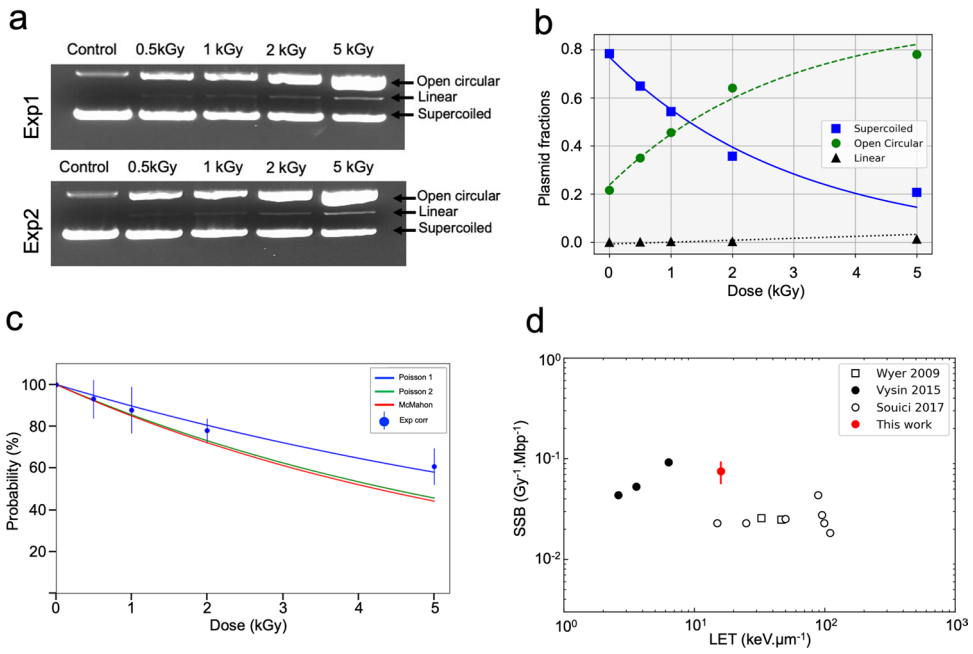

Figure 2. Quantitative analysis of radiation-induced plasmid DNA fragmentation. (**a**) Effect of 3 MeV proton irradiation on two sets of air-dried pBR322 plasmid DNA at 0, 0.5, 1, 2, and 5 kGy. Band sizes separated on 1% (w/v) agarose gel and visualized by ethidium bromide staining gels from Exp1 and Exp2. (**b**) Ratios of plasmid conformations observed on the agarose gel fitted on the McMahon statistical model. (**c**) Intact plasmid strand yields as a function of the proton dose on DNA. The average undamaged strand yields of both sequencing runs are calculated and corrected for the sequencer's response function and are fitted according to the Poisson regression (Poisson 1). The expected intact strand yields are calculated from gel electrophoresis data fitted according to Poisson (Poisson 2) and McMahon's models. (**d**) Comparison of SSB yields with published data. The SSB yields calculated using McMahon's fitting model are compared with previously published data for irradiation of dry plasmid with hydrogen and proton beams. Experimental irradiation conditions: sealed symbols represent irradiation performed in air, while open symbols represent irradiation conducted in vacuum. Our AGE results (red) are in the expected range of damage according to the linear energy transfer of the proton beams (~ 16 keV μm^{-1} in DNA).

$$\beta_{\text{damage}} = (50 \pm 12) \times 10^{-3}\ \text{Mbp}^{-1}\ \text{Gy}^{-1}\ \ (\textit{per}\ \text{double stranded DNA}).$$

Since the sequencing protocol does not distinguish between SSB and DSB as illustrated in Fig. 1b, agarose electrophoresis analysis (AGE) was carried out on the same samples but without *BamHI* restriction (Fig. 2a and Supplementary Fig. S3). AGE is based on changes in plasmid conformation following DNA damage as plasmid DNA can exist in three forms: supercoiled, open circular, and linear, each represented by distinct bands on an agarose gel (Fig. 2a and Supplementary Fig. S1a). In our control samples, pBR322 DNA is predominantly in the supercoiled conformation (~ 70%) with a smaller proportion in the open circular conformation (~ 30%). Briefly, a native plasmid is mainly supercoiled and it is generally considered that when an SSB is induced, the supercoiled plasmid converts to its open circular form, whereas when a single or multiple DSBs are produced, it converts to the full linear form or fragments. Due to variations in ethidium bromide binding efficiency based on plasmid conformation, we use a correction factor of 1.4 to account for the superhelical density when quantifying the supercoiled form of closed circular DNA[29] (Fig. 2b). In our experiments, at a dose of 5 kGy, there is a significant decrease in the proportion of supercoiled DNA (from ~ 70 to ~ 16%) accompanied by an increase in the proportion of open circular DNA (from ~ 30 to ~ 83%) (Fig. 2a). Linear DNA molecules were observed in very low proportions even at the highest doses (~ 1% at 5 kGy) (Fig. 2a,b).

These electrophoretic gel migration experiments provide information on the content of irradiated samples, notably with regards to the yields of SSB and DSB *per* irradiation dose using McMahon's fitting model[10], which is based on two important assumptions: (1) the change from supercoiled to open circular isoforms is systematically ensured if at least one SSB occurs on one strand, (2) two independent SSBs in the two opposite DNA strands spaced less than 10 bp apart are equivalent to one DSB. The relative proportions of these forms were converted into SSB and DSB yields by fitting them to the following Eqs. (3), (4) and (5):

$$S = S_0 e^{-(\beta_S+\beta_D)ND} \tag{3}$$

$$C = e^{-\beta_D DN}\left[C_0 e^{-\frac{1}{2}\beta_S^2 \rho D^2 N^2} + S_0\left(e^{-\frac{1}{2}\beta_S^2 \rho D^2 N^2} - e^{-\beta_S DN}\right)\right] \tag{4}$$

$$L = 1 - (C_0 + S_0) e^{-\left(\beta_D DN + \frac{1}{2}\beta_S^2 \rho D^2 N^2\right)} \tag{5}$$

with S, C, and L the relative proportions of supercoiled, open circular, and linear DNA, respectively. The parameter $\rho = \frac{10}{4{,}361} = 2.3 \times 10^{-3}$ is the ratio between the maximum distance between SSBs on opposite strands that could result in a DSB considered as 10 bp and the length of the plasmid.

