RadiSeq: a single- and bulk-cell whole-genome DNA sequencing simulator for radiation-damaged cell models

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

CrossMark

OPEN ACCESS

RECEIVED
5 February 2025

REVISED
21 July 2025

ACCEPTED FOR PUBLICATION
31 July 2025

PUBLISHED
19 August 2025

PAPER

RadiSeq: a single- and bulk-cell whole-genome DNA sequencing simulator for radiation-damaged cell models

F Mathew[1,2,*], Luc Galarneau[1,2] and J Kildea[1,2]

[1] Medical Physics Unit, McGill University, Montreal, QC, Canada
[2] Research Institute, McGill University Health Centre, Montreal, QC, Canada
[*] Author to whom any correspondence should be addressed.

E-mail: felix.mathew@mail.mcgill.ca

Keywords: RadiSeq, ScWGS and BcWGS simulator, sequencer for radiation-damaged cell models, Monte Carlo DNA sequencing, tool for radiobiology study

Supplementary material for this article is available online

Abstract

Objective. To build and validate a simulation framework to perform single-cell and bulk-cell whole genome sequencing simulation of radiation-exposed Monte Carlo (MC) cell models to assist radiation genomics studies. *Approach.* Sequencing the genomes of radiation-damaged cells can provide useful insight into radiation action for radiobiology research. However, carrying out post-irradiation sequencing experiments can often be challenging, expensive, and time-consuming. Although computational simulations have the potential to provide solutions to these experimental challenges, and aid in designing optimal experiments, the absence of tools currently limits such application. MC toolkits exist to simulate radiation exposures of cell models but there are no tools to simulate single- and bulk-cell sequencing of cell models containing radiation-damaged DNA. Therefore, we aimed to develop a MC simulation framework to address this gap by designing a tool capable of simulating sequencing processes for radiation-damaged cells. *Main results.* We developed RadiSeq—a multi-threaded whole-genome DNA sequencing simulator written in C++. RadiSeq can be used to simulate Illumina sequencing of radiation-damaged cell models produced by MC simulations. RadiSeq has been validated through comparative analysis, where simulated data were matched against experimentally obtained data, demonstrating reasonable agreement between the two. Additionally, it comes with numerous features designed to closely resemble actual whole-genome sequencing. RadiSeq is also highly customizable with a single input parameter file. *Significance.* RadiSeq enables the research community to perform complex simulations of radiation-exposed DNA sequencing, supporting the optimization, planning, and validation of costly and time-intensive radiation biology experiments. This framework provides a powerful tool for advancing radiation genomics research.

1. Introduction

Genome sequencing studies have provided valuable insights into the biophysical mechanisms of ionizing radiation (IR) mutagenesis (Adewoye *et al* 2015, Youk *et al* 2024). However, the high cost and technical challenges associated with these types of experiments, particularly single-cell whole-genome sequencing (ScWGS), often limit their feasibility (Gawad *et al* 2016). To minimize the cost and uncertainty associated with post-irradiation genome sequencing studies, careful experimental design is crucial. To this end, we have developed a Monte Carlo (MC) simulation framework designed to simulate DNA sequencing of IR-damaged cells. This framework can be used to optimize experimental parameters and predict experimental outcomes, thereby reducing the risk of wasted resources and ensuring more efficient and informative experiments.

MC simulations are extensively used to model IR interactions probabilistically. Geant4 (Agostinelli *et al* 2003) and its user-friendly wrapper TOPAS (Faddegon *et al* 2020), PHITS (Niita *et al* 2006), FLUKA

(Ferrari *et al* 2005), and EGSnrc (National Research Council of Canada. Metrology Research Centre. Ionizing Radiation Standards 2021) are some examples among the numerous MC simulation toolkits used in radiotherapy-related research. With the goal of advancing the science and understanding of radiobiological effects at the cellular and subcellular levels, some of these toolkits have been extended with track structure modeling to simulate IR interactions at the DNA scale; Geant4-DNA (Incerti *et al* 2010), TOPAS-nBio (Schuemann *et al* 2019a), and PARTRAC (Friedland *et al* 2011) are some examples. Various geometric nuclear DNA models have also been developed and are used in MC-based radiobiology investigations (Bernal and Liendo 2009, Bernal *et al* 2013, McNamara *et al* 2018, Sakata *et al* 2020, Montgomery *et al* 2021, Bertolet *et al* 2022). With the abundance of MC toolkits and DNA models now available, it is possible to simulate in detail the interaction of IR in cells to obtain a complete description of the DNA damages introduced. A recently-published data standard, called the standard DNA damage (SDD) format, has also been developed to report the DNA damage information from such simulations (Schuemann *et al* 2019b).

Genome sequencing has matured into a valuable biological technique with numerous applications, including in the field of IR research. Various sequencing approaches and bioinformatics techniques have been employed to investigate IR-induced genomic effects including mutation signatures (Behjati *et al* 2016, Kageyama *et al* 2021), DNA damage (Murray *et al* 2019), and genomic alterations (Nguyen *et al* 2016, Youk *et al* 2024). While these approaches have typically focused on using tumors or clonal populations of cells in which all cells harbor the same mutations, we hypothesized that ScWGS of heterogeneous IR-damaged cell samples could provide additional insights into IR-induced genomic alterations. Our novel experiments yielded promising results (Mathew *et al* 2023). However, we found that conducting novel post-irradiation ScWGS experiments is complex, expensive, and time-consuming. In this context, we postulated that the ability to simulate ScWGS and other sequencing approaches *in silico* can help with experimental design and can serve as a valuable complement to the experiments themselves. However, despite our ability to model IR-induced DNA damage, we were faced with a lack of tools to simulate the sequencing of damaged DNA.

