Long read sequencing applied to a multiscale analysis of biological effects induced by ionizing radiation and metal oxide nanoparticle

Pierre Beaudier

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ

SPÉCIALITÉ : BIOINFORMATIQUE

Par Pierre BEAUDIER

Séquençage de 3ème génération appliqué à une analyse multi-échelles des effets biologiques induits par les rayonnements ionisants et les nanoparticules d'oxyde métallique

Sous la direction de : Hervé SEZNEC

Soutenue le 20 mars 2023

Membres du jury :

Mme. NIKOLSKI, Macha	Directrice de recherche, CNRS	Présidente
Mme. ADAM-GUILLERMIN, Christelle	Directrice de recherche, IRSN	Rapporteur
M. GALAS, Simon	Professeur, Université Montpellier	Rapporteur
M. LAURENT, Patrick	Chargé de recherche, Université Bruxelles	Examinateur
M. SEZNEC, Hervé	Directeur de recherche, CNRS	Directeur
M. DUPUY, Denis	Chargé de recherche, INSERM	Membre invité

Remerciements

Aujourd'hui ma thèse s'est achevée. Ou peut-être hier, je ne sais pas. J'ai reçu un message du jury : "Manuscrit accepté. Chômage demain. Sentiments distingués." Cela ne veut rien dire. C'était peut-être hier.

Je tiens tout d'abord à remercier les membres du jury, Mme. Macha Nikolski, M. Simon Galas, Mme. Christelle Adam-Guillermin et M. Patrick Laurent, pour le temps qu'ils ont consacré à l'évaluation de mes travaux de thèse.

Un très grand merci à mes deux directeurs de thèse Hervé Seznec et Denis Dupuy pour leur encadrement au cours de ces années de thèse. Merci Hervé pour tout ce que tu m'as appris que ce soit sur la radiobiologie ou sur la culture bretonne. Je suis convaincu que tu deviendras un bioinformaticien confirmé en un rien de temps en mon absence et que les disques durs de l'équipe ne craindront rien. Merci Denis pour tes conseils et suggestions qui m'ont été d'une aide précieuse pour progresser dans mes analyses tout comme dans ma méthode de travail.

Je remercie Philippe, Guillaume et Laurent de l'équipe iRiBio pour tout leur travail qui a permis de produire une grande partie des données sur lesquelles j'ai travaillé et a ainsi permis de faire progresser ma thèse malgré toutes les intempéries rencontrées. Merci également au reste de l'équipe iRiBio auprès de qui j'ai eu le plaisir de travailler et d'apprendre (Franck, Sara, Nathalie, Claire, etc.)

Merci également aux membres de l'équipe de Denis (Camille, Delphine, Sara) dont la présence a égayé les murs gris et moroses de l'IECB par toutes les discussions que nous avons pu avoir.

Je remercie également tous les coureurs de l'AS CENBG passés comme présent (Guillaume, Xalbat, Fred, Guillaume2, Vincent, Antoine, etc.) Faire du sport ensemble durant ces années a été un super moment et bénéfique aussi bien physiquement que mentalement. Merci également à tous les autres doctorants et post-docs du LP2iB pour les bons moments passés ensemble au labo comme en dehors.

Merci enfin à mes parents, au reste de ma famille et à mes amis qui ont cru en moi et m'ont soutenu durant toutes ces années de thèse.

Abstract

DNA/RNA sequencing methods have undergone swift technological progress and are now major analysis tools. Their use allows the study of the cellular impact caused by exposure to environmental factors from the DNA sequence to the expression of genes. This technology is applied here in the particular case of the study of radiation and nano-induced damage. However, due to the particular and random nature of the physicochemical interactions of ionizing radiation (IR) and nanoparticles (NPs) with living organisms at different molecular and cellular scales (heterogeneity of energy deposition, heterogeneity of NPs internalization), it remains difficult to define precisely the radio- and nano-induced mechanisms. Thus, the opportunity to combine targeted and controlled micro-irradiation experiments, quantitative chemical micro-analysis and Monte Carlo simulations/modeling (Geant4/Geant4DNA) allows to better characterize the doses delivered at the cellular level in *in vitro* and *in vivo* conditions.

In this interdisciplinary context, my thesis project consisted in integrating all the tools and methods necessary for the application of a 3rd generation sequencing technology (MinION, Oxford Nanopore Technologies) in order to study the consequences on DNA and RNA of radiation and nano-induced exposures.

I introduced and validated bioinformatics tools that allowed: (i) to exploit the potentialities of long-read sequencing on reference DNA molecules in order to quantify radiation-induced fragmentation and to compare these experimental data with Monte Carlo simulation data; (ii) to analyze the transcriptomic responses of *Caenorhabditis elegans* nematode populations selectively irradiated in controlled dose (gonad progenitor stem cells) or exposed to titanium dioxide nanoparticles; (iii) to evaluate on genetically characterized sarcoma lines, the cellular responses induced in vitro by combined exposures to ionizing radiation and metal oxide nanoparticles; (iv) to approach the study of the transcriptome at the single-cell level ("Single-Cell RNA-Seq") with the objective of future applications to the analysis of radio- and nano-induced responses at the organism's cell level.

Keywords: Radiobiology, Nanotoxicology, Sequencing, Long-read, Multiscale, Single-cell

Résumé

Les méthodes de séquençage ADN/ARN ont connu des progrès technologiques fulgurants et sont aujourd'hui des outils majeurs d'analyse. Leur utilisation permet ainsi d'étudier l'impact cellulaire causé par l'exposition à des facteurs environnementaux depuis la séquence de l'ADN jusqu'à l'expression des gènes. Cette technologie est appliquée ici au cas particulier de l'étude des dommages radio- et nano-induits. Cependant, de par la nature particulière et aléatoire des interactions physico-chimiques des rayonnements ionisants (RI) et des nanoparticules (NPs) avec les organismes vivants aux différentes échelles moléculaires et cellulaires (hétérogénéité du dépôt d'énergie, hétérogénéité de l'internalisation des NPs), il reste difficile de définir précisément les mécanismes biologiques induits. Ainsi, l'opportunité de combiner des expériences de micro-irradiation ciblée et contrôlée, de micro-analyse chimique quantitative et des simulations/modélisations Monte Carlo (Geant4/Geant4DNA) permet de mieux caractériser les doses délivrées à l'échelle cellulaire en conditions *in vitro* et *in vivo*.

Dans ce contexte interdisciplinaire, mon projet de thèse a consisté à intégrer l'ensemble des outils et méthodes nécessaires à l'application d'une technologie de séquençage de 3ème génération (MinION, Oxford Nanopore Technologies) afin d'étudier les conséquences sur l'ADN et sur l'ARN d'expositions radio- et nano-induites.

J'ai ainsi introduit et validé les outils bio-informatiques qui m'ont permis : (i) d'exploiter les potentialités du séquençage « long-read » sur des molécules ADN de référence afin de quantifier la fragmentation radio-induite et de confronter ces données expérimentales aux données de simulation Monte Carlo ; (ii) d'analyser les réponses transcriptomiques de populations de nématodes *Caenorhabditis elegans* sélectivement irradiées en dose contrôlée (cellules souches progénitrices des gonades) ou bien exposées à des nanoparticules de dioxyde de titane ; (iii) d'évaluer sur des lignées de sarcomes génétiquement caractérisées, les réponses cellulaires induites *in vitro* par des expositions combinées aux rayonnements ionisants et aux nanoparticules d'oxydes métalliques ; (iv) d'aborder l'étude du transcriptome à l'échelle de la cellule unique (« Single-Cell RNA-Seq ») dans l'objectif de futures applications à l'analyse des réponses radio- et nano-induite à l'échelle de la cellule d'un organisme.

Mots-clés : Radiobiologie, Nanotoxicologie, Séquençage, « Long-read », Multi-échelles, Cellule unique

Table of Contents

Introduction

Ionizing radiations (IR) are subatomic particles or electromagnetic waves of sufficient energy (13.6 eV) to ionize atoms along their trajectory, their interactions with living matter can cause direct damage (ionization of target) or indirect damage (ionization of intermediary molecule resulting in highly reactive products, *e.g.*, water radiolysis). They are naturally present in the environment (cosmic rays, natural radioactivity, etc.) but can also originate from human sources (nuclear detonations, industries, etc.) which led to the creation of public institutions dedicated to the surveillance of IR exposure and the associated radio-induced effects. Several categories of population are also routinely exposed to increased doses of IR and require additional monitoring: nuclear workers, nuclear medicine personnel, pilots, astronauts, etc.

However, the models used for computing the risk associated to a given dose of IR exposure mainly take in consideration the risks of DNA breakage ("targeted effects"), mostly because of the risk of long-term tumor development in case of non-repair, and tend to neglect the other aspects of potential radio-induced damage. This historical focus has led to a lack of characterization of these ancillary damage on other potential targets as well as the consequences they could cause on cellular function and survival, especially for organisms routinely exposed to low doses of exposures.

Similarly, metal oxide nanoparticles (NPs), due to their massive use in industry and subsequent deposition in nature, constitute a regular low dose exposure in the population. If the dangerousness of these NPs is characterized during internalization in a cell, the precise cellular mechanisms differ according to the chemical composition and the shape of the NPs, thus influencing the cellular response.

For both physical agents, the magnitude of the cellular response to their exposure is dose-dependent. However, the physical deposition of these agents could be random considering the subcellular scale, since the IR-matter interactions as well as the internalization and the fate of the NPs do not follow a deterministic process. It is therefore complex to establish a defined cellular response to this stress for a given exposure dose.

In the focus of achieving a complete characterization of the radio- and nano-induced damage down to the individual cell scale, the goal of my thesis was thus to integrate modern DNA and RNA sequencing techniques in the interdisciplinary research environment of the iRiBio team working on the interactions between IR, metal oxide nanoparticles and living matter. Micro-irradiation and chemical imaging techniques by charged-particle microbeam have been developed within the team over the last decade in order to achieve analytical precision at the scale of the cell or sub-cellular compartments and these techniques are coupled with simulation

models of particle-matter interactions developed within the Geant4/Geant4-DNA collaboration. In this context, the objective was to introduce and validate sequencing methods as an analysis tool of the radio- and nano-induced cellular responses in complex samples to produce data that can then be used to compare and validate the codes developed in Geant4-DNA.

In this introduction, I will provide a brief summary of the history of sequencing from the inception of the technology to the current methods and what they can contribute to a biological study. I will then detail the interest of their application in the study of the effects of specific elements of contamination and environmental exposure, IR and metal oxide nanoparticles. The research carried out in the iRiBio team and how sequencing can be integrated will be detailed and I will finish by detailing the main research objectives of this thesis project.

I. A brief history of the main advances in nucleotide sequencing methods

The initial discovery of inherited traits in generations of common pea plants by Gregor Mendel in 1866 marked the first event in the foundation of what would become the science of genetics[1]. The first introduction of the "gene" term by Wilhelm Johannsen did not happen until 1909 at a time when DNA had yet to be identified as the carrier of this genetic expression[2]. This discovery did not come until much later in 1944 as a result the Avery-MacLeod-McCarty experiment which first found DNA to be the hereditary material in viruses[3]. Subsequently, James Watson and Francis Crick solved the three-dimensional structure of DNA in 1953 from crystallized X-ray structures produced by Rosalind Franklin and Maurice Wilkins[4,5]. They described it as composed of two coiled strands of nucleotides paired and linked by hydrogen bonds, with 4 different nucleotides found in two possible pairings: Adenosine (A) – Thymine (T) and Cytosine (C) – Guanine(G). This configuration revealed the structure of genes as ordered molecular patterns of nucleotides which could be "read" to identify them. The identification of messenger RNAs (mRNA) functioning as intermediaries between DNA and protein synthesis[6,7] with a similar nucleotide-based structure helped cement the "Central Dogma of Molecular Biology" explaining the flow of genetic information in a biological system[8].

Since then, the sequencing methods designed to obtain the order of nucleotides in DNA (genomics) and RNA (transcriptomics) molecules has become an integral component of genetics and more globally of biological research, and has been applied in a wide array of

research fields (medical diagnosis, virology, forensic biology and-so-on). It has been used to identify the sequence of entire genomes, detect quantitative changes in gene expression, identify organisms in complex environments and the most recent technological advances in this field makes it possible to reach new levels of precision in the definition of molecular and cellular mechanisms.

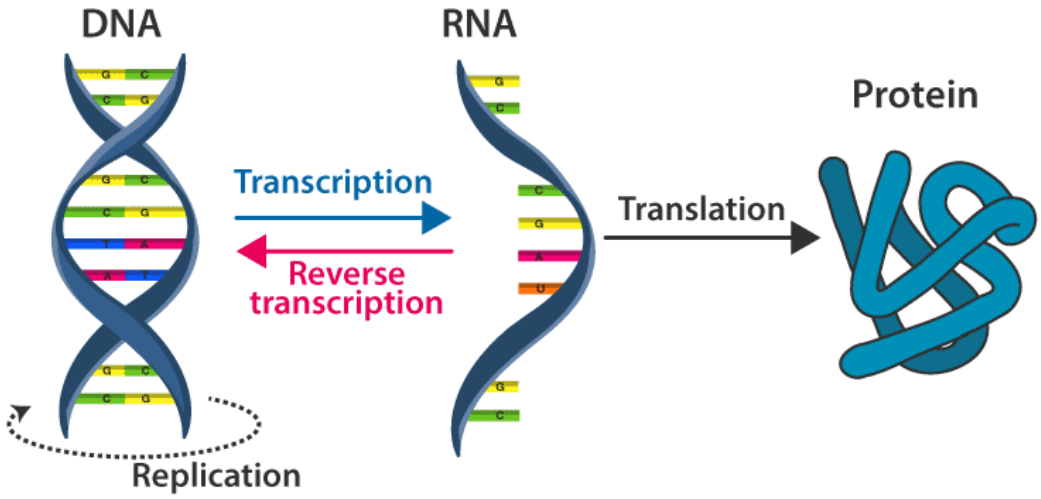


Figure 1. Scheme of the Central Dogma of Molecular Biology representing the three major classes of information-carrying biopolymers in living organisms: DNA and RNA (nucleic acids), and proteins.[9]

While the first complete protein sequencing using partition chromatography was achieved in 1952 by Frederick Sanger on insulin[10,11], the nucleotide sequencing techniques did not follow for some time because of the more important sizes as well as the important similarity between the four nucleotides[12].

The first complete sequence of a gene was obtained in 1972 by Walter Fiers by using RNAses to separate individual oligonucleotides which were then identified *via* electrophoresis and chromatography[13]. However, sequencing did not become a common experimental lab procedure until a few years later.

A) First generation sequencing

In 1977, two methods which became widely popular were published:

-The chemical cleavage method also known as “Maxam-Gilbert sequencing” and published by Allan Maxam and Walter Gilbert[14,15].

-The dideoxy chain-termination method also known as “Sanger sequencing” and published by Frederick Sanger[16,17].

While both of these techniques faced widespread use, the Maxam-Gilbert sequencing has fallen out of use, mainly due to the necessary use of large amounts of hazardous chemicals[18], unlike the Sanger sequencing which is still commonly used to this day for application that do not require high throughput and came to be remembered as the main sequencing method of the first generation of sequencing.

In this Sanger sequencing method, a short synthetic primer of complementary sequence to the region of interest is made to hybridize to its known starting point and will direct the DNA polymerase to synthetize a new single DNA strand of the primer 3’ hydroxyl group by incorporating nucleotides. This reaction is done in four reaction buffers containing normal deoxyribonucleotides (dATP, dTTP, dGTP and dCTP) as well as a small proportion of one of the di-deoxyribonucleotides (ddATP, ddTTP, ddGTP and ddCTP) which are undistinguishable to the DNA polymerase and will get stochastically incorporated in the synthetized strand. These ddNTPs lack the 3’ hydroxyl group which effectively terminates the DNA synthesis process[19]. These new DNA strands’ outputs are extracted and placed on electrophoresis gels for migration, a simple lining up of fragments per length then reveals the nucleotide sequence of the targeted DNA molecule.

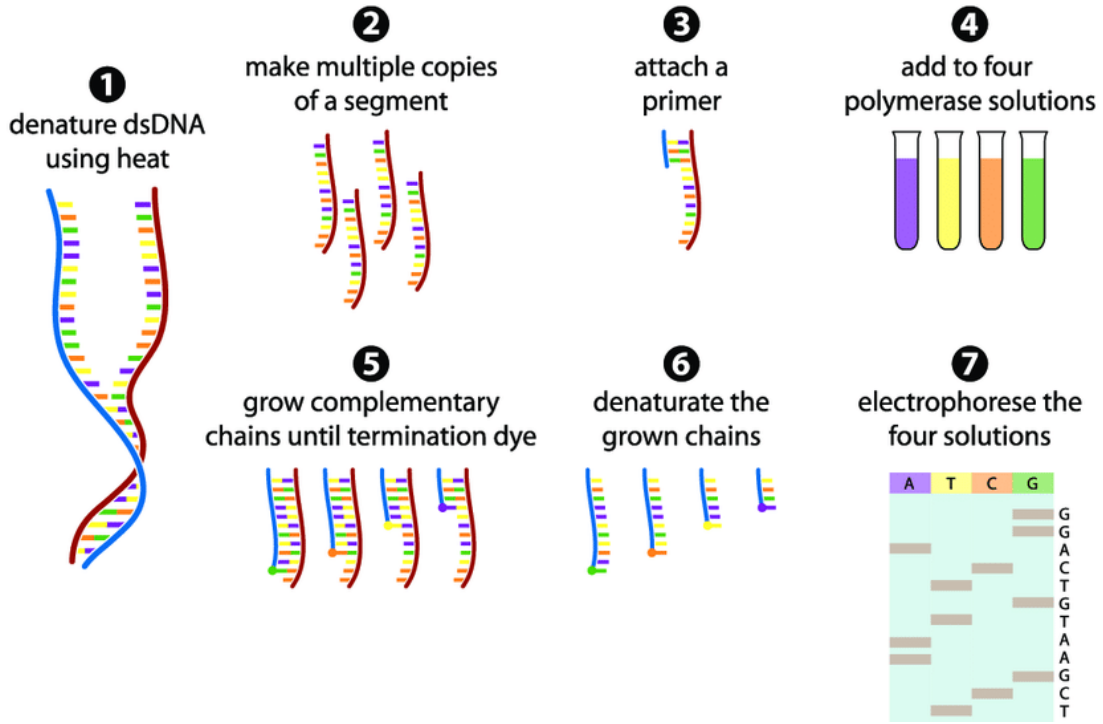


Figure 2. Scheme description of Sanger sequencing[20].

This technique has proven to be very reliable for reads lengths around 300-1000 bp with an error rate of 0.001% [21], excluding the first 15-40 bp due to primer binding, and remains to this day as the gold standard of sequencing accuracy[22,23].

However, the read lengths limitations meant that strategies had to be developed for the sequencing of whole genomes. The most popular of these and which is still currently used is the "shotgun sequencing", a method in which the DNA molecule is randomly fragmented in small segments to sequence individually, the reads produced are then re-assembled into contiguous sequences using their overlapping ends to reconstruct the original DNA molecule[24]. Major projects such as the Human Genome Project[25,26] made use of this method in combination with Sanger sequencing to process larger genomes, but faced technical limitations particularly on long repetitive sequences which proved to be too confusing to piece together using assembly methods and were thus left as gaps in the final genomes[27].

To sum up, the first generation of sequencing methods offered a near-perfect accuracy on reads of relatively short lengths but faced limitations to scale up the amount of DNA fragments sequenced due to electrophoresis time constraints as well as the total cost per nucleotide, the first human genome having cost around $3 billion and taken 13 years[28].
New technologies emerged during the 1990s to meet these challenges and pushed the field of sequencing towards its second revolution, with the introduction of several high-throughput methods capable of rapidly sequencing multiple DNA molecules at a time, which came to be known as "Massive parallel sequencing" or "Next-Generation Sequencing" (NGS)[29].

B) Second generation sequencing

The sequencing methods of this generation can be grouped in two major categories.

- Sequencing by hybridization[30]: known DNA fragments are hybridized on the DNA molecule to sequence in repeated cycles where non-hybridized DNA is washed away. The final DNA molecule is then reconstructed using overlaps between fragments.
- Sequencing by synthesis (SBS)[31,32]: techniques making use of the DNA polymerase to synthetize a new DNA strand, similarly to Sanger sequencing, and the sequence of which is then analyzed.

1/ Sequencing by hybridization

Sequencing by hybridization, initially developed in the 1980s, found some use in specific contexts[33,34] like the identification of disease-related SNPs (Single Nucleotide Polymorphisms) and chromosome abnormalities but has mostly been displaced by other methods, and has not found widespread use unlike alternative SBS methods which were at the heart of the second-generation sequencing, the main two being Pyrosequencing and Illumina dye sequencing

2/ Sequencing by synthesis

a) Pyrosequencing

Pyrosequencing, first described by Bertil Pettersson, Mathias Uhlen and Pål Nyren in 1993 [35,36], became the first popular method of this second wave of sequencing technologies when it was commercialized in 2005 by 454 Life Sciences. Contrarily to previous SBS methods, the DNA sequence is not inferred by electrophoresis migration but rather by luminescence measuring during pyrophosphate synthesis.

The inclusion of each nucleotide in the new strand by the DNA polymerase will cause the release of a nucleotide-specific pyrophosphate group PPi which is converted into ATP by an ATP sulfurylase, it is then used as the substrate for the luciferase enzyme which will produce light proportionally to the amount of pyrophosphate carried by the ATP molecule[37,38]. The resulting light flowgram (or pyrogram) is then analyzed to extract the underlying DNA sequence.

The sequencing machines produced by 454 Life Sciences (later Roche) for commercial use parallelized this process on plates containing wells[39] in which individual DNA molecules separated by emulsion PCR[40,41] (random DNA fragmentation followed by segregation of fragments using beads and insertion in droplets serving as PCR microreactors) undergo sequencing, the plates could fit 96 wells on the first iterations and up to millions of picoliter-scale wells on later versions.

The significant yield increase brought by this technique over Sanger sequencing did however come with some drawbacks in terms of sequence length (maximum of 400 bp) and error rate[42] (around 0.49% on average) but the gains in cost per genome, less than $1 million (at the time of release) for a complete human genome[43], made it very popular and increased interest for massive parallel sequencing.