By doing so, we obtained: $\beta_S = (75 \pm 19) \times 10^{-3}\ \text{Mbp}^{-1}\ \text{Gy}^{-1}$ and $\beta_D = (1.7 \pm 4.6) \times 10^{-3}\ \text{Mbp}^{-1}\ \text{Gy}^{-1}$ (Table 1).

The linearization of plasmids is primarily caused by DSB, which occur at a rare frequency even at high doses in our experiment and we can neglect DSB events induced by two independent SSB in less than 10 base pairs as $\frac{1}{2}\beta_S^2 \rho D^2 N \ll \beta_D D$. Thus, we revisited the AGE data on the basis of a second Poisson model in order to reduce uncertainties in the fitting parameters by considering $\rho = 0$. Under this simplification, the proportion of the linear form is now a function of the DSB yield β_D only, thus:

$$L = 1 - e^{-\beta_D N D} \tag{6}$$

whereas the proportion of supercoiled form still depends on the SSB and DSB yields as shown in Eq. (3). Considering that the linear form is in the minority compared to the supercoiled and circular forms, the proportion of this latter can then be approximated by:

Fitting model	SSB (β_S) – DSB (β_D) yields	10^{-3} Mbp^{-1} Gy^{-1}
Electrophoresis, McMahon	β_D	1.7 ± 4.6
	β_S	75 ± 19
Electrophoresis, Poisson regression	β_D	0.54 ± 0.06
	β_S	72 ± 10
Sequencing, Poisson regression	$\beta_D + \beta_S$	50 ± 12

Table 1. SSB and DSB yields calculated from fitting experimental data of gel electrophoresis and sequencing. Relative proportions of supercoiled, open circular and linear electrophoretic conformations of plasmids at different doses are fitted according to the McMahon's model and the SSB (β_S) and DSB (β_D) yields were deduced based on Eqs. (3), (4) and (5). The same electrophoretic isoform proportions were also fitted according to the Poisson regression using the Eqs. (6) and (7) to calculate the SSB and DSB yields. The total strand breaks yield without distinction between SSB or DSB was calculated according to the Poisson regression based on the sequencing data using Eq. (2). The results are obtained with a correction factor of 1.4 associated to the ethidium bromide staining of the supercoiled intensity.

$$C = C_0 + S_0\left(1 - e^{-(\beta_D+\beta_S)N\,D}\right) \quad (7)$$

The relative proportions of the pBR322 DNA forms were fitted to these equations and gave a second set of yield values with much lower error bars using the same electrophoresis data (corrected with the factor of 1.4): $\beta_S = (72 \pm 10) \times 10^{-3}$ Mbp^{-1} Gy^{-1} and $\beta_D = (0.54 \pm 0.06) \times 10^{-3}$ Mbp^{-1} Gy^{-1} (Table1).

All yield values extracted from our experiments are reported in Table 1. In addition, values obtain with the different correction factor found in the literature have been calculated (Supplementary Fig. S4 and Table S1). Regardless of the AGE data adjustment method, the DSB yield remains much lower than the SSB yield. Thus, the total damage yield measured through sequencing mainly represents the SSB yield caused by 3 MeV protons. The expected probabilities of intact strands are slightly different when we calculate them with the different values of SSB and DSB yields reported in Table 1. While the Poisson regression of sequencing data (curve labelled Poisson 1 in Fig. 2c) aligns well with corrected probabilities of intact strands, the expected curves (labelled Poisson 2 and McMahon in Fig. 2c) weakly derived from electrophoresis' SSB and DSB yield values. These differences are probably due to a lack of precision in the correction factor to be applied in the supercoiled form (Supplementary Fig. S4 and Table S1).

Finally, as shown by comparisons with previous experiments irradiating dry plasmids with hydrogen[30] and protons beams[29,31,32], our results fall within the range of expected damage based on linear proton beam energy transfer (~ 16 keV µm^{-1} in DNA) (Fig. 2d). All these measurements were carried out using a post-irradiation electrophoresis technique, and SSB yields were extracted using the McMahon method. Nonetheless, there is a wide dispersion with a factor of up to 4 between the various results on SSB rates for experimental protocols that are, in a first approach, fairly similar. In particular, around the LET value of our study, Vysin et al. found SSB yields of $(92.9 \pm 7.1) \times 10^{-3}$ Mbp^{-1} Gy^{-1} at LET = 6.4 keV µm^{-1} in DNA with 10 MeV protons[31], while Souici et al. measured 22.9×10^{-3} SSB Mbp^{-1} Gy^{-1} with 3 MeV protons (LET = 15 keV µm^{-1} in DNA)[32]. From this point of view, our results are closer to those of Vysin et al.[31]. Several possible reasons could be behind these disparities which support the interest in performing sequencing measurements in the future and will be presented in the "Discussion" section. The main reason likely lies in the ionizing irradiation conditions at the DNA level that influences the type of interacting particles depending on the experimental protocols which would be better understood with Monte Carlo simulations.

From long-read sequencing of radiation-induced plasmid DNA fragmentation to Monte Carlo simulations

As reported in Fig. 2d, we identify two types of dry DNA irradiation protocols from all these works. Some authors (*open symbols* in Fig. 2d) irradiate DNA deposited on a plastic or glass support under vacuum[30,32], while others (*sealed symbols* in Fig. 2d) perform irradiations in air with a sample sealed with a micrometric film of plastic, as in our study (Fig. 3a)[29,31,33].

Data reported in Fig. 2d clearly shows a systematic difference depending on the irradiation protocol: irradiations carried out in vacuum systematically show lower SSB yields. The distribution of proton-induced secondary particles within DNA could differ according to the irradiation protocol. This point could be estimated and tested using Monte Carlo simulations (Geant4 and Geant4DNA).

Thus, Geant4 simulations were employed to calculate the energy distributions of incident 3 MeV protons and secondary electrons generated in experiments performed under air or vacuum conditions. Geant4 simulations demonstrate that with 3 MeV incident protons, the distribution of proton-induced secondary particles within DNA varies based on the different irradiation protocols. In our experimental condition (air), dry DNA is placed on a 4-µm polypropylene foil, and the protons must pass through this material, interacting with it (Fig. 3a). The simulations reveal significant secondary electron production exiting the polypropylene foil. Different simulation configurations using the Livermore electromagnetic physics constructor with 10^7 primary protons were applied to obtain the corresponding energy spectra of incident protons and secondary emitted electrons (Supplementary Fig. S5). In our conditions, 10% of protons generate scattered electrons with expected energies up to 6 keV emitted from the DNA support as protons traverse it (Fig. 3b). In vacuum experiment conditions, where protons interact with dry DNA before impinging on the foil, calculations show that about only 1% of protons generate an electron that backscatters towards the DNA. The presence of a secondary source of electrons under air conditions raises the question of its effect on DNA fragmentation yield.