A plethora of tools have emerged for simulating next-generation sequencing (NGS) techniques, offering various models for different sequencing technologies and protocols. Notable recent reviews by Alosaimi *et al* (2020), Escalona *et al* (2016), and Zhao *et al* (2017) have assessed many simulators, evaluating their strengths and weaknesses. While most simulators can reasonably imitate real-world NGS procedures, there is no single-best solution suitable for all experimental scenarios (Milhaven and Pfeifer 2023). Notably, none of the existing tools can simulate the sequencing of diploid cells with strand breaks, which is a requisite for modeling the sequencing of IR-damaged cells since IR will most likely create somatic heterozygous mutations. Strand breaks change the size distribution of DNA fragments, which can affect how sequencing reads are mapped to the reference genome. This may introduce variability in sequencing coverage and influence variant calling, potentially changing the number of genomic alterations detected (Normand and Yanai 2013, Pommerenke *et al* 2016). Thus, the development of a new sequencing simulation framework, specifically to integrate with MC-generated IR-damaged cell models, was a prerequisite for us to use sequencing simulations to guide our irradiation and ScWGS experiments.

Here, we present RadiSeq, an open-source simulation framework (Mathew and Kildea 2024) that we developed for Illumina (Illumina Inc. San Diego, California, USA) ScWGS and bulk-cell whole-genome sequencing (BcWGS) of IR-damaged MC-generated cell models that describe their DNA damage in the SDD format.

2. Materials and methods

RadiSeq was developed in C++ with a high degree of end-user customizability in mind. Figure 1 presents a schematic of RadiSeq's processing pipeline. The main goal of developing RadiSeq was to emulate real experiments and predict sequencing outcomes. This stands in contrast to the purpose of most existing sequencing simulation tools that aim to produce 'gold standard' sequencing data for the purposes of validating and assessing bioinformatics analysis pipelines. Consequently, RadiSeq was designed to allow users to customize the simulation by using an input parameter file that contains all of the adjustable parameters needed by the models provided in the simulation framework.

In a typical whole-genome DNA sequencing workflow (Normand and Yanai 2013), the first step is sample preparation, which includes extracting DNA from cells, fragmenting the DNA, amplifying it, and attaching unique identifiers—specific to each sample in BcWGS, or to each cell in ScWGS. The prepared 'DNA library' is then loaded into a sequencing machine, which identifies the nucleotide bases in each fragment to generate sequencing reads based on the selected sequencing technology, typically ranging from 50 to 300 bp in length, with 150 bp being standard for whole genome sequencing. This step may introduce additional errors due to the underlying sequencing chemistry. Finally, the reads are computationally aligned to a reference genome, and genomic alterations are identified using variant calling algorithms (Xu 2018).

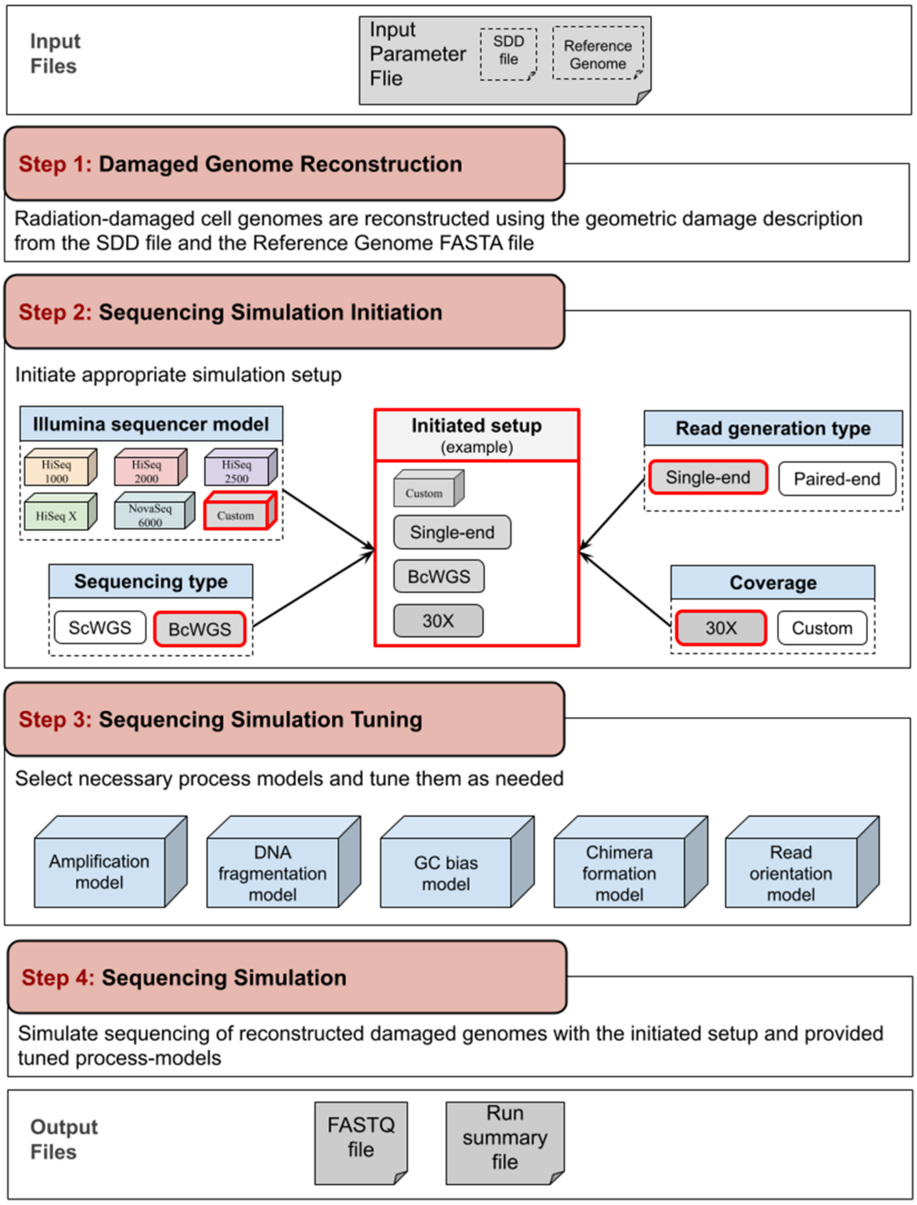
Input Files
Input Parameter Flie
SDD file
Reference Genome
Step 1: Damaged Genome Reconstruction
Radiation-damaged cell genomes are reconstructed using the geometric damage description from the SDD file and the Reference Genome FASTA file
Step 2: Sequencing Simulation Initiation
Initiate appropriate simulation setup
Illumina sequencer model
HiSeq 1000
HiSeq 2000
HiSeq 2500
HiSeq X
NovaSeq 6000
Custom
Initiated setup (example)
Custom
Single-end
BcWGS
30X
Read generation type
Single-end
Paired-end
Sequencing type
ScWGS
BcWGS
Coverage
30X
Custom
Step 3: Sequencing Simulation Tuning
Select necessary process models and tune them as needed
Amplification model
DNA fragmentation model
GC bias model
Chimera formation model
Read orientation model
Step 4: Sequencing Simulation
Simulate sequencing of reconstructed damaged genomes with the initiated setup and provided tuned process-models
Output Files
FASTQ file
Run summary file

Figure 1. A schematic representation of RadiSeq's sequencing simulation pipeline showing the four steps in a simulation. The red outline is used to delineate an example use case in step 2.