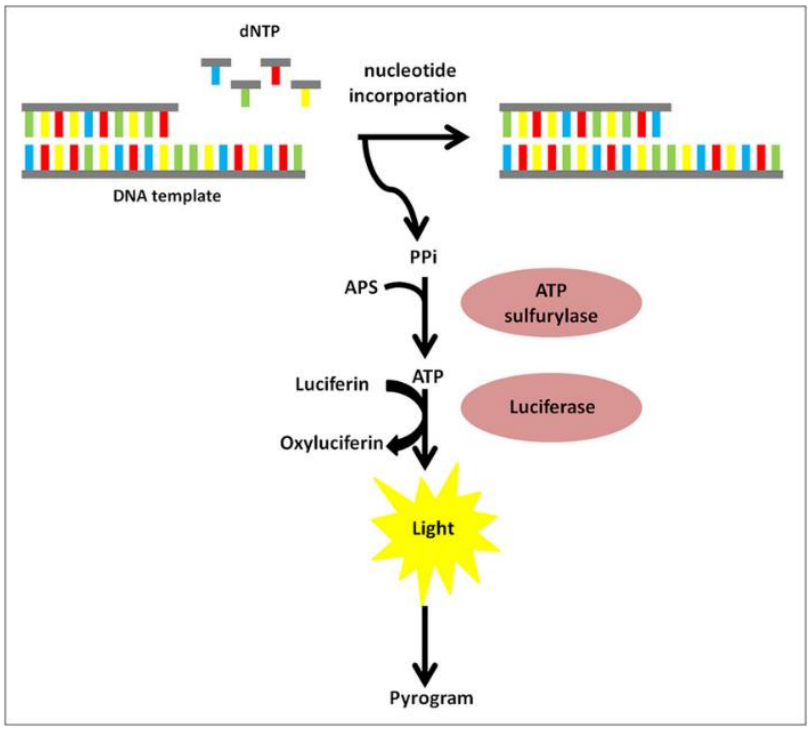


Figure 3. Scheme description of pyrosequencing.[44]

b) Illumina dye sequencing

The second major method in this generation of sequencing is the Illumina sequencing[45,46], initially developed by Solexa and Lynx Therapeutics, which relies on a process dubbed "bridge amplification"[47]. Instead of parallel emulsion PCR reactions, the DNA fragments are passed on a flow cell containing bound complementary oligonucleotides on which they will hybridize and undergo repeated rounds of amplification and washing of the reverse strands to create clusters of identical copies of the original DNA fragments. The synthesis is done in rounds using fluorescent 'reversible-terminator' dNTPs, their linked fluorophore acts as a blocking group which prevents the addition of new nucleotides in order to be effectively detected by the machine and is washed away at the next round to continue synthesis[48]. The DNA fragments sequence is then obtained from the resulting light flowgram, similarly to pyrosequencing.

This method benefitted from an overall similar error rate to pyrosequencing[49,50], but most importantly a significant cost decrease, outpacing Moore's law[51], due to the lack of necessary enzymes, other than the DNA polymerase, which are expensive and account for a major part of the sequencing run price. Multiple protocols making use of this technology have been developed over the years for varying uses and levels of throughput desired with reads size ranging from 150 to 400 bp and outputs ranging from tens of millions to hundreds of millions of reads per sequencing run[52].

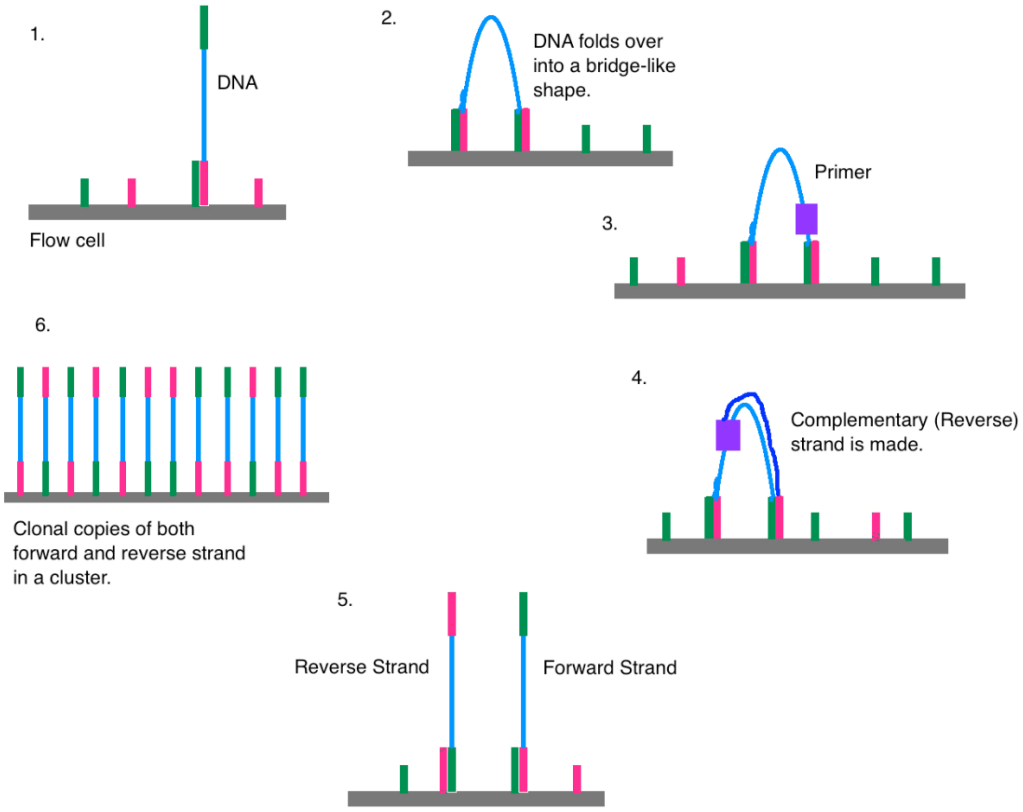


Figure 4. Scheme description of Illumina dye sequencing.[53]

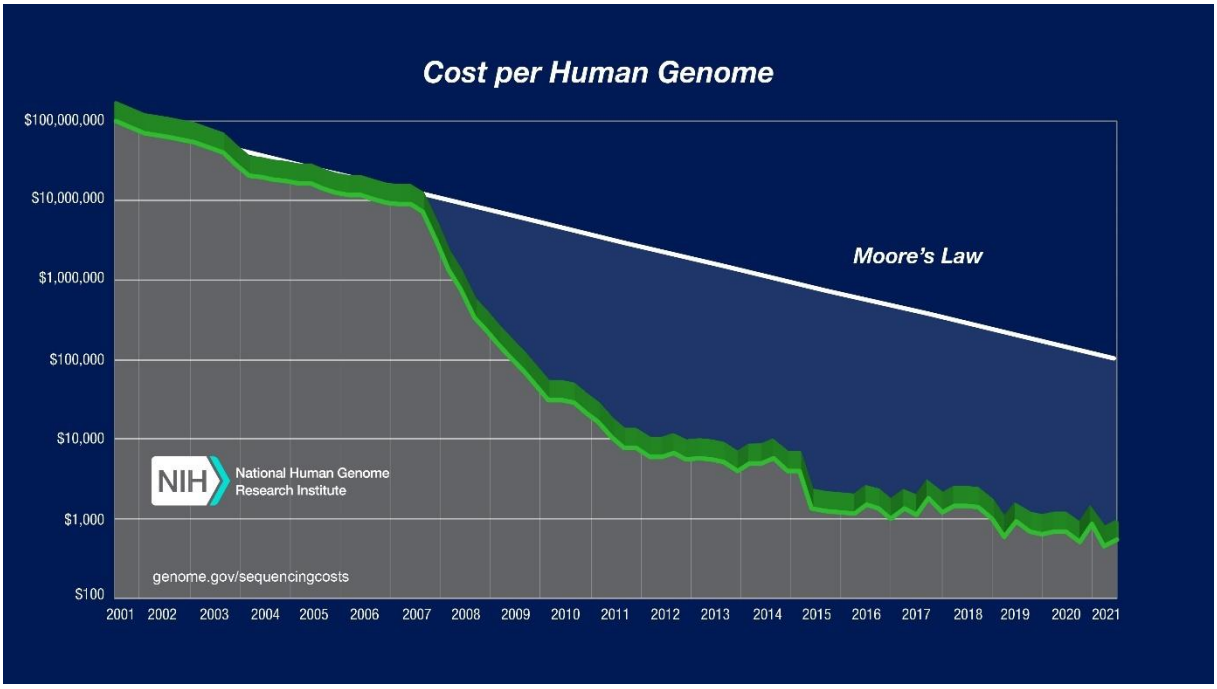


Figure 5. Cost of genome sequencing over the years compared to Moore's law.[54]

This important shift in throughput from sequencing runs have led to the production of vast amounts of data, thus bringing biology into the concept of "Big Data"[55,56] and led to its first place in total cloud storage used among other sciences. It led logically to a vital need for new performant tools to be developed to de-correlate the amount of data produced and the required time to analyze it. This need led in turn to the explosion of the bioinformatics field[57], which had first emerged in the 1960s as a tool for protein sequence analysis[58], and helped the emergence of new degrees and formations based solely on bioinformatic work[59,60], ultimately making it a staple of modern laboratories for studies using sequencing technologies.

To sum up this second generation of sequencing compared to the first generation, the throughput of sequencing runs increased drastically while becoming excessively cheaper per base sequenced, but it came at the price of a slightly lower precision. This massive increase in throughput and number of molecules sequenced not only made genome sequencing and assembly easier, but also opened the door for the analysis of transcriptomics expression through the sequencing of RNAs[61], retro-transcripted into cDNA beforehand, and most particularly the messenger RNAs (mRNA) which are the carriers of the cellular expression from the DNA to the proteins. The transcriptomics expression reveals the levels of expression for each gene, as opposed to previous methods which targeted specific genes, and can detect the statistically significant shifts in expression when the selected organism is exposed to different conditions.

One of the main limitations that persisted in this generation of sequencing was the identification of complex regions in genomes. The short-read technologies cannot adequately sequence these regions due to inherent difficulties at the chemistry level (e.g., repeated sequences, high GC content) therefore hiding their potential usefulness in genetic disease diagnostics[62]. The human genome for example, while first considered complete in 2003 through the Human Genome Project, had 8% of its sequence that had yet to be revealed.
The methods that have been developed to tackle this challenge have been considered as the third generation of sequencing[63], they have in common their focus on long-read methods (>10 000 bp) and single molecule sequencing[64] (being able to sequence a molecule without prior amplification) while also trying to keep the high-throughput of second-generation methods.

C) Third generation sequencing

Numerous methods have been developed, with the first ones emerging around 2008-2009, but two main actors are actually dominating the sector of long-read sequencing:

- Pacific Biosciences (PacBio)
- Oxford Nanopore Technologies (ONT)

1/ PacBio: Single Molecule Real Time (SMRT)

PacBio's long read technology SMRT uses a chip (or flow cell) with a metallic film, which is covered in microfabricated nano-wells called zero-mode waveguides[65,66] (ZMWs). These ZMWs guide light energy through apertures of smaller diameter than its wavelength, causing it to decay exponentially and only illuminate the bottom of the wells in which is located the DNA fragment and the DNA polymerase that will incorporate fluorescent dNTPs into a new synthetized strand. The phospho-linked fluorophores are emitted after incorporation and detected through imaging at a millisecond time scale, and then float away from the bottom of the well which make them undetectable to future imaging rounds. The time between each round can be measured as it is timed with the rate of nucleotide incorporation (inter-pulse duration, IPD) into the new strand, this provides the ability to detect multiple types of base modifications causing important shifts in IPD[67,68].

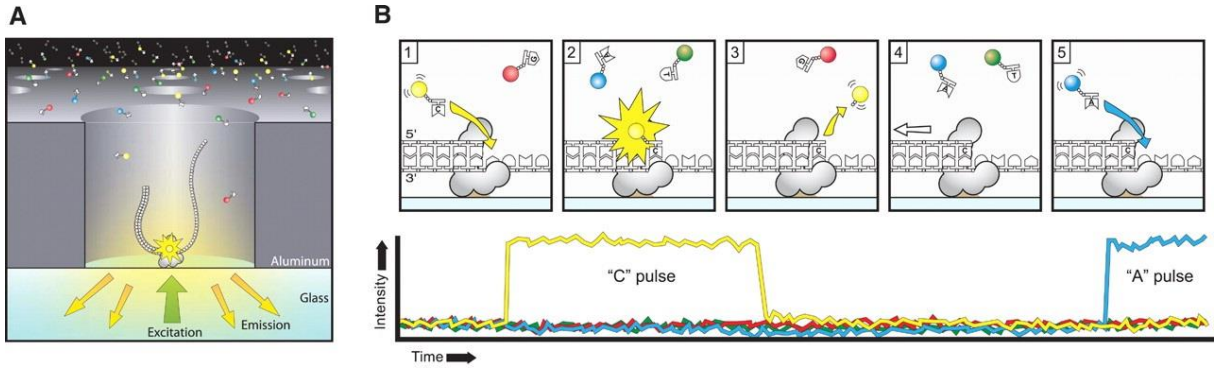


Figure 6. Scheme description of SMRT sequencing.[69]

This technology initially produced reads up to around 10 kb [70] but has been significantly improved since its release in 2011 and can now consistently hit averages read length of 10-25 kb with some reads spanning up to a potential maximum of 64 kb. The accuracy dropped severely in this technique from previous generations with an error rate of 10-15% [71], that are

fortunately randomly distributed and can be reduced through the use of consensus sequences. The introduction of HiFi[72] reads by PacBio, a Circular Consensus Sequencing[73] library preparation protocol for preparing circular reads which are sequenced in repeated passes to create final consensus sequences, has cemented this approach of error correction for SMRT sequencing. This method has reduced the error rate to 0.1-0.2% thus placing it at a similar level than short read methods.

In terms of pricing, the total protocol costs around 1600€ with ~375€ for the library preparation and ~1200€ for the chip which is single-use. It places this technique around the same levels of expense as the cheaper Illumina protocols with an overall similar error rate, lower throughput but a 100x to 250x average increase in read length[74], thus unlocking regions of the genomes that were not available to the NGS techniques.

2/ Oxford Nanopore Technologies: Nanopore sequencing

In the ONT Nanopore sequencing technology[75], the molecules to be sequenced go through a flow cell containing a membrane riddled with nano-scaled pores[76] which is traversed by a constant electric signal, an idea first emitted in the 1980s[77]. On each pore, a motor protein captures one end of the DNA or RNA molecule to be sequenced by recognizing an adapter sequence added during the library preparation, it also unzips the molecule in the case of dsDNA, and translocate it through a pore protein present on the membrane. The passage of each nucleotide will create a characteristic local disruption of the electric current, the final raw electric signal levels will then be interpreted for each pore through a process called basecalling[78] which will output the final molecule sequence. This strategy of sequencing enables the possibility to do direct-RNA sequencing[79] (sequencing RNA without prior reverse-transcription into cDNA) at a high throughput thus allowing for a more precise analysis of transcriptome expression by avoiding the eventual biases produced by reverse transcription. This technology is also ideal for the detection of base modifications, as the electric signal disruption caused by the nucleotides shift through the membrane is also impacted by small chemical modifications.

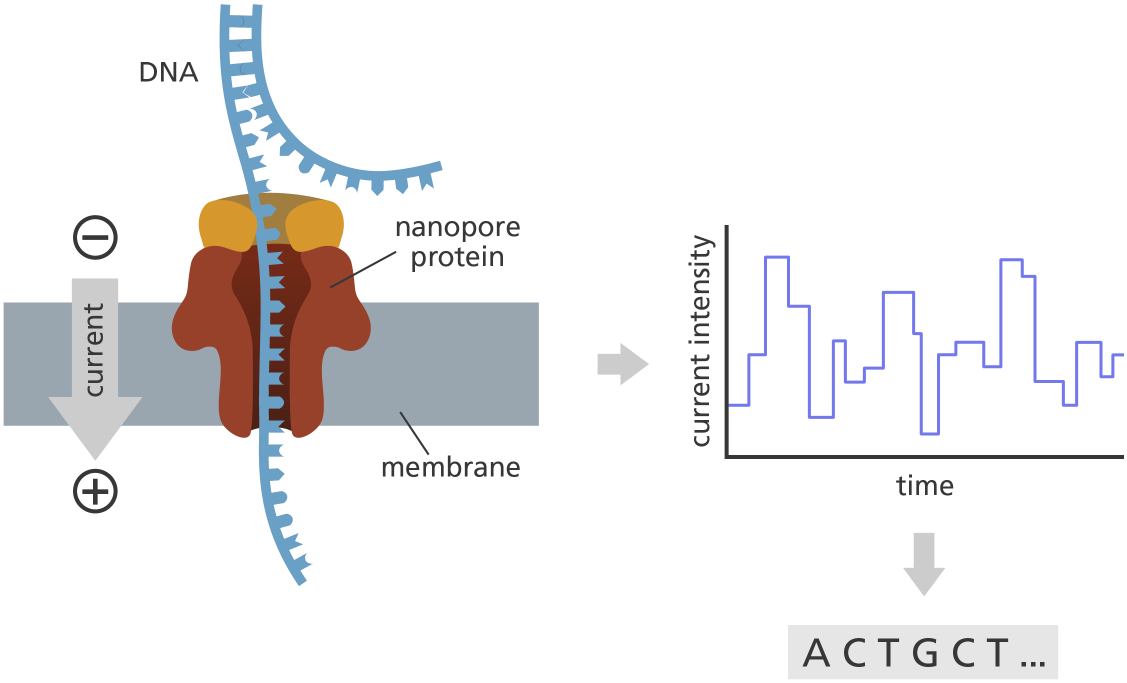


Figure 7. Scheme description of a nanopore sequencing.[80]

The interpretation of this signal remains complex however, especially for direct-RNA, as there are important levels of noise which can hide small differences between nucleotides. At the time of writing, 5-Methylcytosine on DNA is the only base modification detectable using ONT official tools[81], although multiple softwares have been developed by non-affiliated teams[82,83] affined to their personal use and are publicly available. ONT have claimed their intention to add new base modifications in the coming software releases (*e.g.*, N6-Methyladenosine, 8-Oxoguanine)[84].

At the time of release, this technology suffered from an extreme error rate going up to 40% [85]. There have been multiple upgrades to the initial method to significantly reduce this number but it remains around 6% on the current generation of flow cells[86], the most recent library preparation kits released are marketed as lowering error rates around 1% but they have yet to be thoroughly tested by the nanopore community. While this error rate is not limiting in most usages and can be circumvented by computing consensus sequence given a sufficient sequencing depth (30-50X)[87,88], it can cause issues when investigating natural variations in genotyping or haplotyping. It can also be problematic when trying to sequence homopolymers as the translocation speed is not constant and the electric signal variation are minor between identical bases, thus resulting in a difficulty to determine the exact length of homopolymers[86].

One of the main strengths of this method lies however in the read length. While the initial pores resulted in lengths similar to PacBio SMRT (~10-25 kb)[89,90], the technology upgrades make it now possible to theoretically sequence reads of any given size with a current record of 4.2 Mb for a single read[91]. This capacity to sequence entire genome in individual reads is an important step up from previous assembly methods but needs to be perfected as sequencing runs of DNA molecules in the hundreds of kb usually output at best a few complete reads, the rest of the sample being involuntarily fragmented during the process of translocation[92].

Another important advantage of this method is the very advantageous pricing as the default sequencing machine (MinION) is priced at 1000$ and easy to include in any lab due to its handheld size and low weight (90g), this small size has led to its use in various environments from jungles to the International Space Station. Different sizes of flow cells have also been released depending on the desired throughput giving a vast price range from low-cost to high-end:

- Flongle, small throughput, 126 nanopores channels, ~55€ per flow cell
- MinION/GridION, medium throughput, 512 nanopores channels, ~750€ per flow cell
- PromthION, high throughput, 2675 nanopores channels, ~1100€ per flow cell, requires however the purchase of a PromethION 24 A100 sequencing machine.

Overall, the nanopore sequencing method while still lacking in some aspects compared to other mainstream technologies offers a unique research opportunity by its singular features: read length, direct-RNA, price, ease of installation in a lab and adapted framework for base modification analysis.

In summary, the DNA/RNA sequencing field has progressed immensely since its first steps in the 1970s, becoming excessively cheaper and more effective (with the notable exception of error rate). This in turn has led to a higher accessibility of this technology in multiple scientific domains and helped to bring new strategies of analysis on scientific questions. It is in this perspective that my thesis project was designed with the project of incorporating the ONT nanopore sequencing in studies of the interactions between ionizing radiation and living matter to bring new resolution at the DNA and RNA level.

II. Applications of sequencing to study the effects of ionizing radiation and metal oxide nanoparticles (titanium oxide)

Technological advances in the sequencing methods described above have made it possible to evolve from laborious sequencing of unknown sequences to high-throughput, low-cost sequencing of predominantly known sequences in order to study differences in genomic sequences (genomics) or in levels of gene expression (transcriptomics) between different conditions, all with the aim of classifying the potential biological effects of studied factors on a given organism. Sequencing has therefore been adopted by public health authorities, for example, as a study tool to estimate the dangerousness of exposure to toxic substances by seeking to establish a link between these factors and the onset of disease[93]. The search for this link can be done from the study of mutations in the DNA sequence[94,95] to the epigenetic regulation[96] and expression of genes[97] (and by extension of proteins[98]).

Among all the existing risk factors for human health, ionizing radiation and metal oxide nanoparticles are elements whose precise impact on living matter and therefore the level of danger during exposure are still under research. However, exposure to these factors is common either because of their natural presence or because of their deliberate use in industry or medicine. We briefly describe here the nature of these elements and their mechanisms leading to the production of cellular damage:

A) Ionizing Radiation

The term ionizing radiation (IR) encompasses several electromagnetic or particulate radiation which can be categorized in two categories:

- (i) directly ionizing radiation (α particles, β rays): capable of ionizing atoms through Coulomb force depending on their kinetic energy
- (ii) indirectly ionizing radiation (photons, UV, neutrons): electrically neutral elements having no strong interactions with matter, the ionization effects are caused by secondary ionization

The physical properties of these elements are reflected in their interactions with matter which differ greatly in terms of penetrance, biological effectiveness, etc.[99]

IR sources can be of natural origin (cosmic rays, radioactivity of the soil, air, water, etc.), of "undesired" human origin (residues of nuclear detonations or nuclear incidents) or controlled for medical purposes (radiotherapy, radiology, *etc*).[100]

The possibility of fragmenting DNA, and the associated risk of tumor behavior, has led to the consideration of DNA as the primary element in the measurement of radiation-induced damage in order to qualify the hazardousness of an exposure. We therefore speak of "targeted effects" to consider damage to DNA and "non-targeted effects" for damage to targets other than DNA (this category also including the phenomena of bystander effect, genomic instability, adaptive response and hypersensitivity), this categorization forging what we call the classic dogma of radiobiology.[101]

Radiation-induced damage can also be subdivided into two categories based on the nature of the damage into "Direct damage" and "Indirect damage" which we describe here.

1/ Direct damages

The first type of damage caused by IR is due to the primary ionization events which are initiated in the 10^{-14}-10^{-12} seconds range after the energy deposit and leads to the breaking of C-H, N-H, O-H and S-H bonds at the molecular level[102]. These bond breaks can result in unstable nucleotides in DNA leading either to strand breaks or chemical modifications in bases[103]. However, these damages are not limited to DNA but can affect all other biological macromolecules, for example by breaking links between amino acids on proteins and thus affecting their structure and function[104].

2/ Indirect damages

Interactions also occur between normal metabolic free radicals and water molecules that have been excited or ionized by IR, these reactions will result in water radiolysis, the dissociation of water molecules into reactive oxygen species (ROS), the main source of indirect damage.