We supported this scenario by calculating the contribution of incident protons to the expected DNA damage yield in our experimental conditions. We compared the measured data of long-read sequencing of radiation-induced DNA fragmentation with the results of Monte Carlo Geant4-DNA toolkit simulations based on the "*molecularDNA*" example. Plasmids are relatively short circular DNA molecules that naturally occur in supercoiled forms. Unlike cellular DNA, which is intricately looped around histones to form complex chromosomes, plasmids have a simpler geometry, making them an ideal choice for realistic modeling in Monte Carlo simulations. In addition, we introduced an approach to incorporate a total of 10,142 circular plasmid copies to closely replicate the experimental concentration, resulting in significantly enhanced statistical accuracy and a closer simulation with experimental realities (Supplementary Fig. S6a and S6b). In water, simulating radiation-induced DNA damage requires the determination of two important parameters, the minimum energy deposition in water required to produce damage, or energy deposit threshold (t) and the direct interaction range (r) corresponding to the maximum distance from the center of the DNA backbone sensitive to the irradiation that might induce DNA damage (Supplementary Fig. S5d). Here, we considered the energy deposit threshold (t) proposed by Nikjoo et al.[34], which stipulates a minimum of 17.5 eV for inducing strand breaks. Energy deposition in the hydration shells surrounding DNA destabilizes the double helix structure and leads to direct damage. A strand break induction is expected when energy is deposited within a supposed distance (r) of about 5.5 Å enveloping

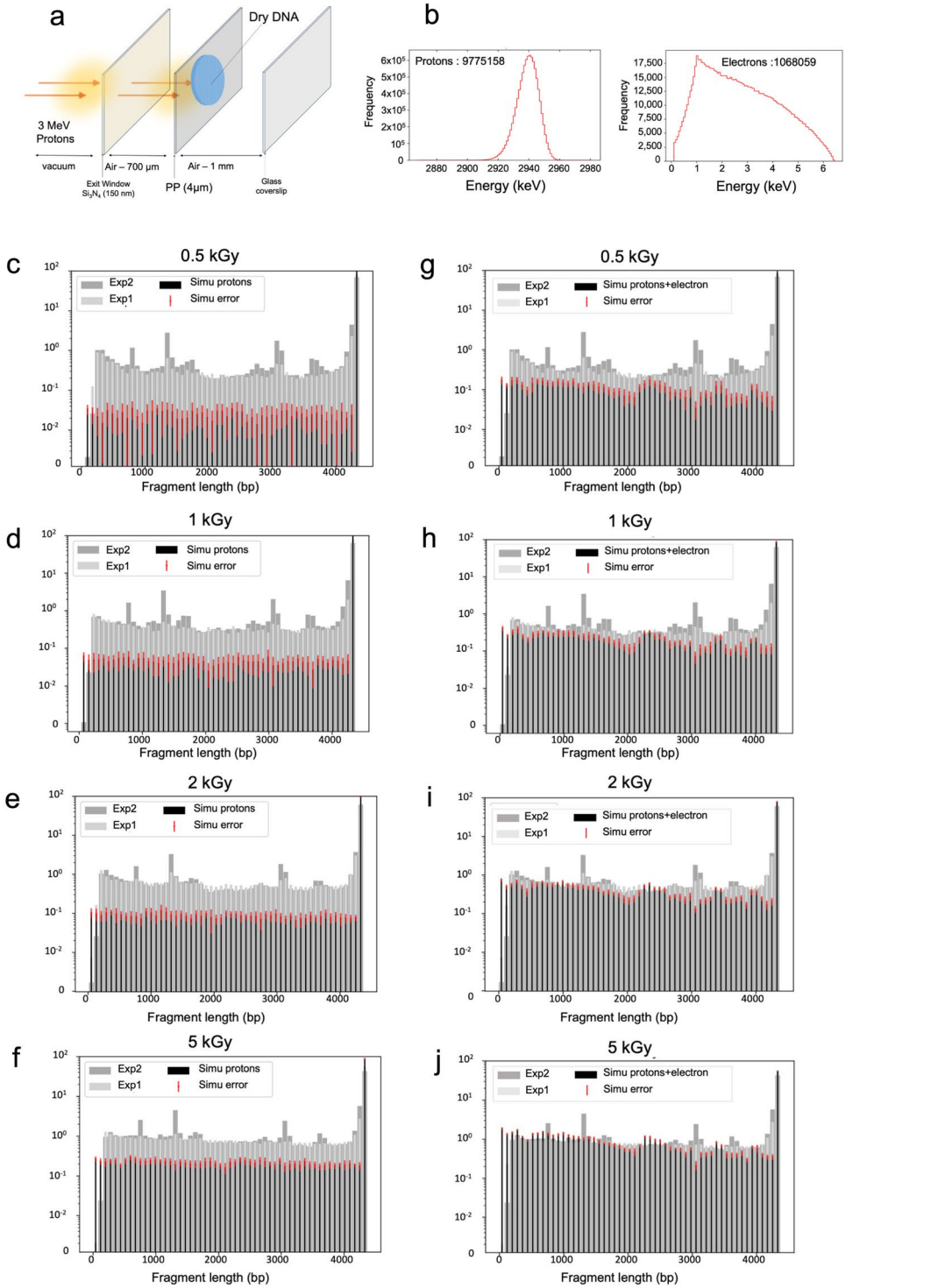

Figure 3. From experience to Monte Carlo simulation. (**a**) Overview of the sample configuration. DNA drop was spotted on a 4 µm thick polypropylene foil (PP). The back-side of the cell dish is closed in air using a glass coverslip. (**b**) Geant4 simulation of the proton and secondary electron energy distributions produced in the experimental conditions. (**c**–**f**) Comparison of fragment length percentage yield between both sequencing runs and the Geant4-DNA simulations for irradiation doses of (**c**) 0.5, (**d**) 1, (**e**) 2 and (**f**) 5 kGy, respectively without the secondary electron contribution. (**g**–**j**) Comparison of fragment length percentage yield between both sequencing runs and the Geant4-DNA simulations for irradiation doses of (**g**) 0.5, (**h**) 1, (**i**) 2 and (**j**) 5 kGy, respectively with the secondary electron contribution. Experimental data from Exp1 and Exp2 are depicted in grey. Simulation results depicted in black were obtained with "molecularDNA" example with 17.5 eV energy threshold (t), and 5.5 Å hydration shell (r). A number of 10 simulations were run for each dose and error bars shown in red were calculated as standard deviation for simulation fragments. Distribution histograms have a bin width of 79 bp.