RadiSeq replicates this pipeline *in silico* for radiation-damaged DNA, beginning with the construction of a damaged cell genome based on a radiation-damaged double-stranded DNA obtained from an external MC simulation. Users can define models within RadiSeq that control DNA fragmentation, amplification, and include error profiles to account for sequencing-related errors and biases in the read data. The simulation process is organized into four steps:

Step 1: Damaged genome reconstruction

RadiSeq's ability to simulate the sequencing of IR-damaged DNA is what sets it apart from other sequencing simulators. DNA models used in MC irradiation simulations are geometric only–they record the location of DNA damage but contain no genomic sequence information. Therefore, RadiSeq's first task is to create a sequenceable damaged genome. To do so, it takes two files as input, one containing the damage-annotated geometric genome of one or more cells in SDD format, as outputted by any external MC pipeline for cell irradiation simulation, and the other containing a complete reference genome in the FASTA format (Lipman and Pearson 1985). As illustrated in figure 2, a simple mapping algorithm then folds the reference genome onto the geometric genome, while maintaining the damage location annotations (base damages and strand breaks; each damage within a clustered damage site is considered separately); the reference genome of any diploid organism, not necessarily human, can be used for this purpose.

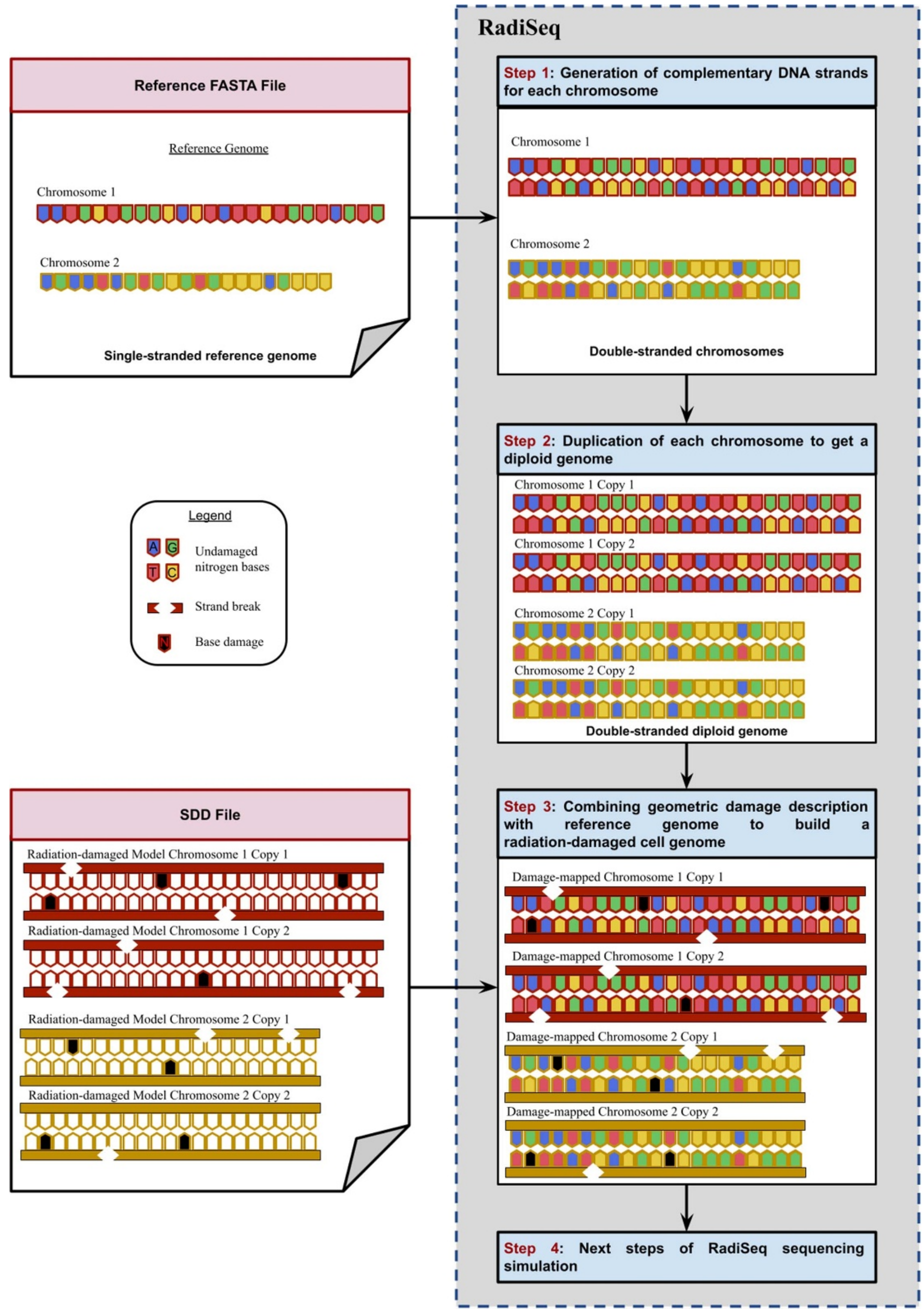
RadiSeq
Reference FASTA File
Reference Genome
Chromosome 1
Chromosome 2
Single-stranded reference genome
Step 1: Generation of complementary DNA strands for each chromosome
Chromosome 1
Chromosome 2
Double-stranded chromosomes
Step 2: Duplication of each chromosome to get a diploid genome
Chromosome 1 Copy 1
Chromosome 1 Copy 2
Chromosome 2 Copy 1
Chromosome 2 Copy 2
Double-stranded diploid genome
Legend
A G T C
Undamaged nitrogen bases
Strand break
Base damage
SDD File
Radiation-damaged Model Chromosome 1 Copy 1
Radiation-damaged Model Chromosome 1 Copy 2
Radiation-damaged Model Chromosome 2 Copy 1
Radiation-damaged Model Chromosome 2 Copy 2
Step 3: Combining geometric damage description with reference genome to build a radiation-damaged cell genome
Damage-mapped Chromosome 1 Copy 1
Damage-mapped Chromosome 1 Copy 2
Damage-mapped Chromosome 2 Copy 1
Damage-mapped Chromosome 2 Copy 2
Step 4: Next steps of RadiSeq sequencing simulation