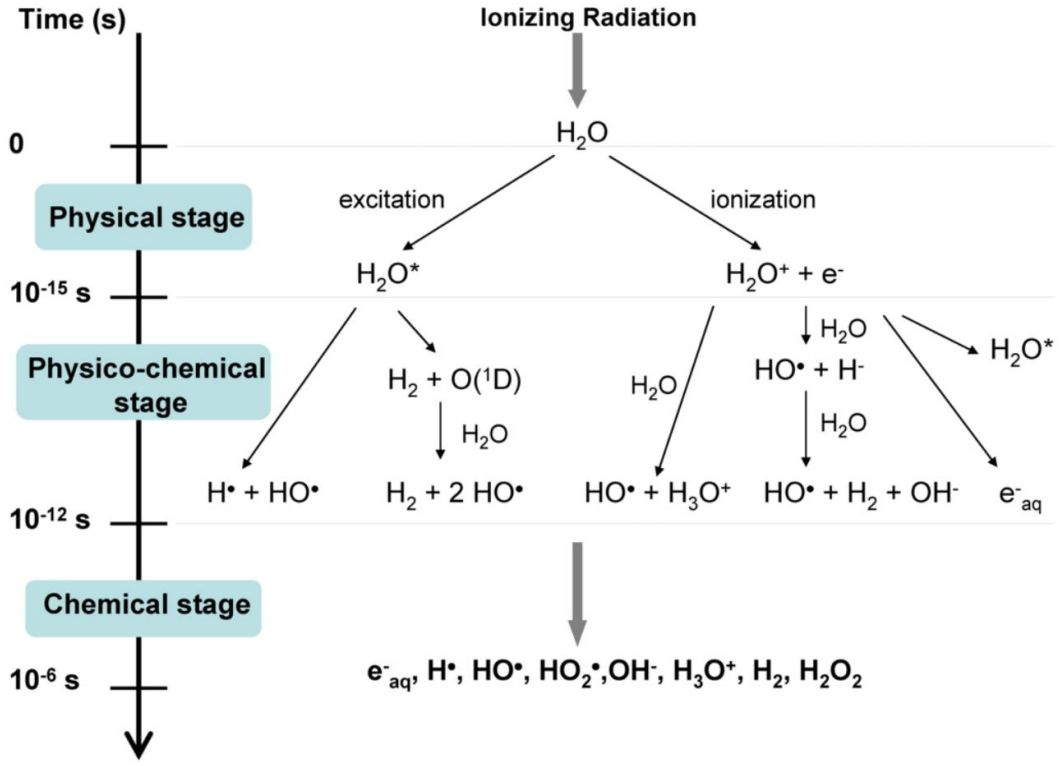


Figure 8. Main reactions occurring during the three stages of water radiolysis.[105]

B) Metal oxide Nanoparticles (Titanium oxide)

Nanoparticles (NPs) are particles of material with a diameter between 1 and 100 nm and can exist in many forms depending on their production method. Although NPs can be of natural origin (atmospheric dust, geological phenomena, etc.), the challenge of understanding the dangerousness of nanoparticles comes mainly from those produced artificially. These NPs are indeed used in plenty of industrial applications (cosmetics, food, chemicals, construction, medical, etc.[106]) which results in increasing quantities of these NPs found in the environment causing regular exposure phenomena and therefore potential health risks.

The physical and chemical properties differ significantly from larger particles resulting in significant surface reactivity. Indeed, the size of an atom being between 0.15 and 0.6 nm, there is a high surface to volume ratio in NPs thus making them more reactive than a larger particle of which a larger fraction of the volume is not located near the surface and is therefore not available to interact with the surrounding medium[107].

The wide variety of NPs types, whether by shape, size or material, translates into significant variability in bioavailability and in the amount of cellular damage caused depending on the

characteristics of the type of particle causing the exposure. However, common types of cellular damage are frequently associated with NPs: oxidative stress, genotoxic damage, endoplasmic reticulum stress, etc [108,109]. The dangerousness of these particles, however, is mitigated by their mechanism of action on the cell, as their internalization is considered necessary to cause cellular damage, although studies have shown damage across barriers.[110]

C) Towards the single-cell analysis of the cellular responses induced by IR and NPs

Numerous studies have been performed on the toxic potential for an organism to be exposed to these two factors, including sequencing methods, whether in patients[111], *in vitro*[112,113] or *in vivo*[114,115,116] models. However, a major challenge in the study of these biological damages lies in the conjugation between cellular damages and a precise quantification of the physical agents (IR, NPs). Indeed, whether it is IR or NPs, cellular deposition can vary greatly within the same organism or between several organisms. The amount of damage induced can thus vary accordingly and the cellular response to these factors may vary from one sample to another. It is with the aim of producing a controlled dose deposition that tools such as charged particle microbeams have been developed.

Charged-particle microbeams are instruments that have been designed since the 1990s to deliver defined number of ions (MeV protons and α particles) at a resolution of a few microns[117]. This accuracy allows for the targeting of individual cells and even specific sub-cellular compartments (mitochondria, nucleus, *etc*) and the study of the effects of IR exposure at a precision unattainable by broad irradiation techniques[118]. This targeting precision also allows the use of chemical imaging by nuclear microprobe methods (Scanning Transmission Ion Microscopy, Particle Induced X-ray Emission, Rutherford Backscattering Spectroscopy, *etc*) which allow to obtain a quantification of the chemical composition of the sample at a micrometric scale according to the type and energy of the charged particle used.[119]

The precision achieved by these tools allows the production of samples in which a single cell or a reduced group of cells will be exposed to the studied factor at an equivalent dose. The use of modern sequencing methods combined with this technology thus offers an unprecedented possibility to study radio and nano-induced damage at the single-cell level, in particular *via* the use of "Single-Cell RNA-Seq" transcriptomic analysis methods in order to obtain the cellular response specific to the irradiated cell or to the cells exposed to NPs.

III. Research context within the iRiBio team

In order to reach the radio- and nano-induced response at the single cell scale, three main tools have been developed within the iRiBio team, each designed to solve a given problematic:

- Nuclear microprobe to perform cellular irradiation at a micrometric scale
- Nuclear microprobe for quantitative chemical analysis at a micrometric scale
- Geant4/Geant4-DNA for characterization of dosimetry at the micrometric scale ("microdosimetry")

Several biological models at different levels of complexity are studied in the team's research projects and need to be adapted for use on these tools depending on the scale (molecular, *in cellulo*, *in vivo*). These micrometric scale tools and the biological models used in the various research projects from *in silico* to *in vivo* models will be described here.

A) AIFIRA facility - microbeam line dedicated for micro-irradiation

A microbeam line has been designed and constructed at the LP2iB (*Laboratoire de Physique des 2 infinis Bordeaux, Gradignan*) onto a particle accelerator from the AIFIRA (*Applications Interdisciplinaire des Faisceaux d'Ions en Région Aquitaine*) facility, starting from the early 2000s under the direction of P. Barberet and P. Moretto[120]. This line can deliver α particles (148 keV/µm) and protons (12 keV/µm) of energy ranging from 1 to 3.5 MeV with a spatial resolution around 0.8 µm. Experiments using solid track detectors in air resulted in around 99.5% of the particles delivered on target for targets under 5 µm from the beam center thus confirming the capacity to irradiate at a cellular and sub-cellular level.

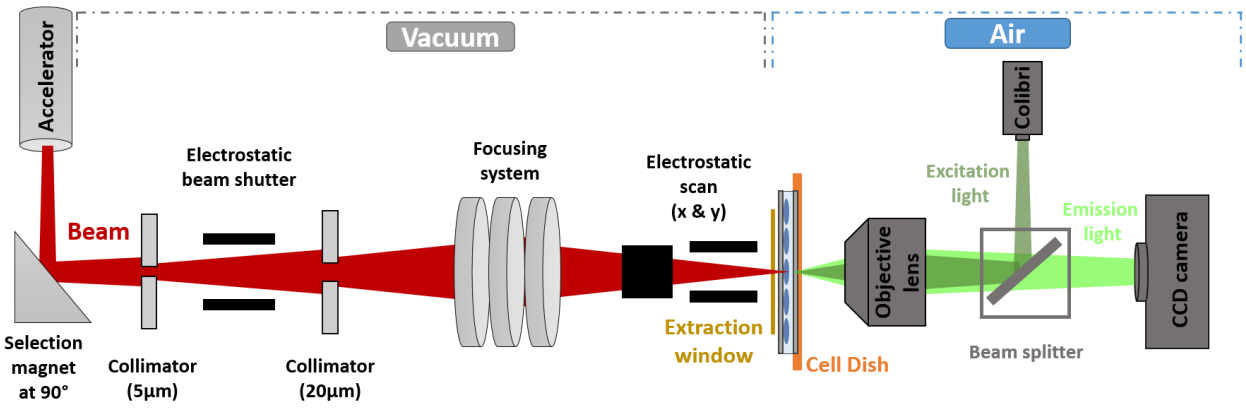


Figure 9. Diagram of the microbeam line @ AIFIRA facility.

This microbeam line has been upgraded with the inclusion of a real-time fluorescence microscopy system, which is also motorized to scan and automate the irradiation of samples, along an ultra-thin diamond proton detector designed for micro-irradiation[121]. This allows for example, the following of early response to radiation-induced DNA damage in *in vitro* cells by monitoring the *in situ* expression of DNA repair proteins tagged with fluorochromes. It can also be used for the targeting of specific regions highlighted by fluorescence in *in vitro* cells or *in vivo* small organisms[122].

Experiments using the AIFIRA microbeam line on cells (*in cellulo*) have made it possible to observe the re-localization of DNA repair proteins such as XRCC1 on the damage sites in irradiated Hela cells[123] and RNF8 that continuously accumulates along the tracks of α particles in the minutes following irradiation[124]. Another experiment at the sub-cellular scale in collaboration with the SNAKE facility (Munich) found membrane potential loss in mitochondria following highly localized targeted mitochondrial irradiation[125]. Irradiations were also carried out *in vivo* on *Caenorhabditis elegans* 2-cell embryos and allowed to follow in real time the radiation-induced damages (appearance of radiation-induced foci)[126]. From these results, the question arises whether other information can be extracted from these original experiments, such as the cellular mechanisms impacted, and correlate them to achieve a better understanding of the nature of radiation-induced damage.

B) AIFIRA facility - microbeam line dedicated to chemical quantification

A second microbeam line has also been developed and was designed to use ion beam analytical techniques (Particle Induced X-Ray Emission, Scanning Transmission In Microscopy, *etc*) by delivering light ion beams in the MeV energy range. These methods allow the quantification of chemical elements at a micrometric precision on biological samples, whether it is the composition of the targeted tissue or the presence of nanoparticles if they are composed of a material detectable by this energy range. This precision allows the study of the distribution of chemical elements at the single cell scale, which was performed on human keratinocytes and *Caenorhabditis elegans* worms with metal oxide nanoparticles[127]. This single-cell quantification of nanoparticles could thus be correlated to a complementary study of the cellular mechanisms involved in the nano-induced cellular response.

C) Geant4/Geant4-DNA

The passage of particles through matter and the subsequent interactions occurs on a scale that is not measurable by current techniques. Yet these interactions and the subsequent energy deposition are the basis of the entire radiation-induced response. In order to overcome this deficiency, softwares simulating the passage of charged particles through matter have been developed using codes based on the Monte Carlo method[128,129]. This technique encompasses multiple computation algorithms relying on random sampling which allows for reproducing of the stochastic nature of IR interactions through random draws.

The Geant4 toolkit, initially released in 1998, is one of the main platforms using these Monte Carlo methods on IR interactions with matter[130]. While it was initially developed solely to model physics experiments for nuclear and particle physics, it quickly evolved to integrate tools for the analysis of interactions in various domains such as space physics, medical applications, microelectronics, etc.[131,132]

One of the extensions to this toolkit is the open-source and publicly available Geant4-DNA project, which focuses on the applications in radiobiology and radiation therapies by developing models for physical interactions at the molecular level with the objective of simulating biological damages induced by IR[133,134,135,136]. The physical interactions of physical particles (electrons, protons, neutrons, etc.) with DNA constituents (backbone and nucleotides) can be simulated step by step using a variety of physics models, while still retaining the micrometer and nanometer precision from the base Geant4 models. This simulation can go up to 1 µs after irradiation and incorporate the physico-chemical and chemical stages of water radiolysis[137,138]. These capacities have made it possible to predict direct and indirect early DNA damage on simple models of bacterial cells and human fibroblasts. Models for small biological organisms, including the L1 larvae *Caenorhabditis elegans* mutant described previously, are currently under development under the direction of Sébastien Incerti and the iRiBio team, with the objective of comparing radio-induced damage observed in experimental data and predicted through Geant4-DNA.

D) Biological models: from the molecule to the organism – multiscale approach

The results obtained in a biological study and their analysis are strongly related to the type of model used and its scale. For example, if *in vitro* models are generally easier to study, the answers they provide can only be partial compared to *in vivo* models. Different study models

have been implemented within the iRiBio team for different study needs and are part of a dynamic of progressing from basic to complex models:

- *in vitro*: DNA molecules in suspension in water allowing to work on the study of DNA fragmentation without risk of activation of repair or protection systems (nucleus, cell membrane, etc.) in order to produce experimental data to be confronted with the Geant4/Geant4-DNA simulation codes
- i*n cellulo*: sarcoma lines to study the radio- and nano-induced cellular response on a type of model whose production and analysis by microbeam are already mastered in the team
- *in vivo*: model *Caenorhabditis elegans* in order to progress towards the study of radio- and nano-induced damage in the more complex context of a multicellular organism.

1/ in vitro approach – from in silico simulation of radiation-induced DNA fragmentation to Long-read sequencing of reference DNA

a) pBR322 plasmid DNA

This plasmid is among the first multipurpose cloning vectors developed to study major genetic mechanisms such as cloning, selection and expression of recombinant molecules, gene expression regulators, etc. It was created in 1977 by Franscisco Bolivar Zapata and Raymond L. Rodriguez (plasmid Bolivar Rodriguez) by deriving it from other known plasmids. Its sequence of 4 361 bp includes an origin of replication (ori) coming from the ColE1-type plasmid pMB1, and two antibiotic resistance genes Amp (ampicillin resistance) and tet (tetracycline resistance). It also contains restriction sites for more than 40 restriction enzymes[139].

In 1979, it became the first fully sequenced plasmid[140] which in addition to its small size and relative simplicity led to its extensive use in numerous studies including the study of DNA fragmentation, as its size allows an easy quantification of the fragmentation by electrophoresis.

For the same reason, we selected this molecule in order to perform our first long-read sequencing on a molecule of accessible size but also to have an easy comparison element via electrophoresis gel migration

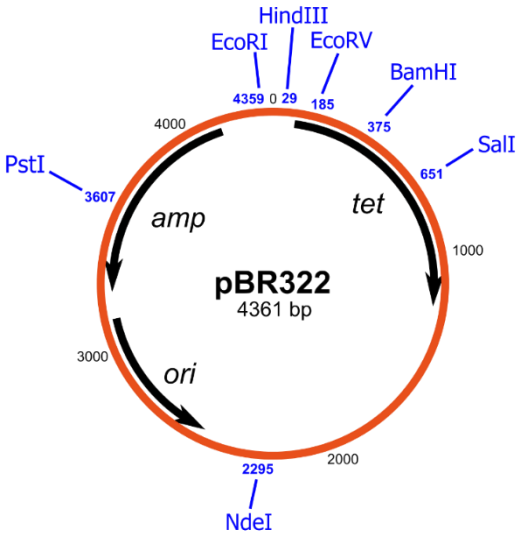


Figure 10. Schematic representation of the pBR322 plasmid genome.

b) Lambda phage DNA

The Enterobacteria phage λ is a double-stranded DNA bacteriophage infecting *Escherichia coli* discovered in 1951 by Esther Lederberg[141]. Its genome can be found in a linear (48 490 bp with 12 bp overhangs at 5' ends) or circular configuration (48 502 bp) and contains numerous restriction sites for restriction enzymes (BamHI, EcoRI, HindIII). Genes coding for 17 proteins have been identified in its genome although some genes have yet to be defined.

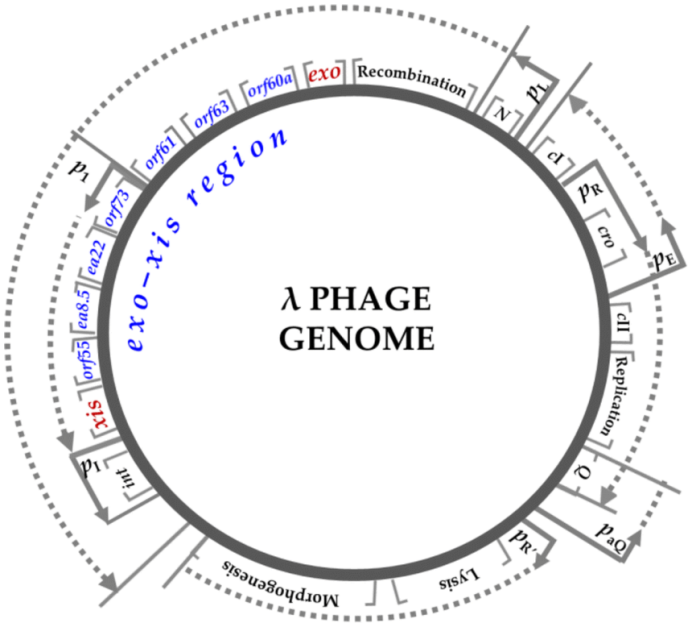


Figure 11. Schematic representation of the Lambda Bacteriophage genome.

From its discovery to the mid-1980s, it became a central model for studies on gene expression mechanisms, viral DNA assembly and lysogeny (one of two cycles of viral reproduction)[142]. While initially used as a model organism in microbial genetics and molecular genetics, it later came to be used as a vector for the cloning of recombinant DNA due to its site-specific recombinase (int) which could be used in the gateway cloning system[143].

We selected this molecule for its interesting size. It is indeed a large molecule (> 10 times the size of pBR322) which is no longer analyzable by classic electrophoresis methods and which length is approx. half the length limit for DNA molecules to be considered as "ultra-long" (>100 kb) for nanopore sequencing[144]. Thus, this molecule gives us a good template to test the feasibility of our experiment on long-read molecules.

c) Phage T4 DNA

The Escherichia virus T4 is a double-stranded DNA bacteriophage infecting *Escherichia coli* discovered around the 1940s. It is a complex virus with 289 open reading frames and coding for more than 40 structural proteins[145].

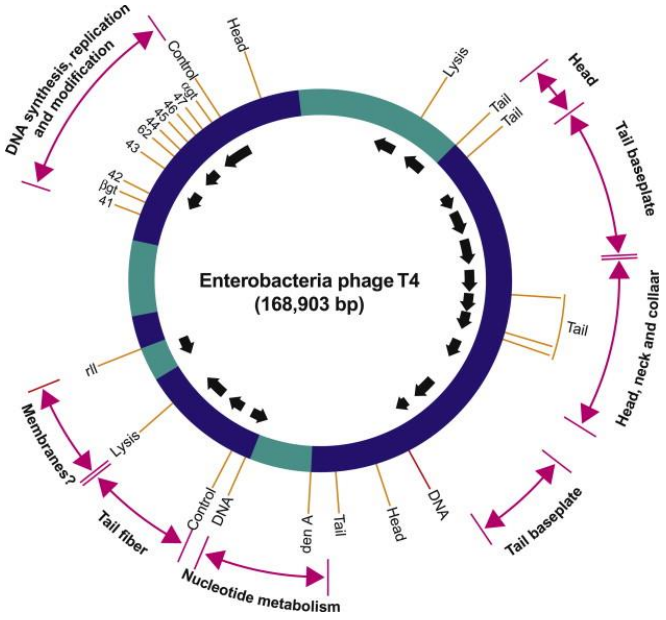


Figure 12. Schematic representation of the Bacteriophage T4 genome

This bacteriophage, along the other T phages, was used as the basis for the studies that determined the role of DNA in the transmission of genes. Numerous studies have also been carried out on the biochemical structure and DNA packaging of the T4 phage which are based on complex protein assemblies[146,147].

We selected this molecule to act as an ultra-long test molecule. Although longer molecules have been sequenced in a single read (up to 8 Mb), these sequencings usually produce only a handful of full reads at best. We therefore wish to evaluate our ability to sequence this ultra-long molecule satisfactorily in order to perform fragmentation studies on it.

2/ Sarcoma – derived patients cell lines

Sarcoma are a heterogenous group of malignant tumors developing in the connective tissue and which can be divided into 2 main categories: bone sarcomas and soft-tissue sarcomas, each with their own staging and treatment approaches. The causes of the development of these cancers remain unclear, although several risk factors have been identified including exposure to ionizing radiation or chemicals. This cell type has been one of the driving forces behind research into cancer mechanisms[148], with the first study dating back to 1909 on hens[149] and remains today a popular model for *in vitro* cancer studies due to the persistent difficulty of treating these types of cancers. Their significant heterogeneity coupled with low prevalence has severely complicated the development of effective treatments, hence the strong interest in studies of these cancers[150]. Although there are many differences between the various types of sarcomas, some molecular factors common to this family of cancers have been identified. These include somatic mutations, intergenic deletions, gene amplifications and reciprocal translations. The majority of high-grade sarcomas with complex karyotypes also possess high frequencies of p53 and pRb (retinoblastoma protein) mutations[151].

Sarcoma cell lines are therefore a biological model of interest in cancer research because of the large number of still unknown mechanisms in their development and the need to develop effective treatments. Two patient-derived lines, IB106 and IB115 are used in the team's reasearch, which despite the innate high genetic variability present in tumor cells have a stable lineage gene expression. IB106 was derived from an undifferentiated sarcoma and presents a highly rearranged genome while IB115 was derived from a dedifferentiated liposarcoma and is characterized by an amplicon profile with a characteristic MDM2 and CDK4 amplification as well as a moderately rearranged genome[152].

3/ Caenorhabditis elegans – a multicellular organism

C.elegans is a small free-living transparent nematode measuring up to 1mm at the adult stage, living in temperate soils and feeding on bacteria. It was first described in 1900 by Émile Maupas[153] who initially named it *Rhabditis elegans* before it was placed in the *Caenorhabditis* genus in 1952 by Günther Osche[154]. It became the first multicellular organism to have its whole genome sequenced in 1998[155] and two Nobel Prizes in Physiology or Medicine were awarded to research on biological mechanisms in *C. elegans*[156,157].

C.elegans embryos are contained in impermeable eggshells after which they hatch and can develop into the four larval stages (L1-L4), each step from one larval stage to another being characterized by a "lethargus" stage during which the worm enters a sleep-like period of inactivity where feeding is stopped, locomotion reduced and a new cuticle is produced after molting the old one[158]. Hermaphrodite individuals (~99.5% of the population vs 0.5% males) have a reproductive cycle of about 2.5-4 days and their lifespan is around 18-20 days when cultured at 20°C [159]. Changes in temperature can have a significant effect on the longevity and development of the worms with a shorter lifespan at high temperatures than at low temperatures[160].

The development cycle of *C. elegans* individuals always follow an identical path and results in individuals with 959 cells for hermaphrodites and 1031 cells for males. This invariability has made it possible to study the fate of somatic cells from the fertilized egg to the adult individual and to catalog them in multiple lineages[161] which in turn made it a choice model for studies on developmental biology, cell cycle, cell death (131 cells are programmed to die during the development of the worm), etc. *C.elegans* has a cylindrical body typical of nematodes with an outer tube (cuticle, hypodermis, neurons, muscles, excretory system) separated from an inner tube (pharynx, intestine, gonads in adults) by the pseudocoelomic space[162].