the DNA according to quantum chemistry calculations[35]. This distance of about 5.5 Å, corresponds to the thickness of two layers of water molecules around nucleotides and sugar-phosphate groups. Additionally, we compared the simulated yields of radiation-induced DNA damage produced by Geant4-DNA with those obtained through sequencing. Calculations with a 17.5 eV threshold yield a SSB rate of $50 \pm 12 \times 10^{-3}$ Mbp^{-1} Gy^{-1} (from sequencing) considering a 5.5 Å layer of water. In the '5–37.5 eV' scenario, the same SSB rate is obtained with a 4 Å layer (Supplementary Fig. S7). Then, according to our previous Geant4 simulations and to only consider the contribution of protons to DNA damage, in a first set of simulation, we kill all secondary electrons produced in the medium during simulations. In a second set of simulation, we calculated DNA damage yields considering the protons and the secondary electrons.

The determination of the fragment lengths from Monte Carlo simulations is performed in a similar manner to long-read sequencing. A random strand is chosen, starting from a specific point in the molecule (such as the *BamHI* site), and read until the first SSB or DSB is encountered, or until the end of the molecule if it remains unfragmented. The size distribution of fragments is then compared to those obtained through long-read sequencing with the contribution of proton only (Fig. 3c–f) and with proton and secondary electrons (Fig. 3g,h). Fragment length determination from Monte Carlo simulations mirrors long-read sequencing. As previously observed, simulations produce a dose-dependent fragmentation rate.

The simulation consistently yields lower fragment rates compared to the sequencing experiment, with the most significant disparity observed at the 0.5 kGy dose (Fig. 3c,g). At this dose, a substantial portion of the fragmented reads is due to the sequencer response rather than being induced by IR. At higher doses, the radiation-induced component has a more significant impact on the experimental data, enabling comparison with the simulation results. At 5 kGy, the calculated fragmentation yields considering only protons are 4 to 5 times lower than the measured ones (Fig. 3f). This ratio is very close to the experimental DNA damage yields obtained under air and vacuum conditions, suggesting that the contribution of the secondary electron source as a consequence of the higher DNA damage yield under air conditions is plausible. In simulations keeping all secondary electrons produced in the water medium surrounding the DNA, the percentage of fragments obtained is higher and in agreement with the experiment data (Fig. 3g–j). Although, this simulation does not reproduce the dry conditions of the experiments in air, it shows a clear and important contribution of the secondary electrons. Whereas a SSB yields of $12 \pm 1 \times 10^{-3}$ Mbp^{-1} Gy^{-1} is reached considering only protons, a calculated yield of $55 \pm 2 \times 10^{-3}$ Mbp^{-1} Gy^{-1} is obtained taking into account also the secondary electrons.

Discussion

We proposed to use the nanopore sequencing technology with the aim of producing a reliable method for comparing DNA fragmentation across different experimental conditions. It offers the capacity to sequence entire DNA molecules without the usual prior digestion and segmentation from other sequencing techniques. Therefore, it would be theoretically possible to carry out sequencing runs in which any radiation-induced fragmentation observed arises solely from the initial state of the DNA molecule and not from the library preparation protocol[13]. It is with this ability in mind that we used this technology to quantitatively measure the radio-induced plasmid DNA fragmentation at different doses of irradiation, compared the results obtained with routine electrophoresis analysis and from Monte Carlo simulation (Geant4-DNA). The analysis of DNA fragments extremities using this technology also giving us information on the genetic position of the reads end allows for the identification of potential preferred sites of breakage.

A series of experiments involved the sequencing of air-dried DNA molecules subjected to irradiation doses ranging from 0.5 to 5 kGy using a 3 MeV proton beam, enabling the quantification of radiation-induced fragmentation. The size distribution of the sequenced reads and the percentage of intact strands served as key indicators of fragmentation. Although this method comes with inherent limitations and biases, including its inability to differentiate between SSB and DSB, the presence of undesired technology- and protocol-linked fragmentation, and variations in complete molecule sizes due to the error rate *per* base, it nonetheless delivers high-throughput data ($> 10^5$ reads *per* conditions) on the size distribution of fragmented DNA for each irradiation dose at a low cost thanks to the use of Flongle flow cells. The long-read sequencing of fragmented molecules does not reveal any specific radiation-induced fragmentation sites that might be associated with extended repetitions of a particular base motif, potentially causing localized reductions in the chain's binding energy. In the case of pBR322 plasmid, the base sequence contains few repeating patterns, which do not exceed 6 bp (GenBank J01749.1). Therefore, the sequencing data shows that this level of repeatability is not sufficient to reduce or increase the strand stability. Longer repetitive sequences are much more widespread in some other non-artificial DNAs, the study of their possible sensitivity to radiation-induced DNA damage is of considerable interest for the eventual finding of preferential sites of radiation-induced DNA fragmentation.

Our AGE results fall within the range of expected damage based on linear proton beam energy transfer, as shown by comparisons with previous experiments irradiating dry plasmids with hydrogen[30] and protons beams[29,31] (Fig. 2). This agreement validates our long-read sequencing experiment as an alternative method to quantify radiation-induced DNA fragmentation and more precisely SSB yields. All these measurements were carried out using a post-irradiation electrophoresis technique, and SSB yields were extracted using the McMahon method. Nonetheless, as reported in the "Results" section, there is a wide dispersion with a factor of up to 4 between the various results on SSB rates for experimental protocols that are, in a first approach, fairly similar. There are a number of possible reasons for these disparities with some of interest for sequencing measurements in the future, which we will now discuss.

First, we found that dose and LET are defined differently by the authors depending on whether they consider DNA or water as the reference material. Nevertheless, these choices of materials lead to a difference in dose and

LET estimates of around 20% at maximum, which cannot lead to differences between fitted yields with a factor of 2–4.