Figure 2. A schematic showing the steps of damaged-genome reconstruction in RadiSeq.

Step 2: Sequencing simulation initiation

Given the variability in error-rate profiles and overall read quality across sequencing platforms, RadiSeq includes five pre-configured Illumina sequencer models and offers an option to specify a custom sequencer to include additional Illumina sequencers using user-provided error profiles. A sequencer profiling tool is bundled with RadiSeq to help users generate error profiles (see supplementary materials section 3.3). Users can choose between ScWGS and BcWGS to be performed on the reconstructed damaged genome from step 1. RadiSeq initiates the handling of reconstructed damaged genomes differently depending on this choice. Both single-end and paired-end (PE) read (Normand and Yanai 2013) generation methods in sequencing are supported.

Step 3: Sequencing simulation tuning

To account for additional variability in PE read generation, RadiSeq includes options for different read-pair orientations such as forward-forward, reverse-forward, and others (see supplementary material). RadiSeq includes two DNA amplification models: (i) a uniform amplification model and (ii) a non-uniform multiple-displacement amplification (MDA) (Lasken 2009) model. RadiSeq's modular design allows for easy expansion with additional models in the future. GC bias (Chen *et al* 2013) is modeled using a unimodal triangular function. Additionally, chimeric read artifact formation (Lu *et al* 2023) can be included in the simulation. All these pre-configured process models are easily customized and can be tuned to meet specific user needs.

Step 4: Read generation

The read-generation process differs between the ScWGS and BcWGS arms of RadiSeq. In BcWGS, each read or read pair is generated from a randomly-selected reconstructed cell within the sample, while in ScWGS, cells are processed sequentially. The desired coverage determines the total number of reads or the number of reads required per cell for BcWGS and ScWGS respectively. To construct a read (the first read in a pair for PE sequencing), RadiSeq evaluates the likelihood of a read originating from various genomic locations, considering user-specified factors such as GC bias and genome amplification. Each location is assigned a weight representing its effective sampling bias, obtained as a dot product of individual sample biases. For PE reads, RadiSeq additionally uses the user-provided fragmentation model to probabilistically determine the location of the second read in the pair, based on an appropriate DNA fragment size sampled according to the model. Read sequences, referred to as template sequences in RadiSeq, are then generated according to the specified read length and orientation. Finally, the generated template sequences are probabilistically modified by introducing chimeric sequences, insertions, deletions, and substitution errors, and their quality is adjusted accordingly to obtain the final reads. A detailed description of read generation is included in the supplementary materials (section: RadiSeq simulation procedure).

Implementation and performance

The generated read data are output as compressed text files that follow the FASTQ format specification (.fastq.gz) along with a RadiSeq run summary text file. A detailed schematic of the RadiSeq processing pipeline and additional details on each of the built-in models are included in the supplementary material.

Multithreading is enabled in RadiSeq and users can set the number of CPU threads for faster processing using the OpenMP parallel programming model (Chandra 2001), and operations are further optimized using memory mapping techniques. The current release of RadiSeq (RadiSeq_v2.0) requires C++ 17 or above, 64-bit (x64) CPU architecture, and is currently only compatible with Unix-based systems.

2.1. Performance evaluation of RadiSeq

We have been using a human geometric nuclear DNA model that that was built in-house and published open source as the 'NICE model' (Montgomery and Manalad 2022) for radiation exposure simulations in TOPAS-nBio. To evaluate the performance of RadiSeq, the NICE model, which represents a human lymphocyte cell in the G0/1 phase, was sham-irradiated (i.e. no actual irradiation) to generate an undamaged SDD file. This file was used to independently simulate both ScWGS and BcWGS in RadiSeq because the simulation pipeline differs for each. During these sequencing simulations, CPU run time was measured to assess performance. Illumina PE reads of length 151 bp were successfully generated for varying genomic coverages using the NovaSeq 6000 sequencer (Illumina Inc.)—one of the models included in RadiSeq. The DNA fragment size distribution was specified using a beta function with a beta value of 10, a mode value of 155 bp, and a maximum of 200 bp. These settings were designed to closely match the conditions of the sequencing experiments previously performed in our lab. Other important parameters used in the simulation that may influence runtime included zero degree of GC bias, a uniform amplification model, and a read artifacts rate of 0.3. Simulations were performed with 40 CPU threads on an Intel Xeon Gold 6148 @ 2.40 GHz processor of the Digital Research Alliance of Canada clusters.

2.2. Verification and validation of RadiSeq

Unit testing and integration testing were performed to ensure that the different components and models of RadiSeq functioned as expected. However, for the acceptance of RadiSeq, it was imperative to validate that various metrics pertaining to the simulated data would precisely mimic the same metrics for the experimental data for both the BcWGS and ScWGS pipelines separately. For this purpose, we used aligned and analyzed experimentally-obtained read data from sham-irradiated human B-lymphoblastoid cells in our lab (Mathew *et al* 2023). These cells belonged to a well-characterized genome of an Ashkenazi individual (Shumate *et al* 2020) and were sequenced both using ScWGS and BcWGS techniques using the Illumina NovaSeq 6000 sequencer. The corresponding experimental conditions were used in RadiSeq to generate simulated data and associated metrics for comparison. For ScWGS, one cell was randomly chosen from the experimental dataset to be used for this comparison.