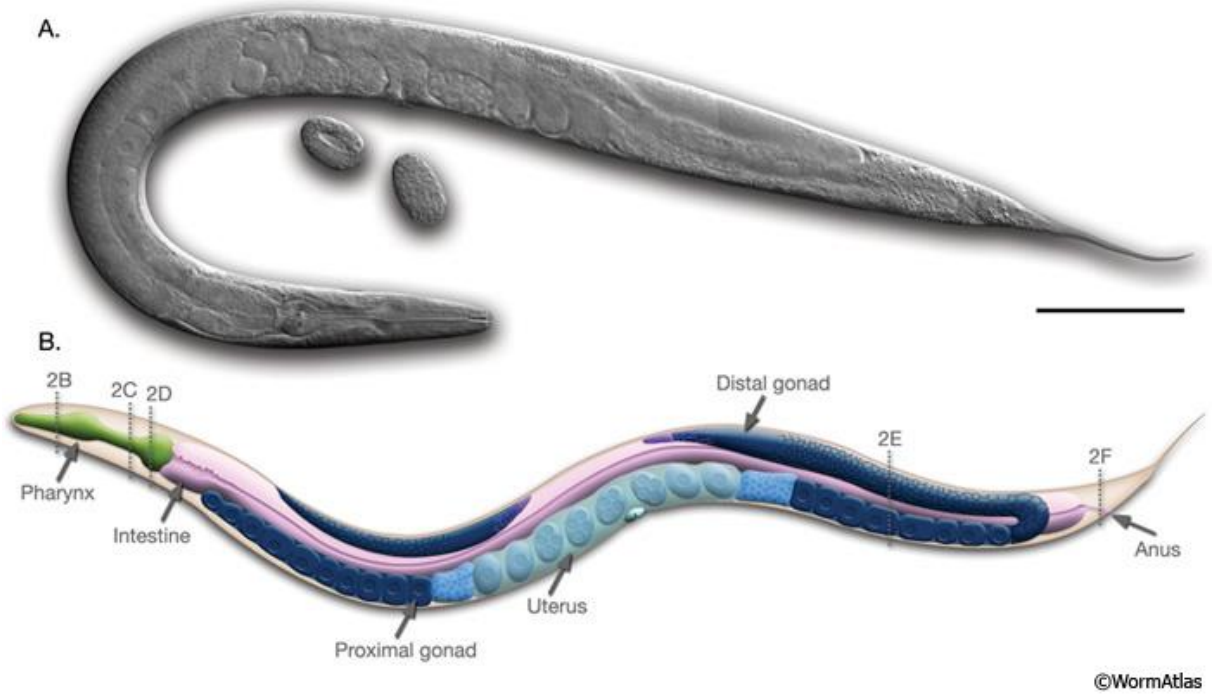


Figure 13. Anatomy of adult hermaphrodite. A. DIC image of an adult hermaphrodite, left lateral side. Scale bar 0.1 mm**. B.** Schematic drawing of anatomical structures, left lateral side.@WormAtlas

The intrinsic biological characteristics of *C. elegans* make it a popular model in the field of life sciences for the following reasons:

- **Genetically homogenous lines**. With over ~99.5% of the natural population being made up of hermaphroditic individuals capable of self-fertilization and producing up to 300 eggs per individual, it is possible from a single individual to create a line whose entire progeny will share an identical genetic heritage because of the absence of genetic mixing at each new generation. This homogeneity added to the invariance of the development of these worms results in adult worms theoretically identical in their development, anatomy and life cycle.

-**Easy and inexpensive to grow**. The rapid developmental cycle of *C. elegans* and the large number of eggs laid per worm means that it is easy to produce large numbers of individuals from a small initial number of worms. The maintenance cost of these worms is relatively low, their main food being OP50 (a strain of *Escherichia coli*) which is also cultivable. Moreover, contrary to animal lines such as the mouse, the individuals can be preserved easily over long periods by freezing them at -80°C. The worms can indeed be stored this way for several months

and resume their normal life cycle once they are brought back to room temperature and placed on a culture medium.

-**Availability of numerous mutant lines**. The previously described ability to create a lineage from a single worm has also made it possible to create numerous mutant lineages via gene knockout or transgenesis. This availability of numerous mutant lines with varied phenotypes makes it easier to study targeted biological mechanisms. These lines are easily accessible thanks to the Caenorhabditis Genetics Center (CGC), which acquires, maintains and distributes stocks of mutant *C. elegans* worms.

-**Synchronization of the development stage**. A natural worm population has individuals at all stages of development. However, the chemical resistance of the eggshell of the laid embryos allows a selection of individuals from this population. Indeed, by introducing bleach in this environment ("bleaching"), all the worms will be destroyed except the embryos, protected by their shell. The result is a synchronized population of individuals all at the same stage of development.

In summary, it is possible to establish genetically identical and synchronized populations of *C. elegans* worms at the same developmental stage, thus minimizing as much as possible the biological variability in a response to a stimulus. These populations can be obtained for a relatively low price, in a short time frame and from a small initial quantity of worms.

E) Thesis objectives: integration of 3rd generation sequencing in research projects

In summary, the iRiBio team has methods to study the biological effects of ionizing radiation and metal oxide nanoparticles at a micrometric precision. Several biological models have been integrated in the research projects with the aim of studying these cellular responses in a complex *in vivo* and multicellular environment at high throughput. However, there is a need for analytical tools of this cellular response that can allow the study of the cellular mechanisms involved with a greater definition in order to complement the existing methods of the team (flow cytometry, qPCR, confocal microscopy, etc.). It is therefore to solve this need that 3rd generation sequencing has been introduced in these research projects.

The integration of the 3rd generation sequencing in those projects has thus been articulated with the aim of being a tool both for the analysis of samples produced on the microbeam line and for the validation of Geant4-DNA simulation codes. Scripts were developed in Python and R to

analyze the sequencing data produced according to the type of molecule sequenced, the model used and the information sought on the dataset. These scripts were also designed to be used routinely in the team by non-bioinformaticians.

This integration was done within several projects already underway from DNA to transcriptome and on several models of studies from DNA in suspension to the *in vivo* model *Caenorhabditis elegans*. An analysis of Single-Cell RNA-Seq data available in the literature was also carried out in order to study whether this technique and the associated analysis methods are compatible with our case study of low quantity samples

1/ Part I: Direct measurements of DNA strand breaks by long-read sequencing

The modeling codes developed under Geant4-DNA allow to simulate DNA molecules in silico, their irradiation by charged particles and the occurrence of radiation-induced damage including direct and indirect damage. These codes contain many parameters governing the interactions between charged particles and DNA molecules, but the values of these parameters are mainly determined from theoretical studies and have not been validated by experimental data. The production of this type of experimental data is, however, complex and has historically been achieved through the use of electrophoresis techniques that offer only limited accuracy.

The long read sequencing technology from Oxford Nanopore Technologies offers the capacity to sequence entire DNA molecules and genomes without the usual prior fragmentation from other sequencing techniques. It would therefore be theoretically possible to carry out sequencing runs in which any observed fragmentation would come from the initial state of the DNA molecule and not from the library preparation protocol. It is with this ability in mind that we used this technology to quantitatively measure the radio-induced DNA fragmentation at different doses of irradiation and compared the results obtained with simulation models The analysis of DNA fragments extremities using this technology also giving us information on the genomic position of the reads end allows for the identification of potential preferred sites of breakage. This study was performed on 4 302 bp (pBR322 plasmid) and 48 502 bp (Lambda phage) DNA molecules to test the feasibility of the study at different sizes. We also sought to introduce a 168 903 bp genome (T4 phage) that falls into the "ultra-long read" category into this analysis by testing different protocols in order to achieve reproducible sequencing.

2/ Part II: Analysis of radio- and nano-induced cellular expression by transcriptomic analysis

Part II.1: Study of radio-induced molecular damage on the RNA metabolism

A microdosimetric study carried out under Geant4-DNA allowed to determine on a 2-cell stage embryo of *Caenorhabditis elegans* modelized numerically in 3D, the deposition of energy occurring during a targeted irradiation on the nucleus. It found that up to 96% of the dose is de posited outside of a DNA molecule depending on the conformation of the chromatin (Torfeh *et al*). Cellular damage caused by IR is thus not limited to DNA damage and can impact the entirety of cellular macromolecules essential for its function within the organism and possibly for its survival even at low doses where no DNA damage is observable. The cellular mechanisms involved in the cellular response to radiation-induced damage to these cellular macromolecules could therefore potentially be characterized by using transcriptome sequencing methods.

A research project was therefore established, aiming at producing micro-irradiated samples on a specific region of an *in vivo* model, precursor cells of the reproductive system in L1-stage *Caenorhabditis elegans* worms, by associating several analysis techniques: confocal microscopy, flow cytometry and long-read sequencing. A major complexity of this project is the low number of worms that can be produced per sample (~200-500), as the targeting by micro-irradiation is done manually. This limitation in the amount of initial material can be problematic especially for transcriptome analysis by sequencing, which requires minimal amounts of RNA to function. The objective was therefore to validate the developed protocol, from sample preparation to the possibility to perform transcriptome analysis, in order to establish a starting point for further study of these micro-irradiated samples. The validation of this protocol would eventually allow to move towards a potential complete characterization of the radiation-induced response on the micro-irradiated regions by analyzing only the transcriptome of the micro-irradiated cells with techniques such a Single-cell RNA-Seq.

Part II.2: Study of the *in vivo* nano-induced cellular response in combination with microscale detection and quantification of nanoparticle exposure

Previous studies conducted with the AIFIRA microbeam line on human cell cultures (keratinocytes, endothelial, cancer, *etc*) *in vitro* combined chemical imaging with biological analyses to detect, follow and determine the effect of titanium dioxide NPs exposure. These studies allowed the identification of the activation of stress pathways by qPCR, in particular of the endoplasmic reticular stress, and a disruption of the calcium homeostasis of the cells while

quantifying the dose of NPs at the scale of the individual cell[163,164]. However, these results have not yet been validated in an *in vivo* model, raising the question of how to study this possible cellular response in a multicellular and more complex model.

In this sense, the objective is to transpose this analysis method to study the effect of exposure to two different shapes of titanium dioxide NPs on the *in vivo* multicellular model *C. elegans*. in combination with chemical imaging, to track the path of the NPs in the organism and their impact on the cellular homeostasis, and by incorporating transcriptome sequencing to have a global view of the impacted cellular pathways. The analyses already carried out on this model made it possible to detect that the calcium homeostasis of the intestinal cells was well impacted and that the worms presented a growth delay but without the NPs being internalized in the cells. The question is therefore whether transcriptome analysis is capable of explaining this phenomenon by detecting the alteration of cellular pathways.

3/ Part III: Evaluation of single cell RNA-Seq in a low yield and high complexity experiment.

The experimental configurations presented previously (micro-irradiation of specific cells, exposure of cells to metal oxide nanoparticles without internalization) are cases in which the cellular response is mainly limited to a few cells in the whole organism. The use of "classical" transcriptome sequencing methods, *i.e.* on the whole organism ("bulk"), can therefore hinder the efficient detection of this specific cellular response.

This is why we need to explore Single-cell sequencing techniques, which offer the possibility to study mRNAs from individual cells and thus potentially to study the transcriptomic response to IR irradiation at the cell level rather than at the organism level with bulk sequencing methods. However, this technology is still recent and so are the bioinformatics analysis methods for this type of data. The interpretation of the results from these analysis methods is sometimes disputed in terms of the algorithms used and the statistical power. Indeed, the current methods offer variable yields per cell which can be problematic when identifying cell types and performing differential expression analysis. Our experimental model currently involves limited quantities of worms for technical reasons which could limit us in the use of this technology. To determine whether this single-cell sequencing technology could still be used in our study, we downloaded public datasets produced on *Caenorhabditis elegans* and replicated the bioinformatics analysis with particular focus on High Content Cells (HCC). These HCC are the largest cells within the downloaded datasets, we studied their expression to determine if these cells best represented cellular expression rather than the multitude of other smaller cells. Through this analysis, we

sought to determine whether single-cell analyses could be performed using smaller quantities of cells but sequenced globally at a greater depth which would allow us to consider single cell sequencing on our micro-irradiated worms to reach a better definition of the radio-induced response to IR irradiation.

4/ Annex: Effect of nano-sensitization on the cellular response of irradiated sarcoma lines

During my thesis, I also had the opportunity to work on a project which, if it does not completely correspond to the central theme of my thesis, also concerns the cellular response to ionizing radiation and metal oxide nanoparticles studied by transcriptomic analysis. This work, which aims to be the subject of an article in the long term, will therefore be placed as an annex.

Radiation therapy is a common method of treating cancers by damaging cancer cells and preventing them from multiplying through high-dose, localized irradiation. However, the effectiveness can vary significantly depending on the type of cancer, with some cancers exhibiting radio-resistance characteristics, which makes them more difficult to treat. The use of nanoparticles has therefore been theorized to use them as catalysts to promote secondary dose deposition in the vicinity of the NPs resulting in a sensitization the cancer cells and thus reducing their radio-resistance. But the team's previous studies showing a disruption of calcium homeostasis led us to investigate the hypothesis that the combined exposure to these two physical agents could promote cell death during irradiation. We therefore sought to study the cellular expression of the IB106 and IB115 sarcoma lines exposed to titanium dioxide nanoparticles, irradiation or both factors at the same time by sequencing their transcriptome. However, this case study raises the question of the possibility of extracting reliable data in such a variable context with the use of cancer lines, which naturally present cellular variability compared to healthy cells, and exposure to physical agents whose dose deposition may vary between samples.

Part I. Direct measurements of DNA strand breaks by long-read sequencing

Introduction

DNA strand breaks are a type of DNA damage that occur naturally and can be caused by endogenous sources (meiosis, free radicals, collapsed replication forks, antigen recombination, etc.)[165,166,167] or exogenous sources (Ionizing radiation, chemicals, UV, etc.)[168,169]. These breaks can involve one strand (SSB, single-strand break) or both (DSB, double-strand break) which can result in subsequent cellular damage. However, these breaks are not usually deleterious to the organism under normal conditions thanks to several cellular DNA repair processes that can be activated depending on the type of damage identified[170,171] and together form the DNA damage response[172].

Figure 14. Simplified representation of Single Strand-Breaks and Double-Strand Breaks.[173]

These cellular mechanisms are essential to the survival of the cell as defective or lack of DNA repair can result in mutations, loss of heterozygosity and chromosome rearrangements potentially leading to cell death[174] or tumor development[175].

The monitoring of sources of DNA damage is therefore the subject of particular attention, especially for ionizing radiation which can be a source of DNA damage through direct (ionization caused by energy deposit on the DNA) or indirect damage (reactive oxygen species from radiolysis of water caused by energy deposition in the water surrounding the DNA)[176]. The reason for this radiation-induced damage is the energy deposited by the charged particles as they pass through the material (Linear Energy Transfer, LET). The probability of DNA damage varies depending on the type of charged particle as well as its energy and the type of tissue irradiated as these factors will impact the energy dose that is deposited. Thus, particles with a high LET will generally be more likely to cause DNA damage because they will transmit more energy per unit distance traveled[177,178], as illustrated by the increased relative biological effectiveness of charged particles when reaching the Bragg peak[179,180].

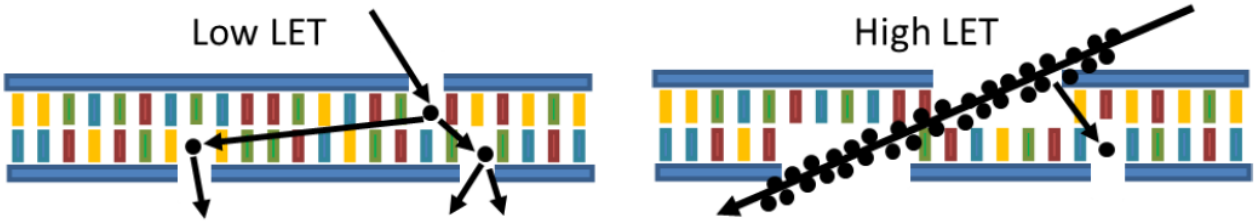


Figure 15. Example of DNA damage for different types of dose deposit.[9]

Public health policies have established exposure limits for the public as well as for the categories of personnel exposed in the course of their work in order to limit as much as possible the probability of suffering unrepaired DNA damage[181]. In order to better measure these risks, models have been developed seeking to accurately estimate the relation between the amount of DNA damage and the dose received. Epidemiology studies have made it possible to calculate the probability of radiation-induced cancer at high radiation doses and have found it to follow a linear relationship which led to the introduction of a so-called linear no-threshold model[182], which remains challenged however because of the lack of concrete evidence of biological effects at low doses[183]. The term low dose refers to radiation levels below those typically associated with biological effects, which include ambient radioactivity to which the general population is exposed. There is no official threshold for energy deposition to classify an irradiation as low dose or not, but the value of 100mGy is frequently used as an upper limit for a human[184,185].

However, this cancer risk approach does not allow a direct estimate of DNA damage for a given dose, as the development of cancer is highly dependent on the damaged chromosomal region and the mutated genes. The precise measurement of DNA fragmentation remains complex today because of the important difficulty to measure fragmentation probabilities. The measurement of DNA fragmentation in living organisms is indeed complex due to several factors: protection provided by a nucleus or a cell membrane, repair systems that can be activated quickly after irradiation, DNA extraction that can damage the molecules, *etc*. Fragmentation measurements tend therefore to be performed on pure DNA samples but difficulties remain due to the small number of measurement methods and their possible bias. In this context, mathematical solutions have started to be developed in the form of modeling this DNA fragmentation. Although theoretical, these mathematical models have the ability to simulate DNA fragmentation without introducing external bias due to an experimental

procedure and can therefore be compared to experimental data and, depending on the degree of certainty in the validity of the experimental data, corrected. Multiple approaches have therefore been developed to assess radiation-induced DNA damage by combining experimental approaches with mathematical models and modelling[186,187,188,189].

One of these projects based on mathematical modelling aiming to achieve a complete characterization of radiation-induced DNA damage by modeling is based on the use of simulation codes for charged particle-DNA interactions. The Geant4-DNA project, built on the basis of Geant4 which simulates the passage of particles through matter by the Monte-Carlo method, seeks to describe the electromagnetic interactions of ionizing particles with biological matter and more particularly DNA. The geometry of DNA molecules in water can thus be modeled and subjected to irradiation simulations in which the trajectory of each particle is followed thanks to the physical modules implemented in Geant4. It is then possible to calculate the energy deposition on the different bases of the studied DNA but also on the surrounding water molecules and thus simulate the direct and indirect ionization phenomena.[190,191]

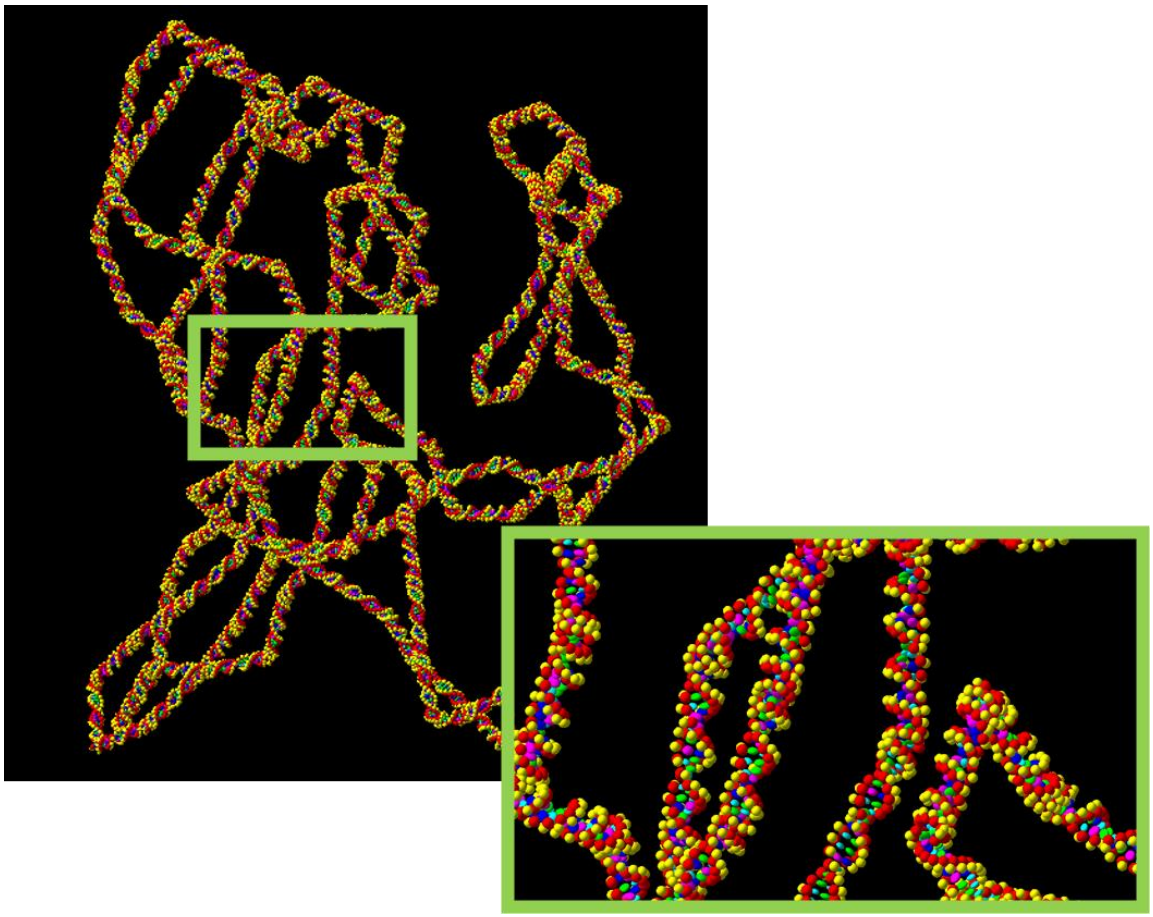


Figure 16. Modeling of a double stranded DNA molecule in double helix conformation under Geant4-DNA. Produced by Konstantinos Chatzipapas, Hoang Tran and Sara Zein of iRiBio team, LP2iB.

In this context, physicists can control all parameters within the model: DNA density, compaction of DNA molecules, type of ionizing particle and its energy, threshold of energy deposition to constitute a strand break, *etc*. [192]

However, these codes are currently based on theoretical models due to the difficulty of obtaining precise experimental results of DNA damage measurement in conditions similar to those used in Geant4-DNA and to the limited number of measurement methods currently available. There is therefore a need for experimental data in order to validate the developed simulation codes but also a difficulty to produce this type of data which we aim to solve by making use of recent technological developments.

Questions addressed in this work

The measurement of DNA fragmentation has historically been performed by gel electrophoresis[193,194] migration due to the high reliability of the method, its low price and the possibility to study several samples on the same gel and thus make a direct comparison. However, this technique suffers from some limitations, notably on the precision of the size of the fragments obtained and on the maximum size supported of the DNA molecules to be migrated. Indeed, beyond about 20 kb, the migration distance of the molecules becomes too small to distinguish the migrated molecules[195]. Variations of the basic method such as pulsed field electrophoresis have been developed to overcome this size limit and allow the study of DNA molecules up to 10 Mb but retain a limited accuracy of fragments size[196]. A few other methods can be used to roughly quantify the amount of DNA damage such as FACS, qPCR or electron microscope imagery but their low accuracy prevents their use as a reliable study method.