Second, the degree of initial conformation of plasmids is different between the experiments of the different authors. For example, in the studies of Souici et al. and Vysin et al., 95% of pBR322 plasmids exhibit supercoiled topology in controls, suggesting a higher order DNA structure with a greater number of twists than in the DNA samples considered in our experimental situation. Different supercoiled topologies can coexist and the rates of transitions from one topology to another induced by SSB remain unknown[36,37]. However, with a similar degree of supercoiled conformation, Souici et al. and Vysin et al., achieve an SSB yield which differs by a factor of 4, which leads us to look for other explanations.

Third, the degree of hydration of DNA can vary between experiments. Various studies have shown that hydration rates range from 4 to 35 water molecules *per* nucleotide, depending on the degree of hygrometry of the air (from dry to water-saturated, respectively)[38]. Irradiation experiments with ^{60}Co X-rays suggest an increase in SSB of around 40% within this range of hydration[39]. Once again, this variable, which may be responsible for a certain level of dispersion in the results, cannot explain a variation of a factor up to 4 between the measured SSB yields.

Finally, we identify two types of dry DNA irradiation protocols from all these works. Some authors irradiate DNA deposited on a plastic or glass support under vacuum[30,32], while others perform irradiations in air with a sample sealed with a micrometric film of plastic[29,31,33], as in our study. A summary of the reported data clearly shows a systematic difference depending on protocols with irradiations carried out under vacuum systematically displaying lower SSB yields (Fig. 3). In our specific case, it is assumed that the physical interactions of protons with the DNA support produced secondary electrons that interfere with the DNA damage production.

Lastly, we decided to compare our experimental data obtained from long-read sequencing with data obtained from Monte Carlo simulation and more precisely the newly developed "molecularDNA" application of Geant4-DNA. As air-dried plasmids are locally hydrated by humidity in the atmosphere, it supports the choice of water as a simulation material. However, the exact value of the simulation parameter (r) representing the hydration shells supporting the DNA molecule cannot be directly measured and could be misestimated in current simulations. From the present study, it seems clear that the sequencing of irradiated DNA molecules in liquid water will provide important data for improving simulation codes in the future. Indeed, the fragment size measurement method was implemented in the Geant4-DNA simulations to replicate the long-read sequencing (random selection of a DNA strand and reading until interrupted by breaks) allowing a direct comparison with sequencing data. The three-dimensional modeling of longer DNA molecules with Geant4-DNA would enable us to compare more complex models and test the ONT sequencer's capabilities regarding the size of the DNA under study. In addition, the irradiation of DNA samples in pure water should make it possible to extend studies to the indirect effects of IR and to treat damage induced by radiation caused by reactive oxygen species, while integrating the chemical simulations of Geant4-DNA. Using long-read sequencing could be an interesting approach to identify the presence of radiation-induced chemically modified bases. Long-read sequencing would also provide a detailed, base-by-base analysis, allowing the detection of such modifications by observing deviations from the expected base patterns. In such perspective, the use of high-resolution MinION Flow Cell proposed by ONT with 512 nanopores with a greatest sensitivity will be mandatory.

In this context, the ONT sequencer capacity to detect mutations and chemical base modifications could broadly widen the scope of DNA damage simulations. Then, it would be interesting to realize long-read sequencing of thin layer of plasmid DNA solution irradiated in water or water complemented with radical scavengers (such as Dimethyl sulfoxide, 2-Mercaptoethanol …) to detect and quantify such DNA modifications. Certain important developments are still required before accurate simulations of radiation-induced DNA fragmentation can be applied, as for example, the implementation of DNA cross-sections, as the current DNA models filled with water may potentially underestimate the damage[5,40–42].

In conclusion, the joint use of ONT long-read sequencing and Geant4-DNA offers numerous possibilities for studying radiation-induced DNA fragmentation. Both technologies are regularly updated and improved which should further enhance the interest in this type of comparative study in the future.

Material and methods

Experimental setup of DNA irradiation

pBR322 plasmids were provided by the Takara company (Cat. # 3050, Takara Bio Europe). The stock concentration of the plasmid DNA was 0.5 µg µL^{-1} in 10 mM Tris–HCl, 1 mM EDTA, pH 8.0. Aliquots of 2 µL plasmid DNA solution were deposited in the center of 4-µm thick polypropylene foil (Cat. #PP30-FM-000140, Polypropylene/PP Film 0.004 mm thick 315 mm coil W PP, Goodfellow Cambridge Ltd, Huntingdon, England) which was sealed with a glass coverslip and placed on a dedicated sample holder specifically designed for the micro-irradiation beamline[43]. Before usage, polypropylene foils were cleaned by three incubations at room temperature for 30 min in ethanol solution (70%, v/v). Afterwards, they were dried and treated for 30 min with UV-C using a germicidal lamp under sterile conditions. DNA samples were kept for 3 h at room temperature under the laminar flow of a microbiology hood in sterile conditions.

Irradiation modalities

Proton irradiation was conducted at the AIFIRA facility, specifically using the Micro-irradiation beamline[43]. The proton beam was adjusted to 3 MeV, with a linear energy transfer (LET) of 12 keV µm^{-1} in water. The LET is 16.4 keV µm^{-1} in DNA, as calculated using SRIM[44] and as proposed by Vysin et al.[31]: the calculated LET for dry DNA samples being 1.3739 higher than the LET for liquid samples. This correction was taken into account in data evaluation. The proton beam was extracted into the air through a 150 nm thick Si_3N_4 window and irradiation was performed at room temperature and atmospheric conditions. The beam is configured into a 500 µm horizontal

line and scanned vertically at a 5 kHz rate to ensure uniform irradiation across a $500 \times 500\ \mu m^2$ square area. This line features a sharp dose drop-off at its edges. For irradiating larger areas uniformly, the sample holder was raster scanned mechanically at a constant speed to cover areas ranging from mm^2 to cm^2. Different doses were achieved by adjusting the speed of the sample holder and using electrostatic means to modulate the beam. The average current during the irradiation process was 270 pA (i.e. $\approx 1.7 \times 10^9$ protons s^{-1}). Assuming a constant LET of the protons (16.4 keV μm^{-1} in DNA) and for a given fluence F, the mean dose was calculated as:

$$D(Gy) = 0.16 \cdot LET\left(\text{keV } \mu m^{-1}\right) \cdot F\left(\mu m^{-2}\right)$$

Irradiations were performed over a $3 \times 3\ mm^2$ area to fully cover the DNA drop, which had a diameter of approximately 1 mm^2. Plasmid samples were placed on a dedicated sample holder and maintained at room temperature and atmospheric conditions. An overview of the sample design is shown in Fig. 3a. The DNA drop was spotted on a 4-µm thick polypropylene foil (Cat. #PP30-FM-000140, Polypropylene/PP Film 0.004 mm thick 315 mm coil W PP, Goodfellow Cambridge Ltd, Huntingdon, England) stretched on a rigid body made in PEEK (Cat. #EK30-RD-000150, Polyether ether ketone, Goodfellow Cambridge Ltd, Huntingdon, England) by means of a clip. The back-side of the cell dish is closed using circular coverslips.