The ultimate accuracy of RadiSeq simulations in describing the genomic effects of radiation damage depends on how precisely the upstream DNA damage is modeled, as this serves as the input for RadiSeq. Therefore, evaluating the absolute accuracy of these post-irradiation sequencing simulations was not the objective of this study. Instead, our goal was to assess how well RadiSeq could replicate sequencing experiments, using the sham-irradiated NICE cell model as a reference. For the ScWGS, the MDA model was used to amplify single-cell genomes with a DNA fragment size distribution comparable to our ScWGS experiments (see supplementary methods). For the BcWGS simulation pipeline, the uniform amplification model and a DNA fragment size distribution comparable to our BcWGS experiments were used in the simulation. For both the ScWGS and BcWGS pipelines, RadiSeq-generated reads were analyzed using the FastQC tool (www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then aligned to the reference human genome using the Bowtie 2 (Langmead and Salzberg 2012) aligner. Alignment statistics were obtained using Samtools (Li *et al* 2009) and compared against the experimentally obtained data.

2.3. Benchmarking of RadiSeq

To benchmark RadiSeq, we selected two of the latest open-source sequencing simulation tools: NGSNGS v0.9.2.2 (Henriksen *et al* 2023) and Sandy v0.25 (Miller *et al* 2023). Both simulators have been previously validated against other existing frameworks. While neither tool supports sequencing directly from computational cell models in SDD format as RadiSeq does, we were still able to emulate Illumina sequencing on a sham-irradiated cell sample using each simulator. To ensure a fair comparison, simulation parameters for both NGSNGS and Sandy were adjusted where possible to match our experimental data. The resulting sequencing reads were aligned using Bowtie 2, and alignment statistics were generated using Samtools following the same pipeline used for RadiSeq. The resulting metrics were then compared across all three tools.

3. Results

Figure 3 shows the RadiSeq computational performance in terms of CPU run time for both BcWGS and ScWGS simulations plotted on a logarithmic scale. A line connecting data points is included to guide the eye.

Tables 1 and 2 compare selected read alignment statistics—chosen to reflect overall performance—generated using Samtools for both RadiSeq-simulated data and the corresponding experimental datasets in ScWGS and BcWGS, respectively. To benchmark RadiSeq's performance, equivalent statistics from two other sequencing simulators, Sandy and NGSNGS, are included for direct comparison. A similarity score, calculated as $\text{Similarity} = \max(1 - |\text{simulated} - \text{experiment}|/\text{experiment}, 0)$, is used to quantify the agreement between simulated and experimental values.

In addition to the read-alignment statistics, Samtools provides several underlying distributions that describe the read-aligned data in more detail. Figure 4 presents two such distributions—(i) read coverage distribution and (ii) insert size distribution—for both experimentally obtained data and simulated data generated using RadiSeq, Sandy, and NGSNGS, shown separately for ScWGS and BcWGS. Additional metrics were also evaluated to assess the agreement between RadiSeq simulated and experimental results; these are included in the supplementary materials for completeness.

4. Discussion

We have built a MC simulation framework called RadiSeq for BcWGS and ScWGS simulation of diploid IR-damaged genomes. To the best of our knowledge, RadiSeq is the first MC framework to simulate the sequencing of radiation-damaged cell models. It is open-source and customizable, and it can generate realistic Illumina sequencing data that can be used for experiment planning and experimental outcome prediction.

Performance testing revealed that simulating the sequencing of sham-irradiated cells takes less than 10 min for up to 10^7 reads with RadiSeq. Beyond this point, the CPU run time increases with the number of reads for both BcWGS and ScWGS simulations, given the same number of CPU threads. For simulations involving up to 10^7 reads, RadiSeq consistently requires 10 min, primarily due to simulation initialization and input file processing. Beyond this threshold, read generation time becomes a more substantial factor, leading to increased overall simulation duration. These two performance regions are visually distinguished in the plot using different shading. It is important to note that CPU run time and memory usage can vary significantly based on model customizations, the number of radiation damages in each cell and the computer resources used.

Both the read alignment statistics and the various distributions obtained using Samtools for the RadiSeq-simulated data closely resembled those of the experimental results under the tested conditions. To

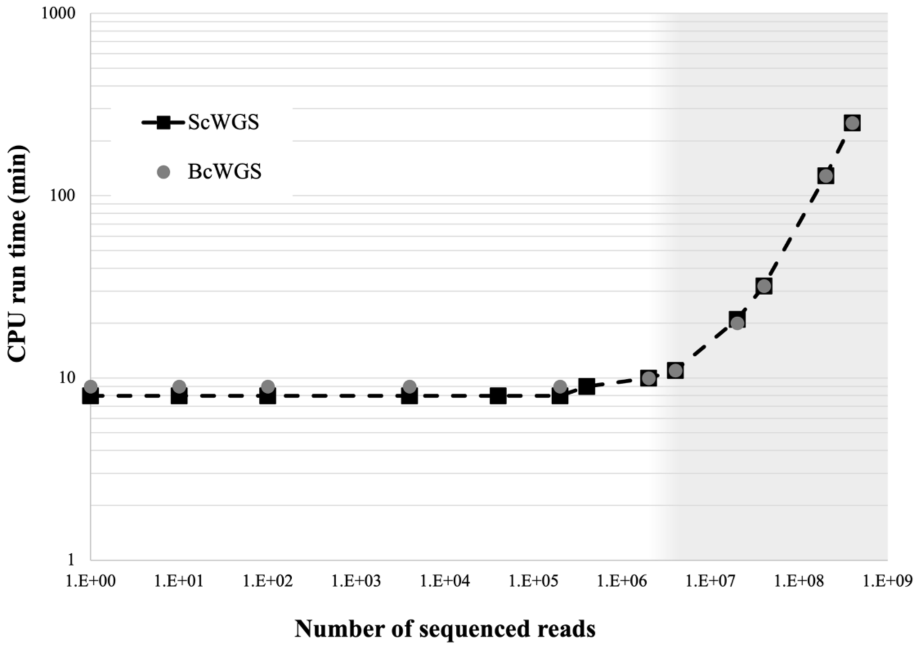

Figure 3. The CPU run time RadiSeq needed to generate different numbers of read sequences in ScWGS and BcWGS simulations are plotted separately. The plot highlights two performance regions with different shading: one where run time remains largely independent of the number of reads, and another where a non-linear dependence is observed.