In this context and with the recent development of long-read sequencing methods, we wanted to study the relevance of the Oxford Nanopore Technologies sequencing method as a way to measure precisely the fragmentation probabilities lyophilized DNA samples after 3 MeV protons irradiation. The basis of our study comes from the theoretical capacity of this sequencer to fully sequence DNA molecules of any length and thus potentially identify every fragment and obtain their precise length thus allowing for an accurate quantification. In addition, obtaining the sequence of DNA fragments from whole molecules can help identify on which nucleotides the fragmentation has occurred and discover whether preferential fragmentation sites emerge from the analysis of all fragments.

In order to evaluate the suitability of this technology as a method for studying DNA fragmentation, three molecules of different sizes were used: **pBR322 plasmid** (4 361 bp, abbreviated as pBR322), **Escherichia virus λ** genome (48 490 bp, abbreviated as Lambda phage) and **Escherichia virus T4** genome (168 903 bp, abbreviated as T4 phage). Thus, we were able to study the capacities of the sequencer to sequence increasingly long molecules and if the DNA fragmentation remained distinguishable on DNA lengths considered as ultra-long (> 50 kb). The results obtained with this sequencing method were compared with gel migration methods to estimate the validity of the results obtained and subsequently included in simulation methods (Geant4-DNA and percolation models) to adjust codes based on physical models.

In this context of this study, my work has been focused on:

- Implementation of a data recovery and processing pipeline for non-bioinformatician team members: data retrieval (via USB or SSH), unzipping, concatenation, alignment on the reference genome, alignment file handling for visualization.

- Analysis of the sequencing runs produced with the different DNA molecules used. Definition of means to measure and compare fragmentation between irradiation conditions. Visualization of the different results by graphs: reads quantity, size distribution, percentages of fragmented molecules.

- Analysis of the alignment files to extract the start and end genomic position of the reads. Identification of potential preferential fragmentation sites.

Materials and Methods

1. DNA molecules

The pBR322 plasmid used is provided by the Takara company (Catalog #: 3050, Entry Name: SYNPBR322, GenBank accession N°J01749). The stock concentration of the plasmid DNA is 0.5 µg/µl in 10 mM Tris-HCl, 1mM EDTA, pH 8.0. According to the manufacturer, the plasmid contained over 70% double-stranded covalently closed circular. The plasmid contained more than 90% supercoiled DNA as examined by gel electrophoresis.

The phage Lambda genome is provided by New England Biolabs (Catalog #: N3011, GenBank accession N° J02459.1). The phage is isolated from the heat-inducible lysogen *E.coli* l *c*I857 S7. The DNA is isolated from the purified phage by phenol extraction and dialyzed against 10 mM Tris-HCl (pH 8.0), 1 mM EDTA at a 0.5 µg/µl concentration.

The phage T4 genome is provided by Nippon Gene (Code N° 318-03971, GenBank accession N°NC_000866). *E.coli* MC1061 was infected with bacteriophage T4 GT7, the phage was separated by cesium chloride density gradient centrifugation, and DNA was extracted from the phage into 10 mmol/l Tris-HCl (pH 8.0) , 1 mM EDTA at 0.2-0.5 µg/µl.

2. Sample preparation

Thin layer of dried DNA was prepared on polypropylene foil (4-µm thick) as follows: polypropylene foil was cleaned by incubation for 30 min in an ethanol solution (70%, v/v), three times at room temperature, then dried and treated with UV-C (germicidal lamp) 30 min at room temperature. Aliquots of 2 µl DNA solution were deposited in the center of the polypropylene foil. Lyophilized samples were then freeze-dried under vacuum (2. 10-5 Bar, -85°C) for 4h before irradiation. Air-dried samples were kept 3 h at room temperature under the laminar flow of a microbiology hood in sterile conditions.

3. Irradiation

Protons irradiation was performed at the AIFIRA facility (External beamline). The proton beam was adjusted to 3 MeV under vacuum and extracted in air through a 200 nm thick Si3N4 window. The beam section is ≈ 1x1 mm² at the position of the window. In order to spread the beam size, the samples were positioned 10 cm away from the window. This leads to an

incoming energy at the position of the DNA molecules of 1.43 ± 0.06 MeV (LET = 20.7 keV.µm-1 in water).

DNA samples were placed on a dedicated sample holder and were maintained at room temperature. DNA samples were also directly irradiated at atmospheric conditions in air. The proton beam was directly targeted on the DNA deposits.

After irradiation, the samples were kept at atmospheric condition, each DNA sample was recovered and resolved in 8 µl TBE buffer (10 mM Tris-Borate, 1 mM EDTA, pH 8.0) at room temperature. Samples were kept for long-term conservation at -20°C in sterile and sealed petri dish before recovering in Tris buffer and until electrophoresis. Control samples were processed in the same way except without being irradiated.

4. Sequencing

The pBR322 plasmid samples were digested beforehand with the BamHI enzyme for 18 hours at 37°C in the dedicated buffer to linearize them.

The libraries were prepared following the standard protocol for the SQK-LSK109 ligation sequencing kit from Oxford Nanopore Technologies. One library was prepared for each sample and they were sequenced on a Mk1C MinION using individual Flongle cells with a min_qscore of 7 and live basecalling for > 21 hours per sequencing run.

5. Bioinformatic analysis

The reads were aligned on reference genome obtained from GenBank using minimap2 2.24[197] with the "--ax map-ont" option. The resulting alignment files were processed using samtools 0.1[198]. The analysis was carried out using Python v3.10.7, the plots were produced using matplotlib v3.6.1 and seaborn v0.12.1. The code is available at:

https://github.com/pelotbdr/iribio_scripts/tree/main/longread_DNA

6. Geant4-DNA simulation

The geant4-DNA Monte Carlo track structure toolkit is used to describe the spatial distribution of energy depositions. The physics interactions with water and inside the biological geometries

were simulated using G4EmDNAPhysics_option2 and the interactions with the DNA molecules are simulated using the water cross-sections. A supercoiled-like molecule of length equal to that of the pBR322 plasmid is built based on the double helix structure and measurements of the molecular position of the B-DNA made by Arnott & Hukins. The nucleobases and the backbone of the DNA were considered as a collection of molecules, rather than discrete atoms. Nucleobases were represented by ellipsoids and sugar and phosphate groups were represented by spheres filled with water. Up to 10 142 plasmids can be modelized in a single simulation, evenly spaced inside a box of 4.42µm x 4.42µm x 4.48µm to reproduced the experimental density (about 0.5µg/µl). Each identical molecule is placed inside a spherical voxel of 200nm in diameter and no overlaps are allowed between voxels. One randomly chosen face of the box is irradiated with a monoenergetic parallel beam of 3 MeV protons. The incident particles were shot until the sum of all energy deposits in the volume led to an absorbed dose of 500, 1000, 2000 and 5000 Gy.

Direct damage occurs when energy from physical processes is deposited on or near a DNA molecule. In this model, we associate damage with a DNA molecule based on a single distance value. The maximum distance (r) from the center of a molecule that can result in any energy deposition tied to that model is called the direct interaction range

Experimental results

1. Long-read sequencing of reference DNA genome: plasmid pBR322 and Lambda phage

The first steps of this study were to investigate our ability to sequence the different DNA molecules selected. In order to establish fragmentation levels according to the deposited dose, it is necessary to establish a reference on intact molecules (control, non-irradiated). Our sequencing runs were thus carried out on DNA molecules without irradiation.

For the achievement of these sequencing runs, we decided to use low-cost Flongle ONT chips. Our needs in terms of quantity of reads are indeed quite low, as we only need to sequence a sufficient quantity of reads (tens of thousands) in order to study the size distribution of the fragments. The capabilities of "classical" arrays in terms of throughput (millions of reads) for the purpose of performing deep genome assemblies or discovering structural variants are therefore not useful in our case.

1/ pBR322 plasmid reference

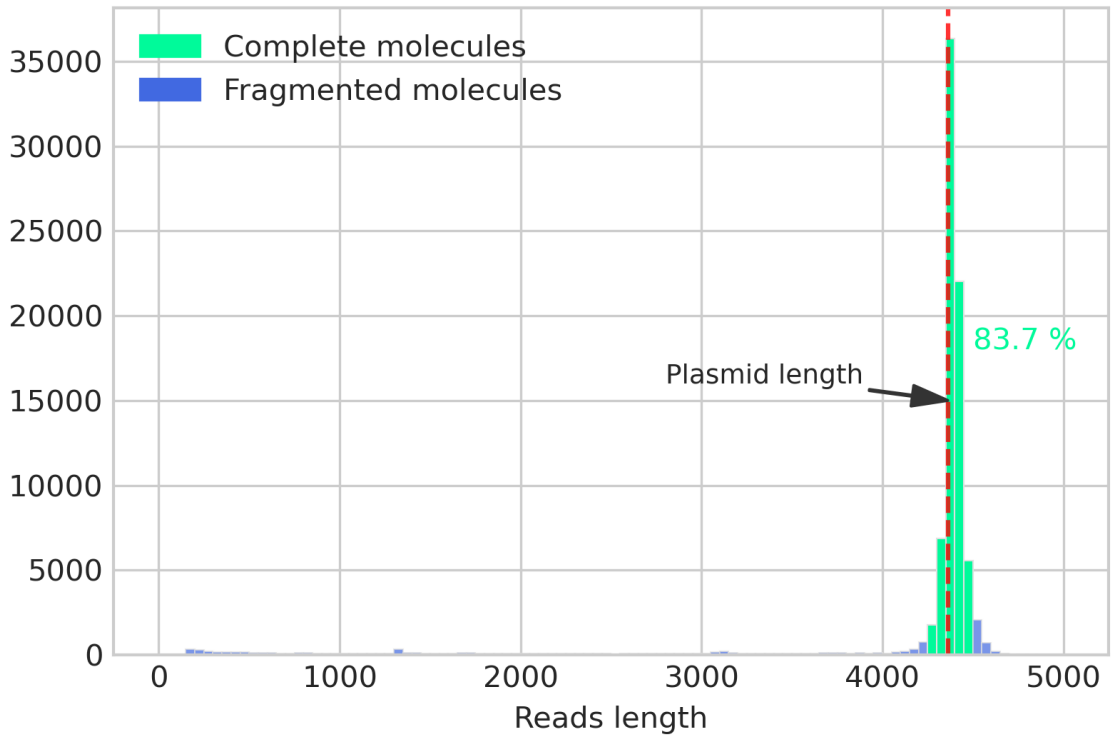


Figure 17. Size distribution of reads sequenced from pBR322 plasmids.

On our sequencing of native pBR322, we obtain a distribution with a clear peak towards the expected size of 4361 bp and containing >80% of reads sequenced. A very low number of fragmented reads can be identified as well at various sizes. However, we observe a variation on this peak with a non-negligible number of reads exceeding the expected size and reaching sizes of 4500-4600 bp (Figure 17). Nanopore technology being still one of the most error-prone sequencing techniques, the CIGAR (Concise Idiosyncratic Gapped Alignment Report) sequence which describes the alignment base by base for each reach was extracted. We found that the majority of the reads present in this peak have an offset (calculated by subtracting the number of deletions from the number of insertions) which can account for most of the peak distribution (Figure 18).

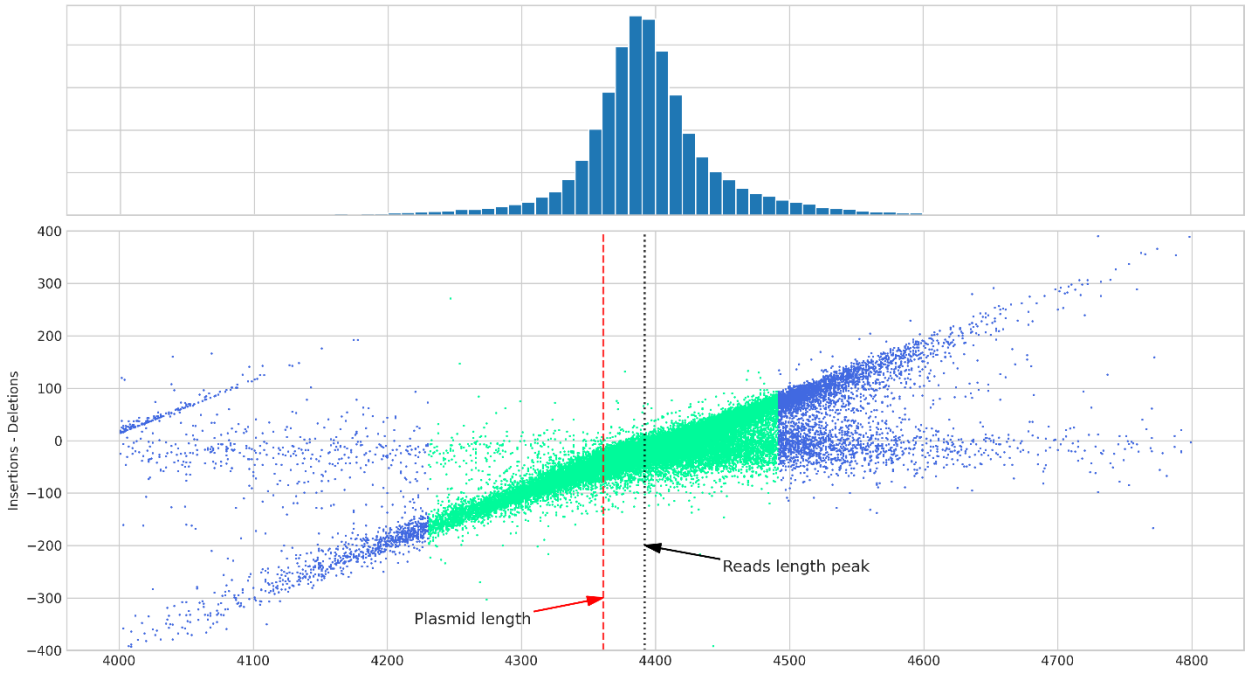


Figure 18. Distribution of sequence length offset (number of Insertions – number of Deletions) by read length.

However, there is still an anomaly to explain which sees the peak at its maximum for a read length of 4392 bp and this despite an offset remaining slightly negative. Our hypothesis is that during digestion by BamHI followed by the adapter ligation, some free-floating nucleotides resulting of the production of overhangs by the enzyme digestion are ligated into the sequence before the adapters thus artificially lengthening the read sequence. The origin of this hypothesis comes from the similar length of the soft clips found in the CIGAR alignment at the start and end of the reads that have an average respective length of 28 and 30 bp (Figures 19-20). This

average addition of 60 bp to the read sequence, in addition to the slightly higher number of deletions compared to the insertions, could offer a satisfactory explanation for the observed shift in the expected peak read size at 4392 bp instead of 4361.

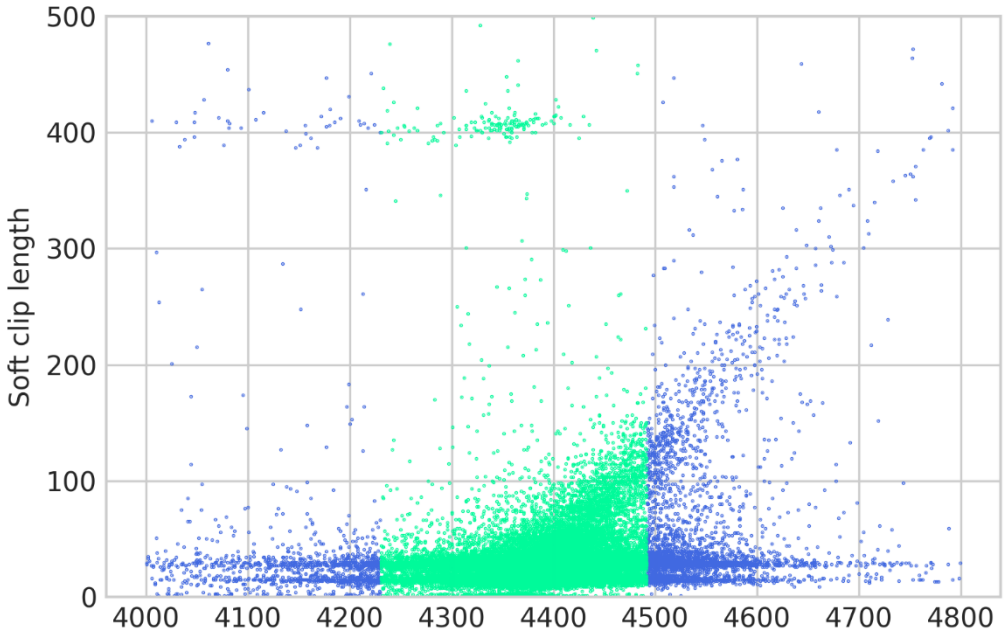


Figure 19. Length of soft clips at the start of the read alignment by read length.

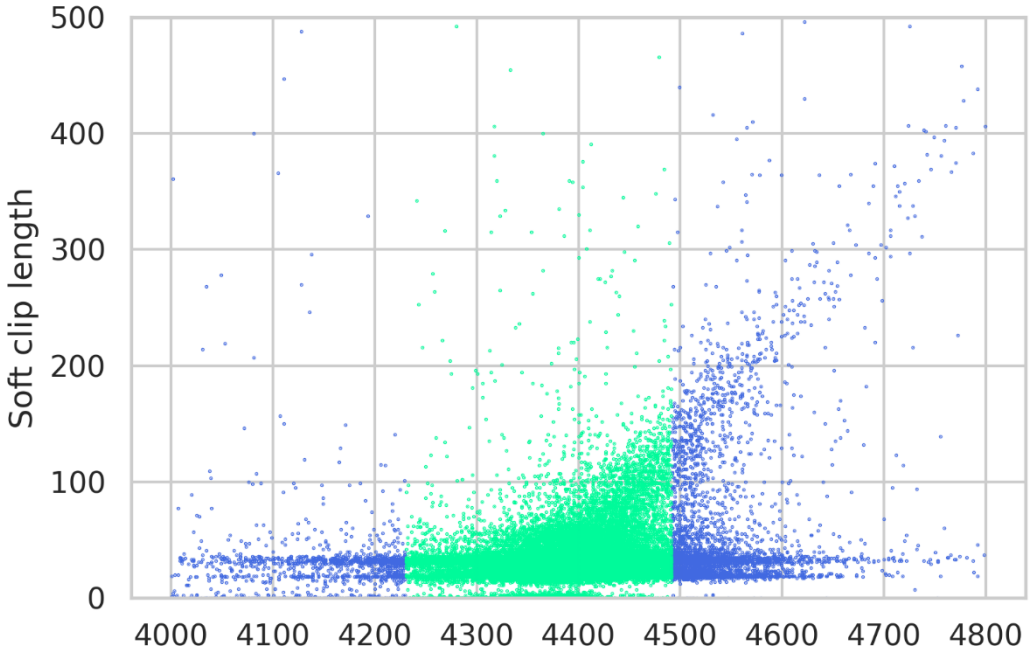


Figure 20. Length of soft clips at the end of the read alignment by read length.

Nonetheless, because of these variations in read length, we had to determine how to categorize which reads are considered as "complete" reads. We found that including reads of reference length ± 3% allowed for the inclusion of 83.7% of all the reads sequenced and included the majority of reads contained in the peak. The choice of using the reference length ± 3% rather than the peak maximum ± 3% which would have included evenly all the reads from the peak comes from the fact that this peak maximum differs between sequencing runs and thus cannot be used a reliable basis to establish the percentage of completed molecules. As for the arbitrary choice of 3% rather than another percentage, we found that a higher percentage, while including the reads of 4500-4600bp would also include reads outside of the peak which could thus be the product of fragmentation, a lower percentage on the other hand would have unjustifiably excluded too much reads from the peak.

Overall, we consider that as long as this condition is similarly applied to irradiated molecules sequencing runs, it offers a satisfactory means of measurement of the percentage of complete molecules and thus of fragmentation quantification.

2/ Lambda phage reference

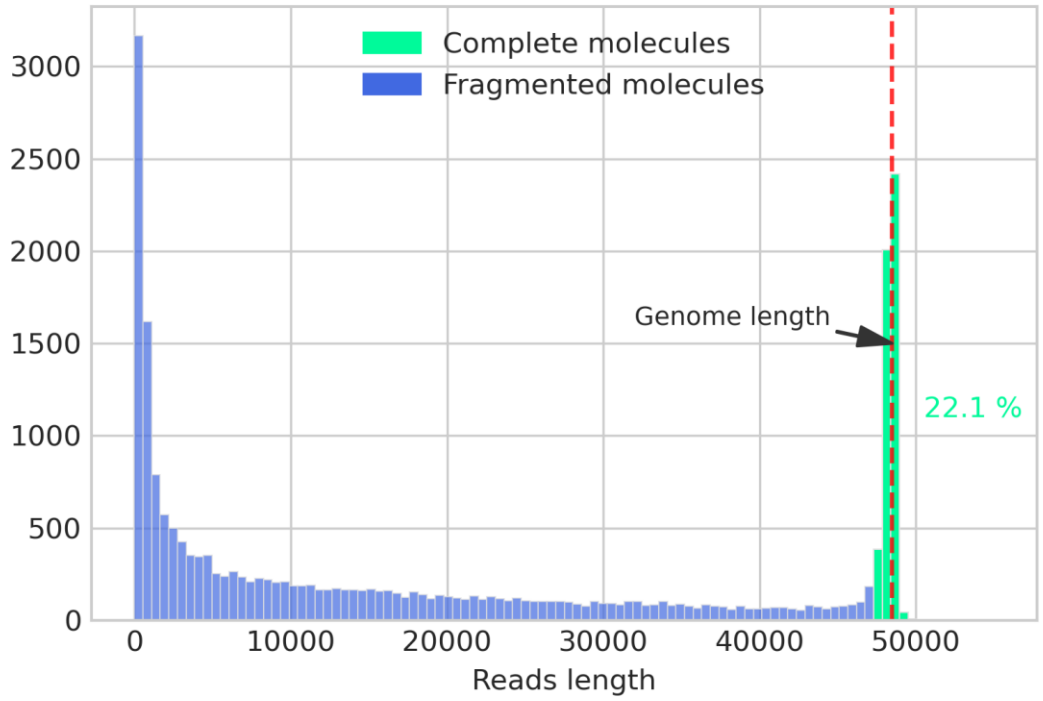


Figure 21. Size distribution of reads sequenced from native Lambda phages.

Contrary to pBR322, the sequencing of native Lambda phage DNA displays an important amount of unexpected reads of lower size than expected with only 22.1% of the reads being categorizable as complete reads despite the use of the same protocol. These unexpected reads result in a distribution were small fragments (~0-2000kb) are in majority and reads of all sizes up to that of the genome length can be found (Figure 21). Although this base can be used to study DNA fragmentation, it is a clear representation of the technical limits of the sequencer, which offers a decrease in efficiency for the sequencing of complete molecules when this size increases with a consequent decrease for molecules approaching the ultra-long size category.