Oxford nanopore sequencing and mapping

The long-read sequencing method developed by Oxford Nanopore Technologies was used to study radiation-induced DNA fragmentation of the pBR322 samples described above. The circular plasmid DNA was digested post-irradiation using the *BamHI* enzyme, for 4 and 18 h at 37 °C in the dedicated buffer, to linearize it for adapter grafting. For sequencing, 250 ng of the digested plasmids were used, following the ligation sequencing kit protocol SQK-LSK109 adapted from the manufacturer's recommendation by eliminating the FFPE repair step. Flongle flow cells were chosen for sequencing as they provided a sufficient number of reads *per* sample and allowed for good statistical analysis of radio-induced fragment reads (with 128 nanopores channels). Additionally, Flongle flow cells are cost-effective, making them suitable for generating numerous samples to ensure the largest accessibility and dissemination of the study. Each library, corresponding to a control sample or an irradiated sample at doses of 0.5, 1, 2, or 5 kGy, was sequenced individually in separate sequencing runs. The sequencing runs lasted a minimum of 21 h with live basecalling enabled (using guppy v6.1.4) and included adapter trimming with a minimum quality score (min_qscore) of 8. Two separate batches were produced for each irradiation condition, ensuring reproducibility and reliability of the results. The resulting fastq files from basecalling were merged, and are mapped onto the pBR322 genome using minimap2 with the "–ax map-ont" option[45]. The read lengths of the mapped reads were obtained from the resulting alignment files, providing valuable information about the length distribution of the sequenced fragments. Fragment size distributions were plotted in Python using the matplotlib histogram function with the 'density' option set to True. To ensure data quality, reads that did not map onto the reference genome were excluded. Additionally, aberrant length values exceeding 110% of the reference length were removed from the dataset. Read lengths, including soft-clipped sequences, were extracted from SAM alignment files. The coverage *per* base (amount of time each base is sequenced) was extracted from the genome alignment file, scaled and visualized using the software "circos v0. 69–9". Base coverage was calculated using UGENE v44.0 and normalized by the read count with the 'max' setting representing the maximum base coverage value observed at 0 kGy (controls).

Agarose Gel Electrophoresis analysis (AGE)

Following irradiation, the samples were stored for long-term preservation at – 20 °C in a sterile and sealed petri dish. For subsequent analysis by agarose gel electrophoresis (AGE), control samples were processed in the same manner but without irradiation and without *BamHI* restriction. Recovery of DNA from the polypropylene foil was done by resuspension in 8 µL of TBE 1X buffer (Cat. #1610770, 10 mM Tris–Borate, 1 mM EDTA, pH 8.0, BioRad) at room temperature (RT). For AGE, the recovered DNA samples (without *BamHI* restriction) were loaded onto an agarose gel (Cat. #1613028, 1% (w/v) TBE Wide Mini ReadyAgarose Precast Gel, 15.6×10 cm, 20-well, with ethidium bromide, BioRad) containing ethidium bromide and electrophoresed at 100 V for 1h at RT in TBE 1X buffer. Following migration, gel images were captured using a Multiplex gel imaging system (Chemidoc MP Imaging System, BioRad). DNA bands were quantified using the manufacturer's image analysis software (Image Lab software, BioRad). By doing so, DNA bands were visualized and quantified to analyze the effects of irradiation on plasmid DNA conformations (supercoiled, open circular and linear).

Geant4 and Geant4-DNA simulations of DNA irradiation

We used Geant4 Monte Carlo simulation toolkit version 11.0 to simulate the secondary electron contribution to plasmid damage following the passage of protons through the polypropylene film covering the dry DNA samples (Fig. 3a). Similar to the experiment and as illustrated in Fig. 3a, 3 MeV protons were shot in vacuum passing through a 150 nm Si_3N_4 exit window traversing a 700 µm distance through air and reaching the 4 µm polypropylene film. The number and kinetic energy of all protons and secondary electrons leaving the film on the DNA sample side were registered to produce the energy spectrum of particles that would interact with the sample.

To test the number of secondary electrons that would backscatter from the polypropylene film, the same simulation was carried out replacing the air traversed by protons with vacuum and the kinetic energy and the number of secondary electrons rebounding back towards the exit window are registered. These simulations were carried out using the Livermore electromagnetic physics constructor with 10^7 primary protons each.

To accurately simulate the physics interactions of protons occurring in water and inside biological geometries, the G4EmDNAPhysics_option2 physics list of the Geant4-DNA was used. This list specifically tracks secondary

electrons of energy ranging from few eV up to 1 MeV[23]. Interactions with the DNA molecules are simulated using the water cross-sections.

The "molecularDNA" example is included in the extended examples category of the Geant4 11.1 public release (June 2022) for Geant4-DNA[31,35,36]. It offers various models of DNA-scale biological targets that can be used for simulations. The geometry parameterization features provided in this example allow users to create multiple biological targets within a single irradiation setup, enabling statistical analysis of the results. Furthermore, the "molecularDNA" example includes a user-friendly application with a comprehensive damage classification scheme. This scheme considers the complexity of the DNA damage and distinguishes between direct and indirect damage. Users can use this complex analysis tool to assign different types of DNA damage, such as single-strand breaks (SSB) and double-strand breaks (DSB), based on the simulation results[37–39]. Simulations were performed using a pBR322 plasmid-like DNA model constructed in a supercoiled shape with a double helix structure[46–48]. The nucleobases were represented by ellipsoids, while the sugar and phosphate groups were represented by spheres filled with water. The supercoiled plasmid geometry consisted of 4361 bp, with 6 bp added to accommodate geometrical constraints and model the loop structure. To reproduce the experimental density of about 0.5 μg μL^{-1}, up to 10,142 identical plasmids were simulated. These plasmids were evenly spaced within a water box measuring 4.42 μm × 4.42 μm × 4.48 μm (Supplementary Fig. S3). Each plasmid was placed inside a spherical voxel with a diameter of 200 nm, and the voxels were randomly oriented in space (Supplementary Fig. S3c). The centers of the voxels were separated by 210 nm to avoid any overlap between plasmids. Irradiation conditions in the simulation setup were kept consistent with the experimental setup, enabling a direct comparison. A 3 MeV proton beam was directed randomly towards one face of the box enclosing the plasmids. 5,000, 10,000, 20,000 and 50,000 incident particles were continuously shot until the cumulative energy depositions within the box resulted in absorbed doses of 0.5, 1, 2, and 5 kGy (Supplementary Fig. S3c). As simulations were performed in a water medium, the DNA molecules were placed in water voxels. In addition, each base was considered as a distinct element within the sequence.