Table 1. Comparison of read alignment statistics from single-cell whole genome sequencing (ScWGS) data, obtained using Samtools. Similarity scores between simulated data—generated with RadiSeq, Sandy, and NGSNGS—and the experimental data are shown, with raw values in parentheses. The highest similarity score for each metric is highlighted in bold.

	ScWGS			
Value in Samtools	Experiment	RadiSeq	Sandy	NGSNGS
Raw total sequences	2776 038	Closest simulated read count to experimental value		
Reads mapped	71%	**0.89** (79%)	0.59 (100%)	0.59 (100%)
Reads paired	100%	**1.00** (100%)	**1.00** (100%)	**1.00** (100%)
Reads properly paired	55%	**0.95** (58%)	0.18 (100%)	0.91 (50%)
Average quality	35.3	**1.00** (35.4)	0.98 (35.9)	0.95 (37.1)
Insert size average	247	0.95 (235)	**0.98** (252)	0.87 (215)
Inward-oriented pairs	28%	**0.96** (29%)	0.00 (100%)	0.54 (52%)

Table 2. Comparison of read alignment statistics from bulk-cell whole genome sequencing (BcWGS) data, obtained using Samtools. Similarity scores between simulated data—generated with RadiSeq, Sandy, and NGSNGS—and the experimental data are shown, with raw values in parentheses. The highest similarity score for each metric is highlighted in bold.

	BcWGS			
Value in Samtools	Experiment	RadiSeq	Sandy	NGSNGS
Raw total sequences	651 326 826	Closest simulated read count to experimental value		
Reads mapped	98%	**0.99** (97%)	0.98 (100%)	0.98 (100%)
Reads paired	100%	**1.00** (100%)	**1.00** (100%)	**1.00** (100%)
Reads properly paired	96%	**0.98** (94%)	0.90 (86%)	0.50 (48%)
Average quality	35.2	**0.99** (35.6)	0.98 (35.9)	0.96 (37.1)
Insert size average	360	**0.94** (338)	0.93 (335)	0.82 (296)
Inward-oriented pairs	48%	**0.98** (47%)	0.00 (99%)	0.92 (52%)

quantitatively assess the similarity between the simulated and experimental distributions shown in figure 4, we calculated the discrete Fréchet distance (Alt and Godau 1995). Within the range of our data, a value of 0 indicates perfect agreement between curves, while values approaching 140 indicate high dissimilarity. For the ScWGS data, the Fréchet distances were 0.005 for the insert size distribution and 0.185 for the coverage distribution. For the BcWGS data, the distances were 0.545 and 0.013 for the insert size and the coverage distributions, respectively. These low values demonstrate a strong similarity between the simulated and experimental results. This supports the capability of RadiSeq to replicate key sequencing metrics under

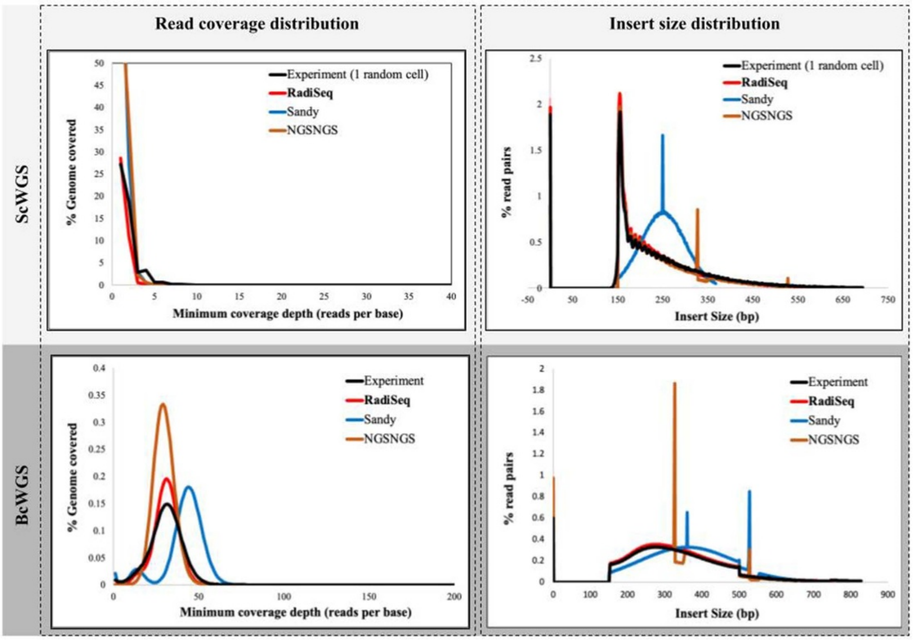

Figure 4. Read coverage distributions (left column) and insert size distributions (right column) for both experimental and simulated data are shown for single-cell whole-genome sequencing (ScWGS, top row) and bulk-cell whole-genome sequencing (BcWGS, bottom row). Simulated data generated using RadiSeq (red), Sandy (blue), and NGSNGS (brown) are compared against experimentally obtained results (black) to evaluate the fidelity of each simulation tool.

realistic experimental conditions. While agreement could be further improved by fine-tuning simulation parameters, accurately reproducing biological and biochemical variability will remain a challenge.

The simulation tools used for benchmarking RadiSeq—Sandy and NGSNGS—were generally intuitive, accessible, and well-documented for simulating the sequencing of sham-irradiated cells. Of the two, NGSNGS offered greater customizability, such as allowing specification of fragment size distribution, which was not possible in Sandy. However, neither tool supported the simulation of GC bias, limiting their ability to fully replicate experimental conditions. As reflected in tables 1 and 2 and the distributions in figure 4, NGSNGS showed better agreement with experimental data than Sandy, likely due to its higher degree of customization. Nonetheless, RadiSeq outperformed both tools in its ability to replicate experimental results. Lower similarity scores in key alignment metrics—such as the number of reads mapped, properly paired reads, and inward-oriented pairs—highlight limitations in the error models of Sandy and NGSNGS. These shortcomings restrict their utility for generating ground truth data for experimental work, although they do not affect their utility for bioinformatics verifications, which, in contrast to RadiSeq, is what they were designed for. RadiSeq, by incorporating a broader range of sequencing error mechanisms, offers a more faithful simulation of real sequencing experiments, making it a more powerful and reliable tool for such applications, including post-irradiation sequencing. In terms of computational performance, NGSNGS was the fastest simulator, completing the simulation in 8 min using a single thread, while RadiSeq and Sandy took 14 and 19 min, respectively, under identical conditions.