However, there is a bias due to the preparation of the sequencing library in this model. Indeed, the standard protocol includes the addition of a quality control in the form of a DNA molecule of about 4kb (*DNA CS*, ONT SQK-LSK109 kit), which was done on these sequencing runs. However, this quality control corresponds to a modified extract of the Lambda phage genome. If the difference with the original reference allows to distinguish the majority of these "*DNA CS*" (ONT reference) from reads coming from the studied Lambda phage, the distinction becomes more complex as soon as the control is not sequenced in its entirety or with a too high error rate on the modified parts of its sequence. As a result, some reads may be erroneously assigned as coming from the studied Lambda phage when they actually come from the DNA CS, creating an artificial fragmentation due to their size. It should also be noted that sequencing was also performed on non-lyophilized native samples to determine if this process was responsible for the low number of complete reads. These sequencings yielded similar results with between 20 and 25% complete molecules sequenced, indicating that freeze-drying does not significantly damage the DNA.

2. Measurement of radio-induced fragmentation on pBR322 plasmid and Lambda phage DNA

From the references obtained by sequencing of the genomic DNA of pBR322 and Lambda phage, we proceeded to irradiations using 3 MeV protons at the AIFIRA facility and sequenced the samples under identical conditions between the two models used. Each sample was sequenced on a single Flongle chip, the quality and number of pores varied from chip to chip, the number of sequenced reads is not a good indicator of the fragmentation rate and so we kept the metric of number of complete molecules. This number was determined by keeping the ±3%

interval rule having been determined on the reference sequencing runs. 2 replicate batches were performed for the pBR322 plasmid and 3 replicate batches for the Lambda phage.

1. Product of sequencing runs on irradiated DNA molecules

(A)

pBR322	Complete molecules (%)	Mean read length (bp)	Median read length (bp)	Number of reads
0 Gy	82.35	4 115.45	4 384	85 490
500 Gy	78.87	3 983.26	4 378	51 464
1000 Gy	78.05	3 927.92	4 372	31 493
2000 Gy	63.43	3 649.08	4 364	69 494
5000 Gy	55.14	3 406.35	4 341	98 005

(B)

pBR322	Complete molecules (%)	Mean read length (bp)	Median read length (bp)	Number of reads
0 Gy	81.43	3 999.62	4342	104 748
500 Gy	71.77	3 752.13	4 349	283 293
1000 Gy	67.46	3 666.20	4 328	185 975
2000 Gy	62.61	3 511.60	4 336	119 536
5000 Gy	47.17	3 134.65	4 132	104 953

Table 1. Read size statistics for sequencing runs of plasmid pBR322 DNA from (A) Batch 1 and (B) Batch 2.

(A)

Lambda	Complete molecules (%)	Mean read length (bp)	Median read length (bp)	Number of reads
0 Gy	20.10	20 074.20	13 501	13 151
500 Gy	7.80	12 407.59	5 696	24 754
1000 Gy	1.77	6 507.08	2 694	1 697
2000 Gy	0.36	3 502.14	1 815	1 937
5000 Gy	0	4 002.17	1 769	691

(B)

Lambda	Complete molecules (%)	Mean read length (bp)	Median read length (bp)	Number of reads
0 Gy	22.09	20 073.24	13 168	22 468
500 Gy	11.58	14 800.59	7 334	27 369
1000 Gy	8.52	14 358.07	8 300	23 320
2000 Gy	0.41	11 883.46	7 234	37 960
5000 Gy	0.30	7 433.31	4 994	43 851

(C)

Lambda	Complete molecules (%)	Mean read length (bp)	Median read length (bp)	Number of reads
0 Gy	9.67	16 417.23	11 328	27 565
500 Gy	8.57	15 974.36	10 962	28 083
1000 Gy	4.42	13 043.08	8 638	32 141
2000 Gy	1.35	9 457.63	6331	54 528
5000 Gy	0.15	6 554.54	4 549	81 935

Table 2. Read size statistics for sequencing runs of Lambda phage DNA from (A) Batch 1, (B) Batch 2 and (C) Batch3

The results of these sequencing runs allow us to observe that the orders of magnitude are globally similar between the batches of the same molecule but that there remains a significant variability. In the replicates of plasmid pBR322, it can be noted that if the 0 Gy and 2000 Gy conditions are at almost identical percentages of complete molecules, the difference between the two batches is quite important for the other conditions, even if still in a similar order of magnitude. For the Lambda phage runs, a very low number of reads can be noted in batch 1, particularly for the 1000, 2000 and 5000 Gy conditions despite a similar sample preparation between all conditions. The number of reads obtained in the other batches shows that this is not a problem due to irradiation and therefore indicates an issue likely caused by a human error during the sample preparation for irradiation or during the library preparation. In batch 3, the 0 Gy condition also appears abnormally fragmented compared to the other batches.

If part of this variability could be explained by a difference of dose deposit during the irradiation which resulted in higher or lower level of DNA damage, this phenomenon does not explain the

degree of variation observed. The rest of the variability can be explained by the sequencing method in which several factors can play a role: (i) the quality of the chip and number of available nanopores, (ii) the stability of the electric current, (iii) the unpredictability of molecules captured or not by the nanopores, (iv) the quality of the library preparation. A solution to erase this variability would be to perform a larger number of sequencing runs in order to "smooth" the results.

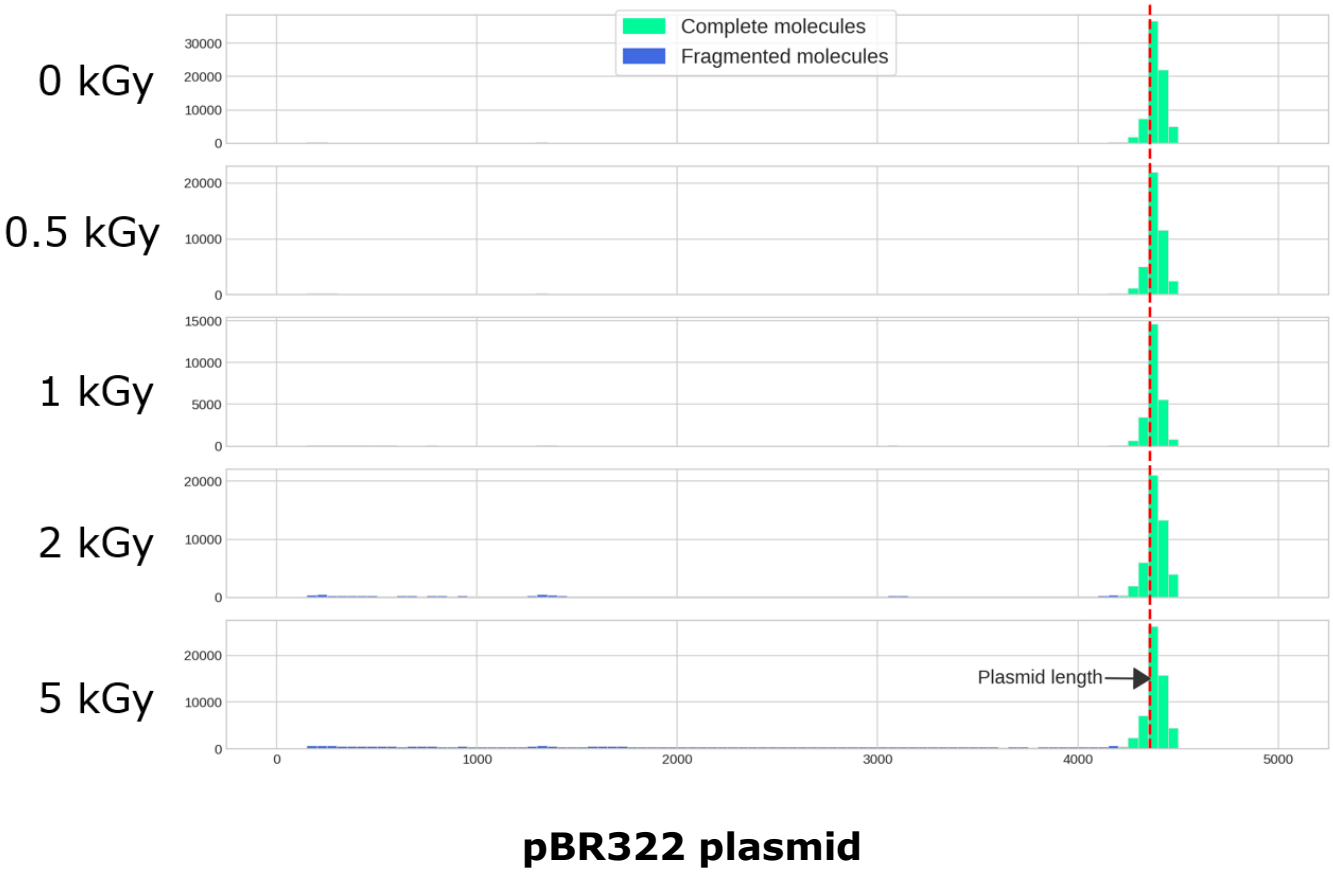


Figure 22. Size distribution of reads sequenced from sequencing batch n°1 of pBR322 plasmids irradiated at different doses.

For the plasmid pBR322 DNA, we can observe that the full molecule peak remains observable and even remains constitutive of a majority of reads even for the highest energy deposition condition at 5kGy. Fragmentation results in reads of all sizes being sequenced which suggests that fragmentation can occur at any point in the plasmid sequence (Figure 22).

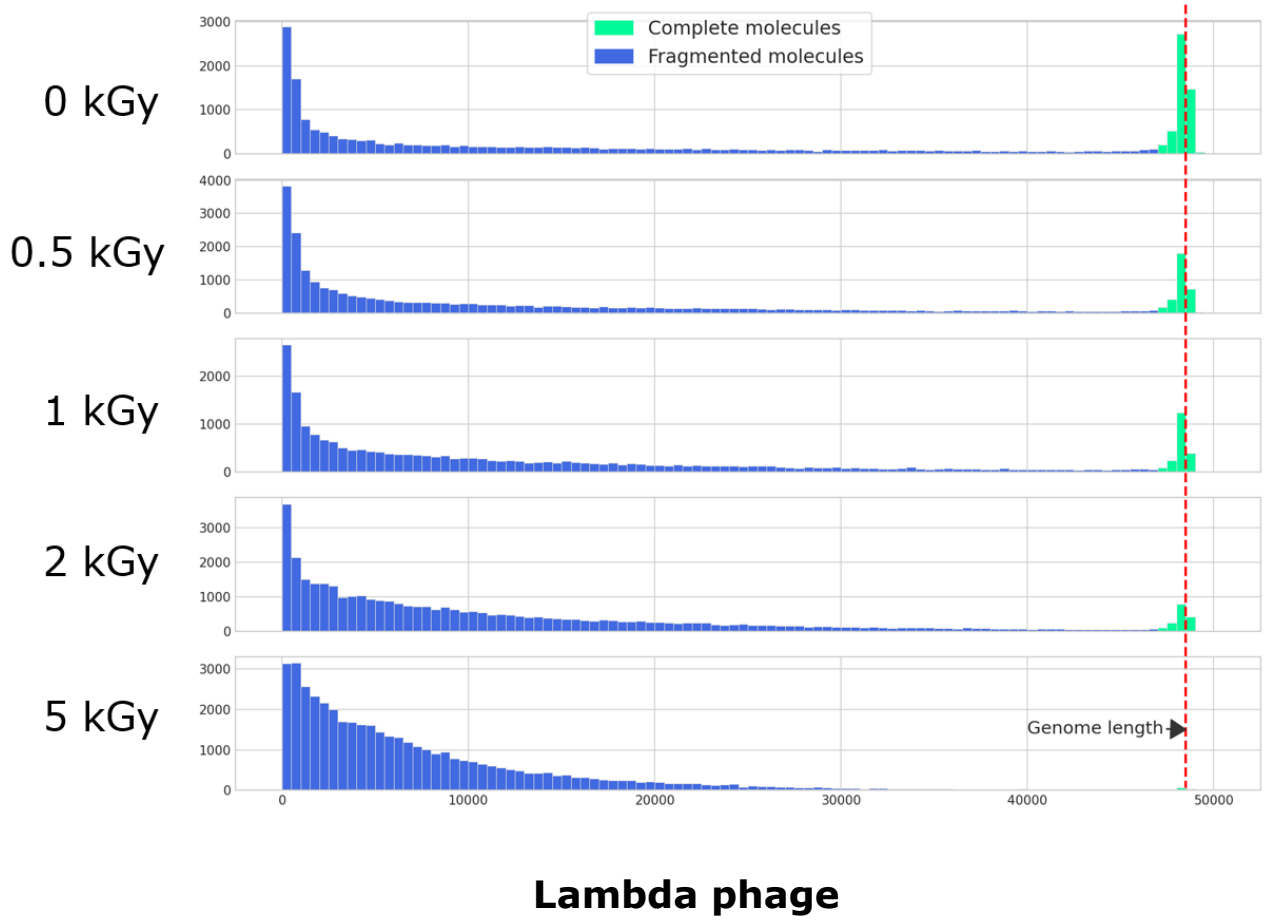


Figure 23. Size distribution of reads sequenced from sequencing batch n°2 of Lambda phages irradiated at different doses.

The Lambda phage DNA sequencings show that fragmentation is more obvious than for pBR322, notably by the progressive disappearance of the peak of complete molecules along the progressive increase of the energy deposit. The decrease of the peak is accompanied by a progressive increase in the number of small reads rather than reads of all sizes as observed previously with the plasmid pBR322 (Figure 23).

2. Calculation of fragmentation probabilities per molecule in an exponential model

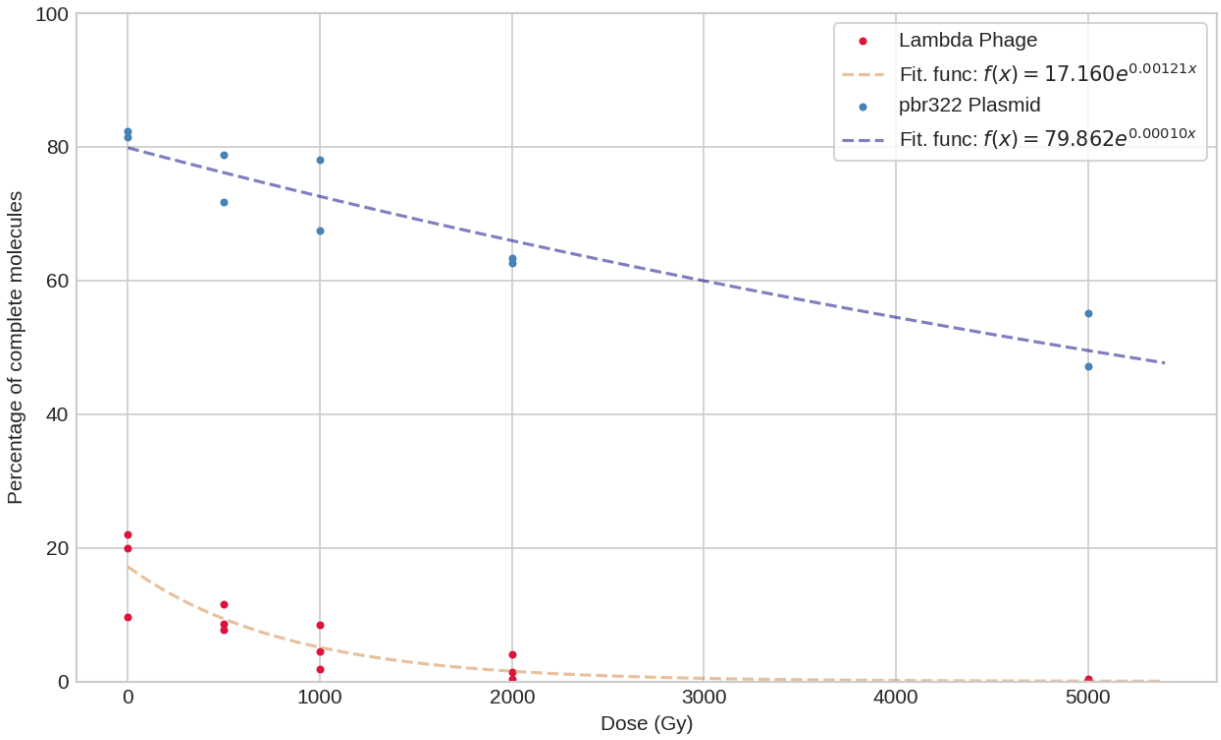


Figure 24. Percentage of complete molecules per dose for pBR322 plasmid (blue) and Lambda phage (red) and exponential function fit for each model.

We plotted the percentage of complete molecules for all replicates of each condition and plotted an exponential curve on the mean of the values for the two models studied (Figure 24). For the definition of this exponential, we consider the following points:

- (i) **Bases are independent**: our system is composed of N links between the bases of the DNA and these links are independent. Each of these links has a probability p of being broken and (1-p) of not being broken, thus creating a 2-state system which allows us to consider that it follows a binomial distribution.

- (ii) **Poisson distribution**: taking into account that N (4 360 or 48 501) >> 1 and that p << 1, we can consider the application of a Poisson distribution with the following formula:

$$P(x) = \frac{(pn)^x * e^{-pn}}{x!}$$

- (iii) **As many reads as molecules**: by considering the probability of DSBs is far smaller to the probability of SSB and that the reading of an SSB by the sequencer interrupts the sequencing of the molecule, we make the approximation that one irradiated molecule corresponds to one read. Given then that we measure the probability that x=0, we can simplify the previous formula as follows:

$$P(x) = \frac{(pn)^x * e^{-pn}}{x!} \qquad \overline{x} = pn + A \text{ (method bias)}$$
$$= \frac{(\overline{x})^x * e^{-\overline{x}}}{x!}$$

$$P(0) = e^{-\overline{x}}$$

The value A, indicated as the method bias, corresponds to the bias induced by the sequencer which fails to sequence the complete molecules under control conditions, resulting in an intercept (value for x=0) different from 1. From this point, np being dependent on the number of links N, the dose D and the fragmentation probability α, the function can be expanded to the following formula:

$$\begin{aligned} \overline{x} &= pn + A \\ &= D * N * \alpha + A \qquad b = N * \alpha \\ &= b * D + A \end{aligned}$$

$$\begin{aligned} P(0) &= e^{-b*D+A} \\ &= e^{-A} * e^{-b*D} \\ &= a * e^{-b*D} \end{aligned}$$

The fit of this exponential function on our data results in the following parameters:

$$y_{pbr322} = (79.9 \pm 1.6) * e^{-(9.6 \pm 1.1)*10^{-5}*D}$$
$$y_{Lambda} = (17.2 \pm 0.3) * e^{-(121.4 \pm 4.2)*10^{-5}*D}$$

We then compared the *b* parameters, which represent the slope of the radiation-induced fragmentation in these equations, between the two models studied and compared it to the ratio between the two models' length.

$$b_{Lambda} = (121.4 \pm 4.2) * 10^{-5}$$
$$b_{pbr322} = (9.6 \pm 1.1) * 10^{-5}$$
$$r = \frac{b_{Lambda}}{b_{pbr322}} \pm \Delta(\frac{b_{Lambda}}{b_{pbr322}})$$
$$= \underline{12.6 \pm 0.2}$$

$$N_{Lambda} = 48501$$
$$N_{pbr322} = 4360$$
$$\frac{N_{Lambda}}{N_{pbr322}} = \underline{11.1}$$

From the parameter *b*, since it can be expressed as *N*alpha*, by dividing it by the size of the molecules we can obtain for each model the fragmentation probability per base per kiloGray.

$$p_{pbr322} = (2.2 \pm 0.3) * 10^{-5}/bp/kGy$$
$$p_{Lambda} = (2.5 \pm 0.2) * 10^{-5}/bp/kGy$$

We thus reach relatively similar ratios between fragmentation rates and molecule sizes and therefore fragmentation probabilities in order of magnitude as well as in value. Although this model is based on several approximations, this result tends to confirm the initial hypothesis postulating that all links were independent with the same probability of being fragmented.

3. Analysis of read ends positions to detect preferred site of fragmentation

We were also interested in the detection of possible preferential fragmentation sites based on the sequence of the reads which is a unique advantage of the sequencing method in the study of DNA fragmentation.

To determine the presence or absence of preferential fragmentation sites, the alignment files were used to determine on which bases the 5' and 3' ends of all reads in a sample are located. The position of the start of the read on the reference file is indicated by default in the alignment details of the SAM files and the end position of the read was determined using the CIGAR sequence. From this sequence, we calculated an offset value based on the number of bases contained in soft-clips, insertions or deletions. For example, one deletion in the read compared to the reference increases the offset value by 1 in order to make up for the "delay" between the two sequences, one insertion will have the opposite effect and will decrease the offset by 1. We can then add the starting position of the read and its size corrected with the offset in order to determine on which base of the reference it ends up. Reads ending with values outside the size of the molecule within a reasonable range (100 nucleotides for the pBR322 plasmid and 1000 for the Lambda phage) are considered to end on the nearest endpoint.

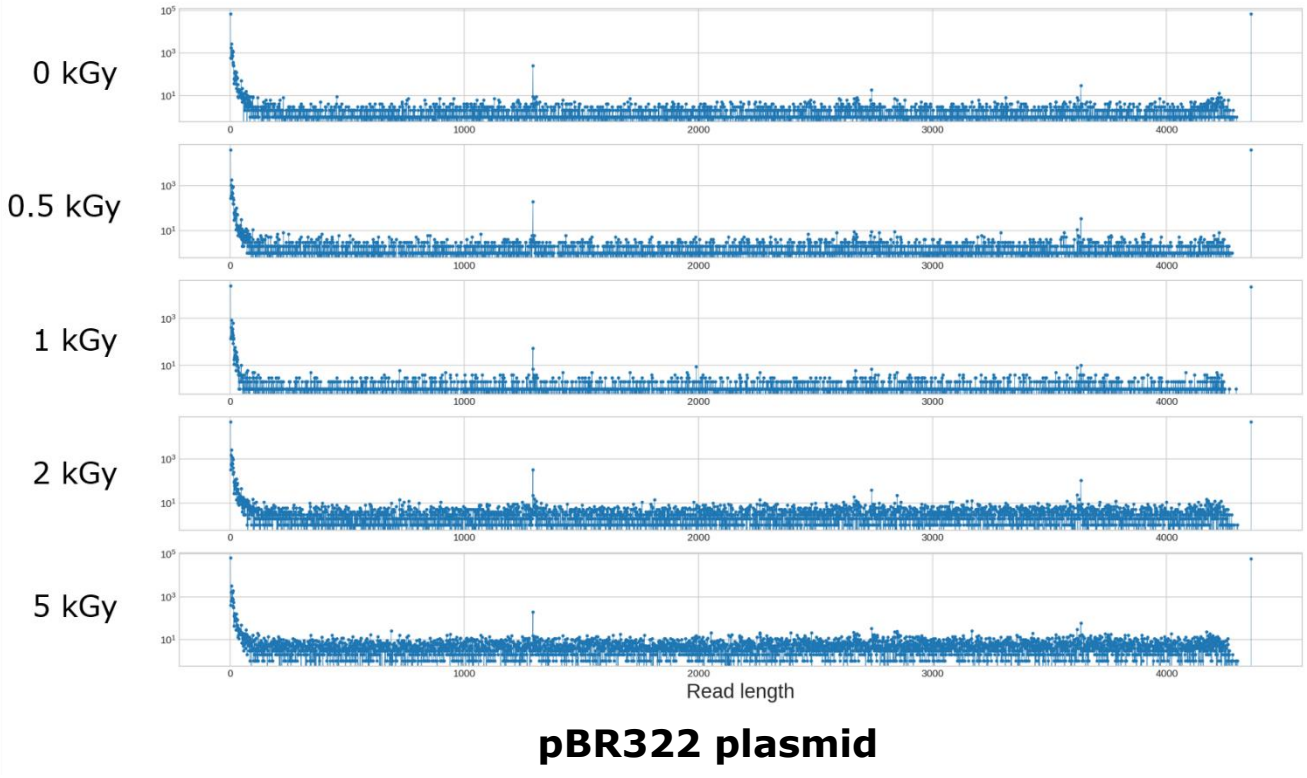


Figure 25. Number per base of reads ending or starting on this base from one sequencing batch of pBR322 plasmid. Y-axis in log10 scale.