In our model, direct damage occurs when energy from physical processes is deposited on or near a DNA molecule. To determine the association of damage with a DNA molecule, we used a single distance value as shown in Supplementary Fig. S3. This value, known as the direct interaction range (r), represents the maximum distance from the center of a molecule where energy deposition can lead to damage. It takes into consideration the hydration shells that support the double helix DNA structure. To assign damage to a specific DNA nucleotide, we considered the volume of its sugar-phosphate group and the surrounding hydration shell with a thickness (r). Any energy deposited within this combined volume above a certain threshold (t) is classified as a single-strand break (SSB) corresponding to that DNA nucleotide. When two SSBs occurred on opposite strands and were separated by 10 bp or less, they constituted a double-strand break (DSB). For the purpose of this simulation, we disregarded damage to nucleobases. Even though the plasmid simulations are performed in water, only the physical interactions are simulated considering the direct damages only to model the "dry" plasmid condition. The simulated plasmids are considered intact in the form of a closed loop prior to irradiation. A DNA fragment length is defined as the distance between a strand break and a specified cut point on the plasmid for linearization. Irradiation doses ranging from 0.5 to 5 kGy are simulated to compare with the experimental data.

Data availability

Raw sequencing data is available under BioProject accession number PRJNA985044.

Code availability

"*molecularDNA*" example is publicly available within the extended examples of Geant4 after downloading and installing the software toolkit from the official website: More details are found on the Geant4-DNA website: http://geant4-dna.org/ and the dedicated documentation site: https://geant4-dna.github.io/molecular-docs/.

Received: 8 April 2024; Accepted: 5 August 2024
Published online: 12 August 2024

References

1. Yousefzadeh, M. *et al.* DNA damage-how and why we age?. *Elife* **10**, e62852 (2021).
2. UNSCEAR 2008 Report Volume I. *United Nations: Scientific Committee on the Effects of Atomic Radiation.* www.unscear.org/unscear/en/publications/2008_1.html.
3. Lord, C. J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287–294 (2012).
4. Shepard, C., Yost, D. C. & Kanai, Y. Electronic excitation response of DNA to high-energy proton radiation in water. *Phys. Rev. Lett.* **130**, 118401 (2023).
5. Shepard, C. & Kanai, Y. Ion-type dependence of DNA electronic excitation in water under proton, α-particle, and carbon ion irradiation: A first-principles simulation study. *J. Phys. Chem. B* **127**, 10700–10709 (2023).
6. Li, L., Story, M. & Legerski, R. J. Cellular responses to ionizing radiation damage. *Int. J. Radiat. Oncol. Biol. Phys.* **49**, 1157–1162 (2001).
7. Heller, C., Duke, T. & Viovy, J. L. Electrophoretic mobility of DNA in gels. II. Systematic experimental study in agarose gels. *Biopolymers* **34**, 249–259 (1994).
8. Viovy, J.-L. Electrophoresis of DNA and other polyelectrolytes: Physical mechanisms. *Rev. Mod. Phys.* **72**, 813–872 (2000).
9. Maayah, Y., Nusrat, H., Pang, G. & Tambasco, M. Assessing the DNA damaging effectiveness of ionizing radiation using plasmid DNA. *Int. J. Mol. Sci.* **23**, 12459 (2022).
10. McMahon, S. J. & Currell, F. J. A robust curve-fitting procedure for the analysis of plasmid DNA strand break data from gel electrophoresis. *Radiat. Res.* **175**, 797–805 (2011).
11. Birren, B. W., Lai, E., Hood, L. & Simon, M. I. Pulsed field gel electrophoresis techniques for separating 1- to 50-kilobase DNA fragments. *Anal. Biochem.* **177**, 282–286 (1989).