Limitations

Overall, RadiSeq demonstrates reasonable computational performance, and acceptable accuracy in simulating read data, closely resembling experimental data. However, RadiSeq has several important limitations.

First, RadiSeq currently only simulates Illumina sequencing chemistry. Although Illumina technology is the most widely used high-throughput NGS technology, incorporating other sequencing technologies into the RadiSeq framework as customizable options is desirable and planned for future development.

Second, RadiSeq employs estimations and simplifications of complex biological processes for various model designs, particularly for the MDA amplification model, GC bias model, and the model for read chimeras. As described in the supplementary materials, these models are rather rudimentary in the current release of RadiSeq and can be enhanced to incorporate greater complexities.

Third, there is a trade-off between simulation tuning and CPU run time. RadiSeq parameters can be fine-tuned to better match experimental data in order to improve simulation accuracy, but increased accuracy necessitates a longer completion time.

Fourth, the accuracy of RadiSeq in simulating the sequencing of IR-damaged cells has not yet been validated. Doing so will require a range of experimental irradiation conditions with contrasting sequencing outcomes in which experiment and simulation can be compared. Ongoing work by our research group is investigating the variation of sequencing outcomes as a function of radiation dose and radiation type (e.g. photons versus neutrons). Nevertheless, RadiSeq uniquely accepts SDD files containing simulated double-stranded radiation-damaged DNA data and it is thus ready to predict experimental outcomes for post-irradiation DNA sequencing.

Lastly, we acknowledge that while our simplistic technique of folding a damaged geometric genome onto a reference genome is useful for our purpose, it cannot capture the intricacies of mutagenesis. Indeed, simulation of mutagenesis is outside the scope of RadiSeq and it should be modeled upstream in the MC irradiation pipeline such that the mutations are already encoded in the SDD file that is used as input to RadiSeq. RadiSeq will not attempt to repair the DNA (to see how RadiSeq handles strand breaks, for instance, see supplementary material). We note that recent work by the TOPAS-nBio collaboration, in developing the MEDRAS (McMahon and Prise 2021) and DaMaRiS (Warmenhoven *et al* 2020) packages that model DNA repair, is an important step towards generating mutations from DNA damage. Future work will incorporate the output of these packages into ongoing validation efforts. While RadiSeq remains independent of upstream SDD file generation, users can adapt the input, such as by using repair-incorporated SDD files or genomes with artificially introduced mutations, based on their experimental goals.

5. Conclusion

We have developed and released RadiSeq, a first-of-its-kind user-friendly framework to model the single-cell and BcWGS of IR-damaged DNA. It provides the radiation biology community with a modular and customizable framework that can be used and built upon to aid in experimental design and outcome prediction for post-irradiation sequencing experiments.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/kildealab/RadiSeq.

Acknowledgment

We gratefully appreciate the insights and support of Stephan O’Brien (McGill Space Institute) on the computational aspects of this work. Some of the in-built Illumina sequencer error profiles and the read generation methods used in RadiSeq were adopted from the ART Illumina sequencing tool published open-source by the US National Institute of Health (Huang *et al* 2012). We sincerely thank the authors Weichun Huang *et al* for making the ART toolkit available under an open-source license. This research was enabled in part by computational resources provided by Calcul Quebec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

Funding

This work was supported by the Collaborative Research and Training Experience grant entitled Responsible Health and Healthcare Data Science (SDRDS) of the Natural Sciences and Engineering Research Council; Canadian Space Agency Grant [19FAMCGB25 to JK]; Discovery Grant of Natural Sciences and Engineering Research Council of Canada held by JK; and a Fonds de Recherche du Québec—Santé Doctoral Training Award to FM.

ORCID iDs

F Mathew 0000-0003-2611-2898
Luc Galarneau 0000-0003-1871-9331
J Kildea 0000-0002-7084-1425

References

Adewoye A B, Lindsay S J, Dubrova Y E and Hurles M E 2015 The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline *Nat. Commun.* **6** 6684

Agostinelli S *et al* 2003 Geant4—a simulation toolkit *Nucl. Instrum. Methods Phys. Res. A* **506** 250–303

Alosaimi S *et al* 2020 A broad survey of DNA sequence data simulation tools *Brief. Funct. Genom.* **19** 49–59

Alt H and Godau M 1995 Computing the Fréchet distance between two polygonal curves *Int. J. Comput. Geom. Appl.* **5** 75–91

Behjati S *et al* 2016 Mutational signatures of ionizing radiation in second malignancies *Nat. Commun.* **7** 12605

Bernal M A and Liendo J A 2009 An investigation on the capabilities of the PENELOPE MC code in nanodosimetry *Med. Phys.* **36** 620–5

Bernal M A, Sikansi D, Cavalcante F, Incerti S, Champion C, Ivanchenko V and Francis Z 2013 An atomistic geometrical model of the B-DNA configuration for DNA–radiation interaction simulations *Comput. Phys. Commun.* **184** 2840–7

Bertolet A *et al* 2022 Impact of DNA geometry and scoring on Monte Carlo track-structure simulations of initial radiation-induced damage *Radiat. Res.* **198** 207–20

Chandra R 2001 *Parallel Programming in OpenMP* (Morgan Kaufmann)

Chen Y-C, Liu T, Yu C-H, Chiang T-Y and Hwang C-C 2013 Effects of GC bias in next-generation-sequencing data on de novo genome assembly *PLoS One* **8** e62856

Escalona M, Rocha S and Posada D 2016 A comparison of tools for the simulation of genomic next-generation sequencing data *Nat. Rev. Genet.* **17** 459–69

Faddegon B, Ramos-Méndez J, Schuemann J, McNamara A, Shin J, Perl J and Paganetti H 2020 The TOPAS tool for particle simulation, a Monte Carlo simulation tool for physics, biology and clinical research *Phys. Med.* **72** 114–21