We first focus on the pBR322 DNA (Figure 25). pBR322 is a circular DNA sequence and the sequencing by ONT needs its linearization by BamHI, thus the BamHI site was used as site 0 of the reference. It can be observed that the majority of reads have as their ends the BamHI restriction site. Very few reads can be found to have their ends on rest of the sequence although this quantity increases slightly with the dose. 3 phenomena of interest are also observable on this distribution:

- At the 5' end (site 0), we observe that some of the reads above the mean have an extremity on the bases adjacent to the BamHI restriction site. We believe that this is due to potential degradation of the ends produced by the restriction enzyme prior to the grafting of the sequencing adapters, the ends produced containing overhangs which are more sensitive to potential breakage.

- The 3' end does not show the same result from grafted overhangs as the 5' end where one can even distinguish a white zone in the nucleotides preceding the end. The most likely hypothesis explaining this result is a bias due to the method. The calculation method is not perfect because of the overall higher length than expected of the reads, causing the 3' end to be globally "behind" the end of the sequence.

- 2 bases appear above the rest (positions ~1250 and ~3600) and this in all conditions including the control which indicates that it is not linked to the irradiation and therefore preferential sites of fragmentations. The study of these regions reveals that they are BamHI "pseudo-sites" of restriction, i.e., nucleotide sequences almost identical to the BamHI restriction site except for one base. We believe that these pseudo-sites are digested by error by the restriction enzyme, perhaps due to an error in the sequence of the plasmid introduced during its production which transforms this "pseudo-site" into a real digestion site.

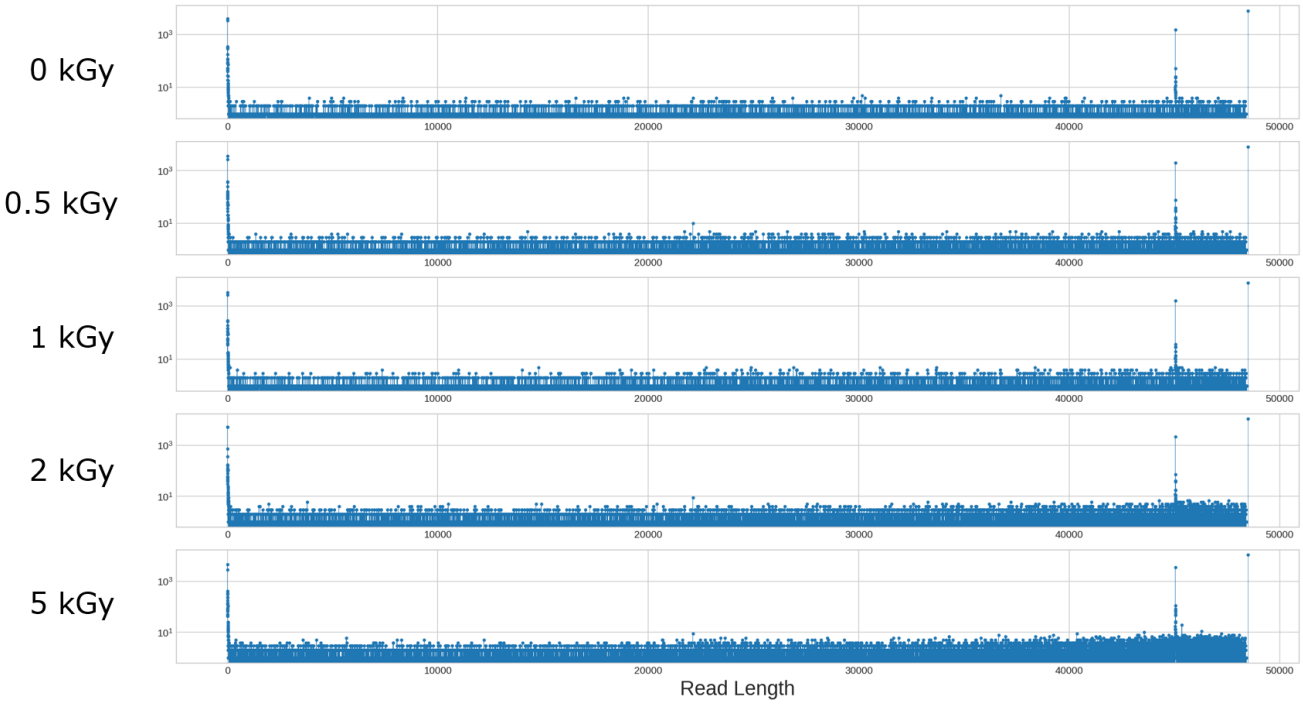


Figure 26. Number per base of reads ending or starting on this base from one sequencing batch of Lambda phage. Y-axis in log10 scale.

Secondly, regarding the analysis of the Lambda phage DNA (Figure 26), we can see that the majority of reads have the beginning and the end of the reference which corresponds well to the expected result, the native molecule being in linear form. However, we can see a similar discrepancy as in plasmid pBR322 but of greater magnitude with some bases largely above the rest around position 46 000. This is actually the starting point of the *DNA CS* sequence and therefore this peak is due to sequences that could not be removed from the Lambda phage read set due to too much sequence similarity. The presence of these *DNA CS* also explains the fact that bases located in the region adjacent to their starting point are slightly higher than other regions of the genome, due to the fact that these molecules terminate on all adjacent bases depending on their fragmentation state.

In summary, whether on the plasmid pBR322 DNA or on the Lambda phage DNA, we do not observe any particular emerging trend that would suggest the presence of preferential fragmentation sites on these molecules.

3. Comparison with a percolation model

A 1D percolation model[199] has been designed by F. Gobet (iRiBio, LP2iB) to simulate the resulting size distribution of reads based only on the previously obtained fragmentation probabilities. In this model, 1000 DNA molecules of a given length are subjected to potential fragmentation, each base is considered as independent and is subjected to the same fragmentation probability. The read sequence is traversed from base to base with a probability test performed on each base to determine if fragmentation occurs. If fragmentation occurs, the size of the traversed read is kept and the rest of the read is ignored, without taking into account whether the other strand is fragmented. The algorithm follows a process similar to the one which takes place during sequencing with the read translocation through the nanopore being stoppable similarly by single strand (SSBs) or double strand (DSBs) breaks. This model was applied on DNA molecules of plasmid pbr322 and Lambda phage length with the associated probabilities ($2.2*10^{-5}$/bp/kGy for pBR322 plasmid DNA and $2.5*10^{-5}$/bp/kGy for Lambda phage DNA) at doses of 1 kGy and 5 kGy. From the distributions obtained with the percolation model, we then compared with those obtained by the sequencing method.

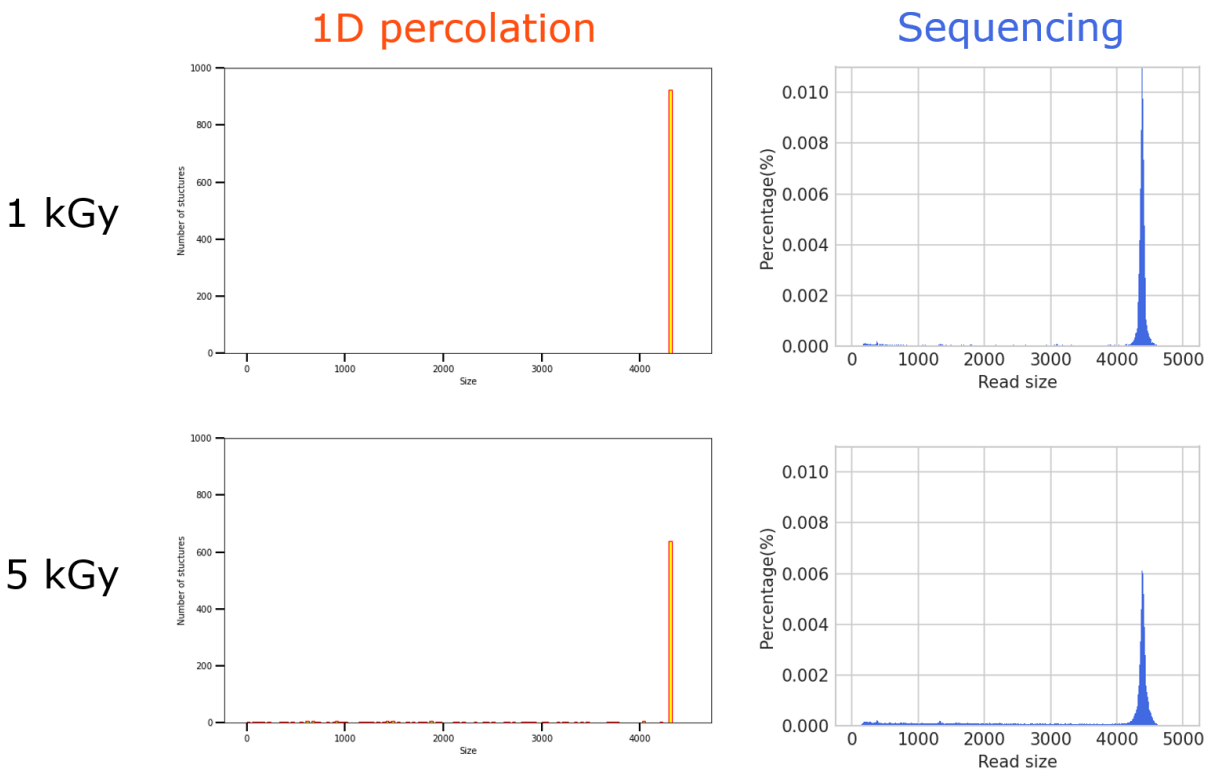


Figure 27. Comparison of read size distribution for sequencing and 1D percolation model at 1 kGy and 5 kGy for the pBR322 plasmid.

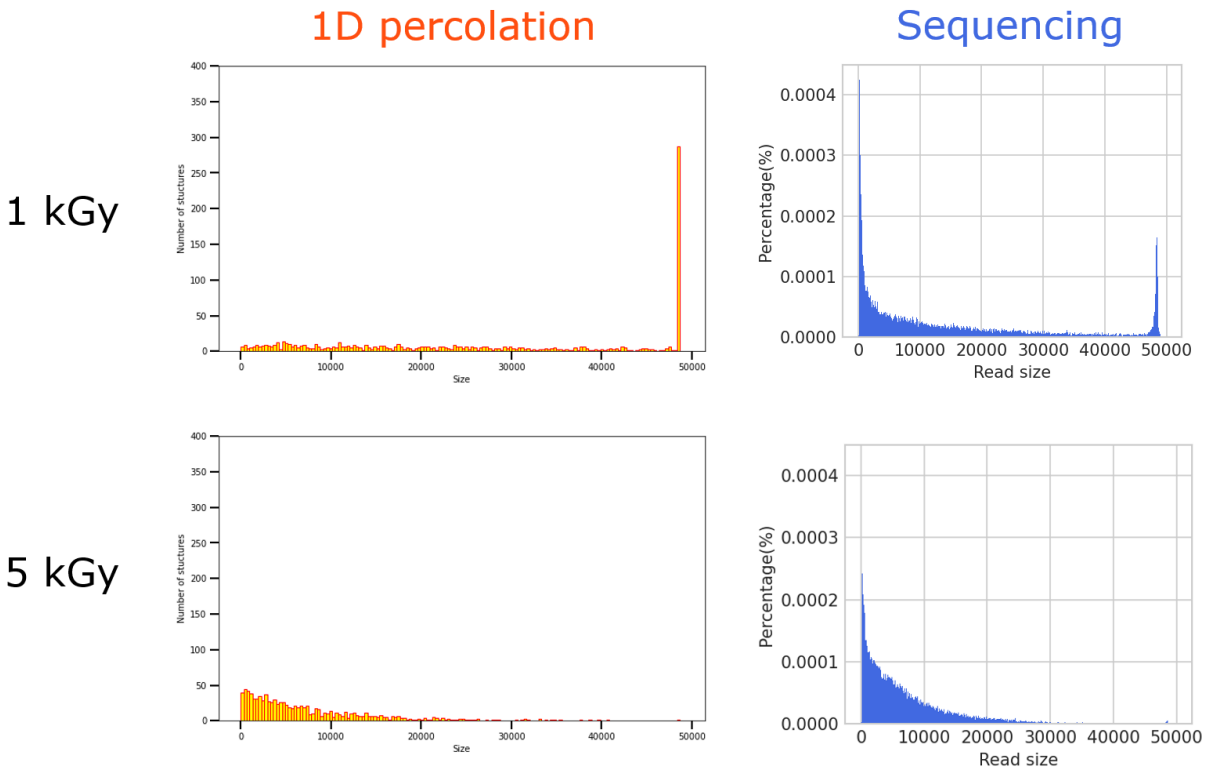


Figure 28. Comparison of read size distribution for sequencing and 1D percolation model at 1 kGy and 5 kGy for the Lambda phage.

For the pBR322 DNA (Figure 27), very little fragmentation is observable at 1 kGy for both methods, and fragmentation at 5 kGy results in fragmented reads of almost any size between 0 and the plasmid size while keeping a consistent peak of complete molecules for both methods as well.

For the Lambda phage DNA (Figure 28), fragmentation at 1 kGy produces reads of all sizes from 0 to the size of the phage with slightly smaller reads than other fragmented sizes while at 5 kGy only small reads remain with the vast majority being smaller than 20 kb.

Overall, this percolation model results in distributions relatively similar to those obtained via sequencing. This is despite the fact that the model is based only on the fragmentation probability, which is itself calculated only from the percentage of complete molecules in sequencing runs and not from the size distribution of the fragmented reads. Since the unwanted fragmentation due to sequencing cannot be replicated by this model, the potential for comparison remains limited, but the relative similarity between the distributions obtained by these two methods is an encouraging sign about the validity of the results obtained by sequencing but also on the ability to simulate DNA fragmentation.

4. Comparison with Geant4-DNA simulations

The geometry of the pBR322 plasmid having been previously developed under Geant4-DNA, the following problem concerned the adjustment of the various parameters related to the irradiation by 3 MeV protons. The two main parameters evaluated here are the threshold of energy deposition causing a strand break and the maximum distance accepted between a base of the plasmid and the energy deposition for it to be considered that the energy has been deposited there (called "hydration shell"). Several irradiation simulations of the pBR322 plasmid were therefore performed testing different parameters from the Geant4/Geant4-DNA literature on DNA fragmentation[200], by K. Chatzipapas (iRiBio, LP2iB), H. Tran (iRiBio, LP2iB) and S. Zein (iRiBio, LP2iB). They also developed codes to measure the size of DNA molecules and fragments produced after irradiation, in order to produce a size distribution of the totality of the irradiated molecules. From the results of these simulations, they selected the parameters that would give DNA fragment size distributions closest to that obtained by the long-read sequencing method.

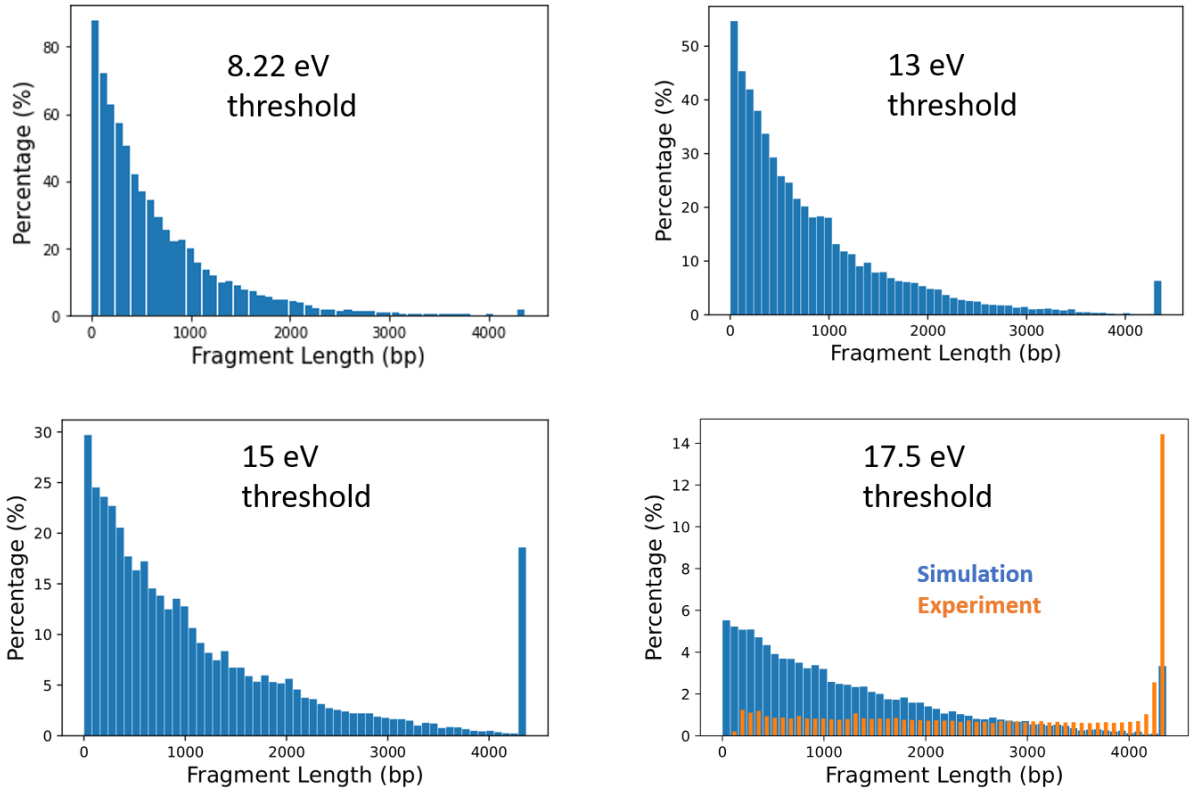


Figure 29. Comparison of fragments size distribution after 3 MeV proton irradiation simulated by Geant4-DNA on pBR322 plasmids for different energy thresholds. Only the first fragments in DNA strands are considered.

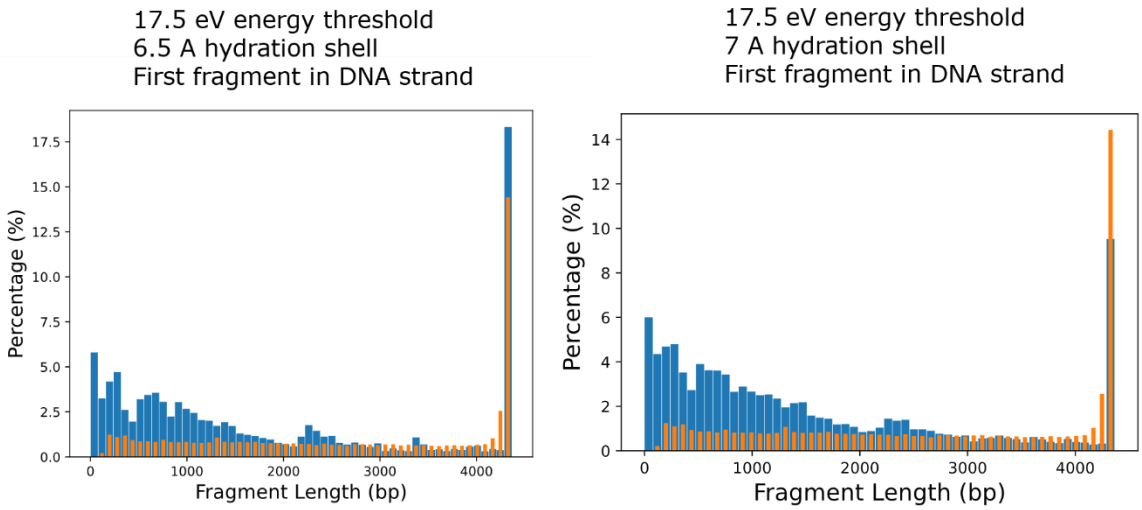


Figure 30. Comparison of fragments size distribution after 3 MeV proton irradiation simulated by Geant4-DNA on pBR322 plasmids for different sizes of hydration shells. Only the first fragments in DNA strands are considered.

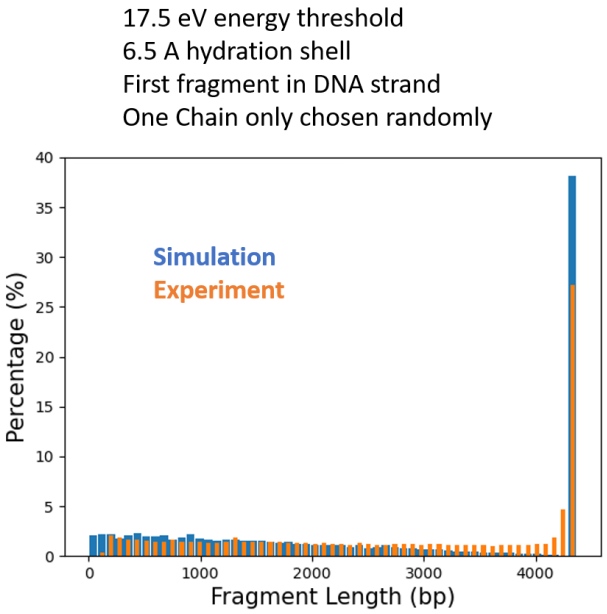


Figure 31. Fragment size distribution after 3 MeV proton irradiation simulated by Geant4-DNA on pBR322 plasmids using selected parameters and a fragment counting method similar to sequencing.

By comparison, the multiple simulations led to values of 17.5 eV for the energy deposition threshold (Figure 29) and a 6.5 Å hydration shell (Figure 30) to best approximate the simulated distribution and the experimental distribution obtained by sequencing. The measurement of the fragment size was also adapted to follow an operation similar to that of sequencing: 1 strand randomly sequenced and stop of the reading at the first SSB or DSB. This configuration results in a highly similar size distribution between simulation and sequencing (Figure 31).

The use of long-read sequencing has therefore provided experimental data that allow the calibration of simulation parameters for radiation-induced DNA fragmentation in Geant4-DNA. Although these results are preliminary, requiring more experimental data and on other DNA molecules to be more trustworthy, they confirm the interest of integrating long-read sequencing in this context of conjugating the development of simulation codes and the production of experimental data.

For the Lambda phage, although the geometry of the molecule is ready in Geant4-DNA, the much larger size of the molecule leads to significant technical constraints that mean that these simulations could not yet be performed at the time of writing.