12. Xu, X. *et al.* Direct observation of damage clustering in irradiated DNA with atomic force microscopy. *Nucleic Acids Res.* **48**, e18 (2020).
13. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).
14. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
15. Jain, M., Olsen, H. E., Akeson, M. & Abu-Shumays, R. Adaptation of human ribosomal RNA for nanopore sequencing of canonical and modified nucleotides. *Methods Mol. Biol.* **2298**, 53–74 (2021).
16. Friedland, W. *et al.* Comprehensive track-structure based evaluation of DNA damage by light ions from radiotherapy-relevant energies down to stopping. *Sci. Rep.* **7**, 45161 (2017).
17. Nikjoo, H., Uehara, S., Emfietzoglou, D. & Cucinotta, F. A. Track-structure codes in radiation research. *Radiat. Meas.* **41**, 1052–1074 (2006).
18. Francis, Z., Villagrasa, C. & Clairand, I. Simulation of DNA damage clustering after proton irradiation using an adapted DBSCAN algorithm. *Comput. Methods Programs Biomed.* **101**, 265–270 (2011).
19. Nikjoo, H., O'Neill, P., Terrissol, M. & Goodhead, D. T. Quantitative modelling of DNA damage using Monte Carlo track structure method. *Radiat. Environ. Biophys.* **38**, 31–38 (1999).
20. Bernal, M. A. *et al.* Track structure modeling in liquid water: A review of the Geant4-DNA very low energy extension of the Geant4 Monte Carlo simulation toolkit. *Phys. Med.* **31**, 861–874 (2015).
21. Incerti, S. *et al.* The geant4-dna project. *Int. J. Model. Simul. Sci. Comput.* **01**, 157–178 (2010).
22. Incerti, S. *et al.* Comparison of GEANT4 very low energy cross section models with experimental data in water. *Med. Phys.* **37**, 4692–4708 (2010).
23. Incerti, S. *et al.* Geant4-DNA example applications for track structure simulations in liquid water: A report from the Geant4-DNA project. *Med. Phys.* **45**, e722–e739 (2018).
24. Chatzipapas, K. P. *et al.* Simulation of DNA damage using Geant4-DNA: An overview of the "molecularDNA" example application. *Precis. Radiat. Oncol.* **7**, 4–14 (2023).
25. Lin, B., Hui, J. & Mao, H. Nanopore technology and its applications in gene sequencing. *Biosensors (Basel)* **11**, 214 (2021).
26. Balbas, P., Soberon, X., Bolivar, F. & Rodriguez, R. L. CHAPTER 1: The plasmid, pBR322. In *Vectors* (eds Rodriguez, R. L. & Denhardt, D. T.) 5–41 (Butterworth-Heinemann, 1988). https://doi.org/10.1016/B978-0-409-90042-2.50007-6.
27. Sahlin, K. & Medvedev, P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nat. Commun.* **12**, 2 (2021).
28. Bourret, S. *et al.* Fluorescence time-lapse imaging of single cells targeted with a focused scanning charged-particle microbeam. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **325**, 27–34 (2014).
29. Ushigome, T. *et al.* Yield of single- and double-strand breaks and nucleobase lesions in fully hydrated plasmid DNA films irradiated with high-LET charged particles. *Radiat. Res.* **177**, 614–627 (2012).
30. Wyer, J. A. *et al.* Fragmentation and plasmid strand breaks in pure and gold-doped DNA irradiated by beams of fast hydrogen atoms. *Phys. Med. Biol.* **54**, 4705–4721 (2009).
31. Vyšín, L. *et al.* Proton-induced direct and indirect damage of plasmid DNA. *Radiat. Environ. Biophys.* **54**, 343–352 (2015).
32. Souici, M. *et al.* Single- and double-strand breaks of dry DNA exposed to protons at Bragg-peak energies. *J. Phys. Chem. B* **121**, 497–507 (2017).
33. Urushibara, A. *et al.* LET dependence of the yield of single-, double-strand breaks and base lesions in fully hydrated plasmid DNA films by $^{4}He^{2+}$ ion irradiation. *Int. J. Radiat. Biol.* **84**, 23–33 (2008).
34. Nikjoo, H. *et al.* Radiation track, DNA damage and response: A review. *Rep. Prog. Phys.* **79**, 116601 (2016).
35. Nakano, M. *et al.* Local thermodynamics of the water molecules around single- and double-stranded DNA studied by grid inhomogeneous solvation theory. *Chem. Phys. Lett.* **660**, 250–255 (2016).
36. Krajina, B. A. & Spakowitz, A. J. Large-scale conformational transitions in supercoiled DNA revealed by coarse-grained simulation. *Biophys. J.* **111**, 1339–1349 (2016).
37. Marko, J. F. & Siggia, E. D. Statistical mechanics of supercoiled DNA. *Phys. Rev. E* **52**, 2912–2938 (1995).
38. Swarts, S. G., Sevilla, M. D., Becker, D., Tokar, C. J. & Wheeler, K. T. Radiation-induced DNA damage as a function of hydration: I. Release of Unaltered Bases. *Radiat. Res.* **129**, 333–344 (1992).
39. Yokoya, A. *et al.* Yields of strand breaks and base lesions induced by soft X-rays in plasmid DNA. *Radiat. Prot. Dosim.* **122**, 86–88 (2006).
40. Zein, S. A. *et al.* Electron transport in DNA bases: An extension of the Geant4-DNA Monte Carlo toolkit. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **488**, 70–82 (2021).
41. Krim, M. *et al.* Cross sections, stopping power and Bragg peak range calculation of proton collisions with the DNA base adenine. *Jpn. J. Appl. Phys.* **58**, 096001 (2019).
42. Tan, Z., Xia, Y., Zhao, M. & Liu, X. Proton stopping power in a group of bioorganic compounds over the energy range of 0.05–10 MeV. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **248**, 1–6 (2006).
43. Barberet, P., Jouve, J., Sorieul, S., Alfaurt, P. & Mathieu, L. AIFIRA: A light ion beam facility for ion beam analysis and irradiation. *Eur. Phys. J. Plus* **136**, 67 (2021).
44. Ziegler, J. F., Ziegler, M. D. & Biersack, J. P. SRIM: The stopping and range of ions in matter (2010). *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **268**, 1818–1823 (2010).
45. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
46. Lampe, N. *et al.* Mechanistic DNA damage simulations in Geant4-DNA part 1: A parameter study in a simplified geometry. *Phys. Med.* **48**, 135–145 (2018).
47. Agostinelli, S. *et al.* Geant4: A simulation toolkit. *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrom. Detect. Assoc. Equip.* **506**, 250–303 (2003).
48. Sakata, D. *et al.* Fully integrated Monte Carlo simulation for evaluating radiation induced DNA damage and subsequent repair using Geant4-DNA. *Sci. Rep.* **10**, 20788 (2020).

Acknowledgements

We thank the technical staff members of the AIFIRA facility J. Jouve, P. Alfaurt and S. Sorieul. The AIFIRA facility is financially supported by the CNRS, the University of Bordeaux and the Région Nouvelle Aquitaine. The Mission pour les Initiatives Transverses et Interdisciplinaires (MITI) of CNRS is also thanked for the funding of the PhD thesis of PiB, in the frame of the "80'PRIME" program. The authors would like to thank F. Chappuis PhD student at the Institute of Radiation Physics, Lausanne, Switzerland for her advice on the fitting model of damage yield calculation. The authors would also like to thank the BioRad III project members funded by the European Spatial Agency. SAZ was supported by the CNRS and the Lausanne University Hospital (CHUV). KC was supported by the BioRAD III projet (European Spatial Agency).

Author contributions

H.S. conceived and supervised the project. Ph.B., G.D. and H.S. designed and performed the radiation biology experiments. L.P., H.S. performed the DNA sample preparation and long-read sequencing experiments. PiB performed the bioinformatic and statistical analysis. S.A.Z., K.C., H.T., S.I. developed the Monte Carlo Geant4-DNA code. S.A.Z., K.C., H.T., S.I. performed Monte Carlo simulations. Pi.B., S.A.Z., K.C., H.T., R.L., F.G., H.S. contributed to the data analysis and interpreted the results. G.D., R.L., F.G., D.D. provided expert review and editing. Pi.B., S.A.Z., F.G. and H.S. wrote the main manuscript and prepared the remaining figures. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-69406-3.

Correspondence and requests for materials should be addressed to H.S.

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

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