Ferrari A, Fasso A, Sala P R and Ranft J 2005 *FLUKA: A Multi-Particle Transport Code* CERN-2005-10 (CERN) (https://doi.org/10.5170/CERN-2005-010)

Friedland W, Dingfelder M, Kundrát P and Jacob P 2011 Track structures, DNA targets and radiation effects in the biophysical Monte Carlo simulation code PARTRAC *Mutat. Res.* **711** 28–40

Gawad C, Koh W and Quake S R 2016 Single-cell genome sequencing: current state of the science *Nat. Rev. Genet.* **17** 175–88

Henriksen R A *et al* 2023 NGSNGS: next-generation simulator for next-generation sequencing data *Bioinformatics* **39** 041

Huang W *et al* 2012 ART: a next-generation sequencing read simulator *Bioinformatics* **28** 593–594

Incerti S, Baldacchino G, Bernal M, Capra R, Champion C, Francis Z and Guèye P 2010 The geant4-DNA project *Int. J. Model. Simul. Sci. Comput.* **01** 157–78

Kageyama S-I *et al* 2021 Identification of the mutation signature of the cancer genome caused by irradiation *Radiother. Oncol.* **155** 10–16

Langmead B and Salzberg S L 2012 Fast gapped-read alignment with Bowtie 2 *Nat. Methods* **9** 357–9

Lasken R S 2009 Genomic DNA amplification by the multiple displacement amplification (MDA) method *Biochem. Soc. Trans.* **37** 450–3

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R 2009 The sequence alignment/map format and SAMtools *Bioinformatics* **25** 2078–9

Lipman D J and Pearson W R 1985 Rapid and sensitive protein similarity searches *Science* **227** 1435–41

Lu N, Qiao Y, Zuhong L and Jing T 2023 Chimera: the spoiler in multiple displacement amplification *Comput. Struct. Biotechnol. J.* **21** 1688–96

Mathew F and Kildea J 2024 *Kildealab/RadiSeq: RadiSeq_v2.0* (Zenodo) (https://doi.org/10.5281/ZENODO.13737532)

Mathew F, Manalad J, Yeo J, Galarneau L, Ybarra N, Wang Y C, Tonin P N, Ragoussis I and Kildea J 2023 Single-cell DNA sequencing-a potential dosimetric tool *Radiat. Protect. Dosim.* **199** 2047–52

McMahon S J and Prise K M 2021 A mechanistic DNA repair and survival model (Medras): applications to intrinsic radiosensitivity, relative biological effectiveness and dose-rate *Front. Oncol.* **11** 689112

McNamara A L *et al* 2018 Geometrical structures for radiation biology research as implemented in the TOPAS-nBio toolkit *Phys. Med. Biol.* **63** 175018

Milhaven M and Pfeifer S P 2023 Performance evaluation of six popular short-read simulators *Heredity* **130** 55–63

Miller T L*et al* 2023 Sandy: A user-friendly and versatile NGS simulator to facilitate sequencing assay design and optimization *bioRxiv* 2023–08

Montgomery L, Lund C M, Landry A and Kildea J 2021 Towards the characterization of neutron carcinogenesis through direct action simulations of clustered DNA damage *Phys. Med. Biol.* **66** 205011

Montgomery L and Manalad J 2022 *McGillMedPhys/topas_clustered_dna_damage: Indirect Damage Scoring Update* (Zenodo) (https://doi.org/10.5281/ZENODO.6972469)

Murray V, Hardie M E and Gautam S D 2019 Comparison of different methods to determine the DNA sequence preference of ionising radiation-induced DNA damage *Genes* **11** 8

National Research Council of Canada. Metrology Research Centre. Ionizing Radiation Standards 2021 *EGSnrc: Software for Monte Carlo Simulation of Ionizing Radiation* (National Research Council of Canada) (https://doi.org/10.4224/40001303)

Nguyen V, Panyutin I V, Panyutin I G and Neumann R D 2016 A genomic study of DNA alteration events caused by ionizing radiation in human embryonic stem cells via next-generation sequencing *Stem Cells Int.* **2016** 1346521

Niita K, Sato T, Iwase H, Nose H, Nakashima H and Sihver L 2006 PHITS—a particle and heavy ion transport code system *Radiat. Meas.* **41** 1080–90

Normand R and Yanai I 2013 An introduction to high-throughput sequencing experiments: design and bioinformatics analysis *Methods in Mol.Biol.* **1038** 1–26

Pommerenke C, Geffers R, Bunk B, Bhuju S, Eberth S, Drexler H G and Quentmeier H 2016 Enhanced whole exome sequencing by higher DNA insert lengths *BMC Genom.* **17** 399

Sakata D *et al* 2020 Fully integrated Monte Carlo simulation for evaluating radiation induced DNA damage and subsequent repair using Geant4-DNA *Sci. Rep.* **10** 20788

Schuemann J *et al* 2019b A new standard DNA damage (SDD) data format *Radiat. Res.* **191** 76–92

Schuemann J, McNamara A L, Ramos-Méndez J, Perl J, Held K D, Paganetti H, Incerti S and Faddegon B 2019a TOPAS-nBio: an extension to the TOPAS simulation toolkit for cellular and sub-cellular radiobiology *Radiat. Res.* **191** 125–38

Shumate A *et al* 2020 Assembly and annotation of an Ashkenazi human reference genome *Genome Biol.* **21** 1–18

Warmenhoven J W, Henthorn N T, Ingram S P, Chadwick A L, Sotiropoulos M, Korabel N, Fedotov S, Mackay R I, Kirkby K J and Merchant M J 2020 Insights into the non-homologous end joining pathway and double strand break end mobility provided by mechanistic in silico modelling *DNA Repair* **85** 102743

Xu C 2018 A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data *Comput. Struct. Biotechnol. J.* **16** 15–24

Youk J *et al* 2024 Quantitative and qualitative mutational impact of ionizing radiation on normal cells *Cell Genom.* **4** 100499

Zhao M, Liu D and Qu H 2017 Systematic review of next-generation sequencing simulators: computational tools, features and perspectives *Brief. Funct. Genom.* **16** 121–8