5. Sequencing of T4 phage DNA perspective

Following these analyses on the plasmid pBR322 and the Lambda phage, we sought to apply this method on a DNA molecule falling into the category of "ultra-long read" (> 100 kb): the T4 phage genome. In the context of the simulation of radiation-induced damage in Geant4-DNA, the objective is to progress towards increasingly long DNA molecules in order to reach sizes close to the genomes of living organisms. It would thus be possible to simulate real cells with precisely defined DNA to obtain a complete characterization of radiation-induced DNA damage in a cellular model. The smallest genome of a living organism that can be studied is 580 kb[201], ~10x longer than the genome of the Lambda phage, but the majority of genomes have sizes in Mb or Gb, which is still a considerable way to reach these scales. The sequencing of the T4 phage genome is therefore part of this progression towards molecule sizes close to those observed in living organisms.

Previous sequencing of the Lambda phage genome had shown that whole molecule sequencing becomes more complex as the size of the sequenced molecule increases. This trend was confirmed by our first sequencing of T4 phage DNA on Flongle, which did not allow us to sequence complete molecules, as the entirety of the reads was smaller than expected. We checked the integrity of our native DNA solution by electrophoresis gel migration to ensure that it was not damaged but we did not detect anything abnormal.

As Flongle chips offer a lower yield than conventional sequencing chips, we performed sequencing on standard R10.4 chips in order to determine if a larger number of pores and therefore a higher yield would allow us to obtain complete molecules.

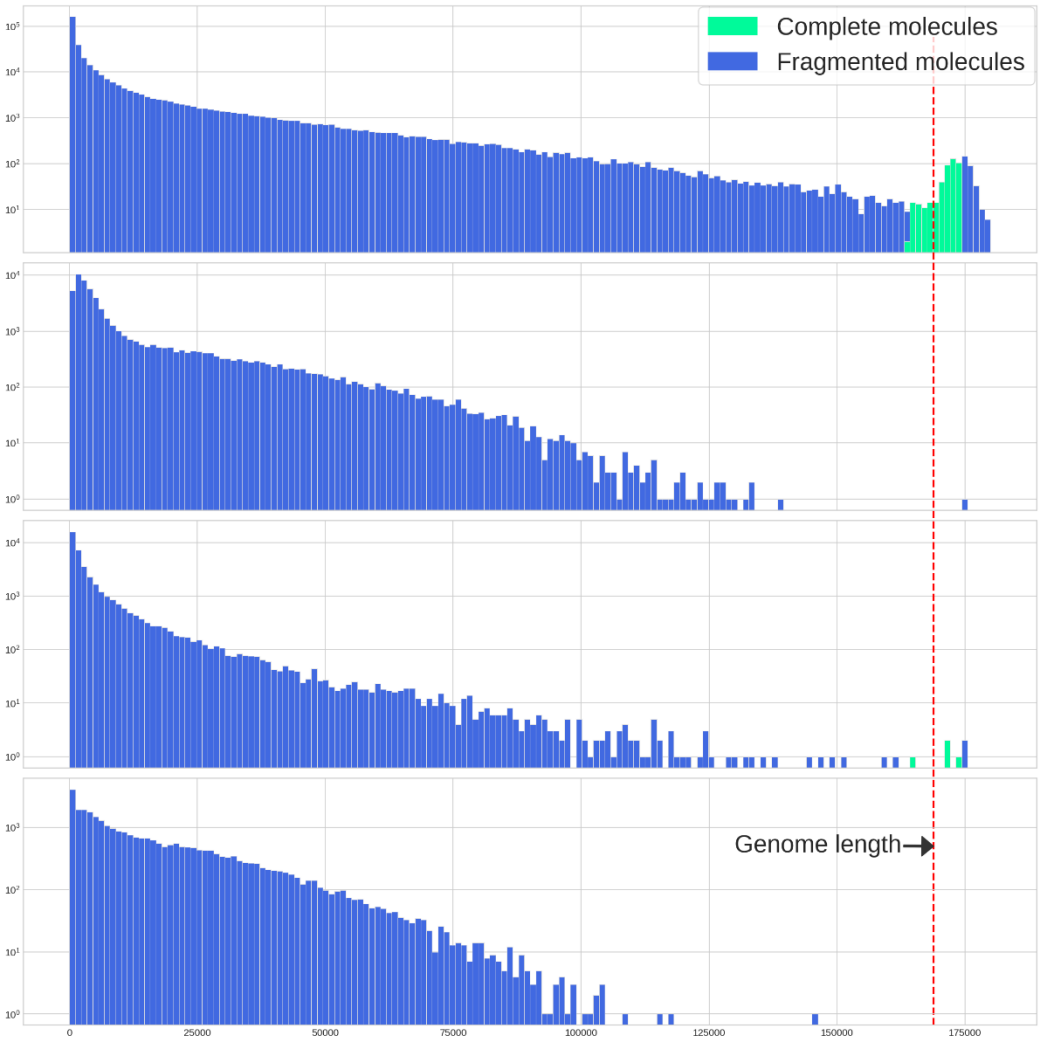


Figure 32. Size distribution of reads sequenced from T4 phages in 4 separate runs performed on standard R10.4flow cells. Y-axis in log10 scale.

Using these flow cells, we obtained one run in which a discernible peak of complete molecules can be found (Figure 32). Although this peak only constitutes 0.1% of the total reads, the sequencing of multiple molecules at this size indicates that this “ultra-long” length is not unreachable for. A similar phenomenon to the one observed on the plasmid pBR322 can be observed with a difference between the peak’s expected molecule length and the peak’s observed molecule length, but more important this time with a peak at about 174 000 instead of 168 903. This difference cannot be explained similarly as for pBR322 by the presence of

short soft clips added at the ends of the reads or an imbalance between the number of insertions and deletions due to the error rate of the machine. Indeed, we find a common pattern on these reads with one end of the read having a short soft clip (10-30bp) and the other end a very long soft clip (up to tens of thousands of kb). Manual examination of these long soft clips reveals the presence of long aberrant sequences in the read sequence with repeats over several hundred/thousand nucleotide bases or nucleotide pairs (e.g. 'AAAAAAAAAA' or 'ATATATATATAT'). The presence of these non-mappable elements on the genome disrupts the mapping process which fails to resume after such a long gap and therefore passes the rest of the sequence in soft clip. Our hypothesis on the origin of these aberrant sequences is the blockage of nucleotides during their translocation through the nanopore, potentially due to a problem related to the helicase, which maintains one or more bases in the ionic current crossing the membrane and whose variations are always measured by the sequencer resulting in the addition of aberrant sequences in the read sequence. This phenomenon is not specific to ultra-long molecules as we have observed it on some cDNA sequencing runs but seems to occur here on a larger quantity of reads in view of the shift observed in the peak of complete molecules. Although this defect results in unexpected sizes of complete molecules, the fact that the peak is relatively well defined and stands out in the distribution allows us to know that these are reads corresponding to complete DNA and therefore does not prevent the analysis of these samples.

We did not manage to reproduce this sequencing run on our next experiments, the result obtained for these runs being similar to those obtained on the Flongle chips, *i.e.* an almost total fragmentation of the sequenced molecules and only a few reads of size > 100-120kb.

We have not yet been able to determine what factor is causing this significant difference between the different sequencing runs, but it seems clear that it is in one or more of the steps of the library preparation protocol. We have not yet been able to identify what needs to be adapted to improve the sequencing, but this project is still in its early stages and this run containing a peak of complete molecules confirms that this is at least an achievable result and can probably be improved in the future.

Discussion

The long-read technology developed by Oxford Nanopore Technologies has been rapidly evolving since its commercialization began in 2015, at which time the low read quality scores and yields allowed little or no exploitation of the long-read concept[202,203]. Technological advances on this method have significantly increased the yield of sequencing chips, the quality of reads produced as well as the size of the molecules sequenced, with the longest reads produced reaching several Mb[204,205] but with the objective of obtaining large genome coverage and not to obtain a maximum of complete molecules.

Through our experiments, we were able to sequence DNA molecules irradiated at different doses and obtain a measure of radiation-induced fragmentation by the size distribution of reads and by the percentage of complete molecules. These results allowed us to extract consistent fragmentation probabilities on two different size models (pB322 and Lambda) using the same computational methods, thus confirming our ability to observe radiation-induced effects by this method. We then compared these results with simulations performed with Geant4-DNA on 3D pBR322 DNA models and this comparison allowed us to determine the parameter values (energy deposition threshold for a break and the minimum distance between the energy deposition and a base for an interaction) that best corresponded. Although more replicates are still needed to obtain more precise average values, these results confirm our ability to combine novel simulation methods of DNA fragmentation with a modern sequencing method that allows us to analyze DNA fragmentation from a perspective never before achieved.

It should be noted, however, that the ability to detect DNA fragmentation with this sequencer is strictly dependent of its method of reading DNA molecules. For example, it is impossible to distinguish double strand breaks (DSBs) from single strand break (SSBs) as they both result in an interruption of the read. There is also the fact that strand breaks will be missed because of the fact that only one of the 2 strands of the DNA molecule is captured by the nanopore and read. Therefore, in the case of a DSB, the molecule passing through the membrane simply stops at the break site but, in the case of an SSB, the result will depend on which strand is sequenced. If the strand carrying the SSB is translocated, the reading of the molecule will stop at this site which will be indistinguishable from a DSB. On the other hand, if the strand not carrying the SSB is sequenced, this break will not be detected and therefore not counted in the final reads. Similarly, any SSB theoretically present on the rest of the molecule would be missed as only 1 SSB could be detected per DNA molecule.

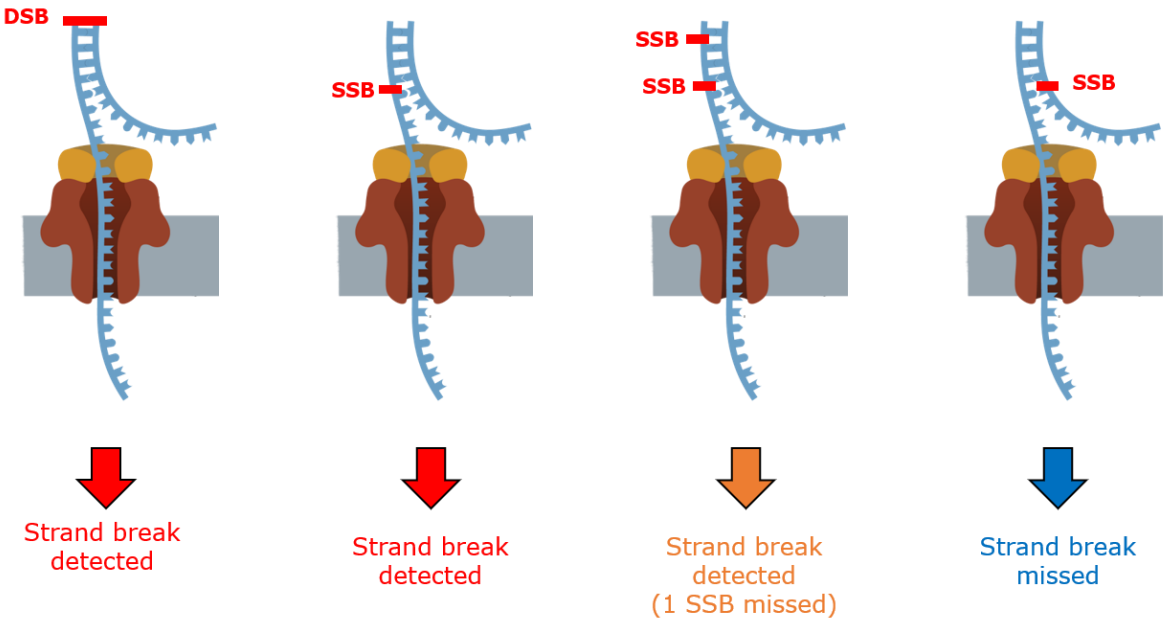


Figure 34. Possible scenarios of DNA strand break detection.

The progression towards the sequencing of "ultra-long" DNA molecules is still in its infancy and requires an adaptation of the classical library preparation protocol to the challenge of sequencing molecules >100kb in quantity and reproducibility. Current sequencing runs of the T4 phage have globally not yielded complete molecules, with the exception of one run that has shown encouraging results on the technical possibility of achieving better sequencing on this molecule. A possible hypothesis that could explain this absence of complete molecules is the circularization of the complete molecules of our sample, in particular during the ligation step of the adapters, the conditions being favourable to a ligation of the 2 ends of the molecule. This would mean that complete molecules are unavailable for sequencing and that only fragmented molecules, either present in small initial quantities or produced by mechanical breakage during library preparation or DNA node formation, are sequenced.

In summary, the Minion long-read sequencer proves to be an effective tool for measuring DNA fragmentation on molecules but only applicable for the moment to DNA molecules of length <50kb. Using the percentage of complete molecules per sample, we were able to trace back to probabilities of fragmentation caused by direct radiation-induced damage and the fragmentation can also be visualized with the naked eye on the read size distributions.

While we used this sequencer with the classical library preparation and sequencing protocols so far, other alternatives are possible such as 2D sequencing which sequences the 2 strands of the DNA molecules and could therefore bring new angles of study about the detection of SSBs compared to DSBs. The rapid technological advances in this technology should also help produce sequencing runs of increasing quality as new versions are released, thus allowing easier analysis of the longest DNA molecules.

It could be possible to consider performing next the same measurements on DNA molecules suspended in water in order to obtain, by comparison, the fragmentation probabilities caused by indirect damage, particularly the radiolysis of water also placing the issue of detection of modified bases, mainly oxidized, as a central issue in the study of DNA damage. The study of DNA fragmentation on longer molecules also remains a challenge. Several teams are working on alternative methods of library preparation in order to optimize the yield of long-read molecules[206]. It is therefore possible to continue with the T4 phage by adapting the protocol until a satisfactory result is obtained in order to unlock the ultra-long-read as a tool for measuring radiation-induced fragmentation.

Part II. Analysis of radio- and nano-induced cellular expression by transcriptomic analysis

When it comes to assessing the magnitude of biological effects caused by a factor on a living organism, DNA has historically been a frequent marker of biological damage, the measure of the risk caused by exposure to a factor being estimated on its capacity to damage or not the DNA because of the short- and long-term danger at the cell or organism level. However, apart from the activation of repair systems, the DNA damage is not representative of the overall cellular response to the stress generated and which may also have caused other types of damage in the rest of the cell. Furthermore, the possible cellular fates after DNA damage (cell death, survival with or without potentially deleterious long-term mutations) do not reflect the short-term state of the organism and the potential impact on cellular functions that may also result in longer-term damage.

The field of transcriptomic, the study of all RNA transcripts produced by transcription from DNA, allows us to approach this question of cellular response in a more global way by studying the cellular expression through mRNAs and thus observe potential changes in gene regulation in response to a given factor in what is called differential gene expression (DGE). Several techniques, the oldest dating from the late 1970s[207], were initially developed to study RNA:

- Expressed Sequence Tag: Sanger sequencing of short transcripts which were used for gene discovery and gene-sequence determination[208,209].

-Digital Differential Display: Quantification of transcripts by electrophoresis visualization[210].

-Serial/Cap Analysis of Gene Expression: Sanger sequencing of concatenated random transcript fragments to determine levels of gene expression[211].

-Microarrays: Measure of abundance of defined sets of transcripts by hybridization to complementary probe[212,213].

However, the most efficient and popular of those protocols quickly became RNA-Seq after it was first developed in the mid-2000s[214,215]. Unlike previous methods which were limited to specific genes, RNA-Seq used high throughput sequencing methods which allows for the sequencing of all mRNAs of the studied sample and a more efficient quantification of the number of transcripts per gene which in turn allowed for more accurate DGE studies. While initially based on the next-generation sequencing methods (NGS) which still constitute a majority of published transcriptomics studies[216], RNA-Seq has been through many developments with the introduction of 3rd generation sequencing technologies like Oxford Nanopore Technologies and PacBio.

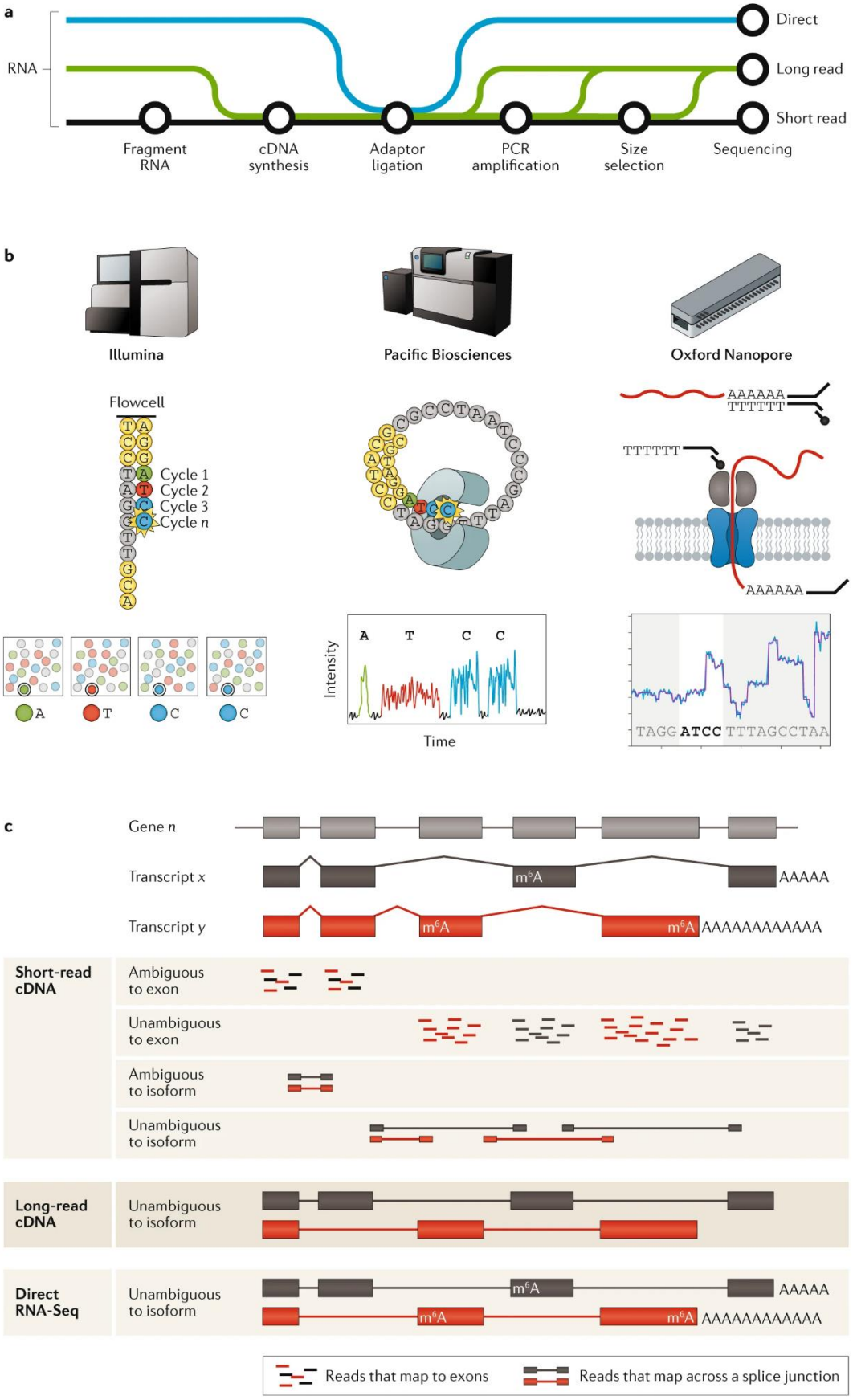


Figure 35. Overview of library preparation, sequencing and analysis for the three main current RNA-Seq methods. (Stark et al, 2019)

These 3rd generation methods, being long-read methods, have the capacity to sequence mRNAs in their entirety, thus allowing more precise exploration of the sequences of these molecules, for example by obtaining the complete sequences of complex RNAs to be sequenced with classic RNA-Seq methods or by having the possibility to measure the length of polyA tails, thus revealing disparities that impact the accessibility of certain genes for sequencing.

Another important advance made possible by this generation is the possibility of performing transcriptome sequencing without prior PCR amplification. The removal of this step allows to eliminate the risks of bias introduced by this technique (sequence artifacts[217,218] or unequal amplification[219,220]) and thus allow a better characterization of the transcriptome and its expression.

Oxford Nanopore Technologies' (ONT) method goes one step further by also eliminating the reverse-transcriptase step in its direct-RNA sequencing protocol that allows for the construction of a sequencing library from a native RNA sample[221]. The absence of alteration of the native sequence of these RNAs also makes it possible to directly detect modified bases on RNAs without prior chemical treatment, a process previously limited to DNA. The study of the epitranscriptome, the set of basic modifications affecting RNA, is a major challenge for the proper understanding of the impact of these modifications on the expression of these RNAs depending on their impact on the reactivity, structure and base-pairing interactions of the molecule[222]. At the time of writing of this document, the detection of these modified bases on the Oxford Nanopore method, although theoretically possible on all base modifications, remains limited to a few modifications and no official detection software has yet been officially developed by the company, although multiple publicly available tools have been developed by the community of users of this sequencer[223,224,225]. The reason for the lack of a reliable method at the present time comes from the increased difficulty to identify the sequenced bases of an RNA and thus the important uncertainty compared to a DNA molecule. Since the method of detection of modified bases is based either on the signal intensity of a base or on basecalling errors, this uncertainty related to RNA handicaps the detection of modified bases[226].

In view of the advances made in the field of transcriptomics, these methods are of great interest in the study of cellular pathways in correlation with results obtained via other analysis methods, in order to allow a more precise understanding of the observed phenomena. Several projects developed within the iRiBio team on the effects of ionizing radiation, metal oxide nanoparticles and the joint effects of these two factors can therefore benefit from the contribution of this transcriptome analysis by sequencing in order to correlate existing methods (micro-irradiation,

chemical imaging, flow cytometry, etc.) and their results with an analysis of the impacted cellular pathways. Oxford Nanopore's 3rd generation sequencing method has been implemented in 3 ongoing research projects in order to bring results to better understand the interactions between ionizing radiation, metal oxide nanoparticles and living organisms:

Questions asked in this work

Experiments in three projects were previously carried out in an interdisciplinary framework *via* the use of analytical methods specific to physics, chemistry and biology, including the analysis of the expression of certain genes by qPCR but not *via* a high-throughput method of transcriptomic expression analysis.

3rd generation sequencing methods were thus integrated in addition to those already practiced. RNA extractions were performed and these RNA libraries were then sequenced to study the cellular response to these experimental conditions studied in the different projects by comparison to the control condition using differential expression analysis.

In the context of this study, my work has been focused on:

- Implementation of a differential expression analysis pipeline for direct RNA-Seq data from raw data to exploitable results and statistics (differentially expressed genes, impacted pathways). Adaptation of this pipeline for use by non-bioinformaticians team members.

- Study of the genes and pathways impacted under the conditions of irradiation or exposure to the metal oxide nanoparticles studied and correlation with results obtained via other analysis techniques to evaluate the biological effects of the conditions studied.

- Evaluation of the relevance and efficiency of sequencing as a method for the analysis of radiation-induced or nano-induced damage in an interdisciplinary framework.

Part II.1: Study of radio-induced molecular damage on the RNA metabolism

Introduction

Ionizing radiation interacts with biological matter by depositing energy along its path resulting in excitations and ionizations. This deposited energy cannot be directly measured and is therefore generally quantified per unit of mass in the volume considered (absorbed dose). However, this quantification per unit of mass becomes less suitable when defining the dose at the scale of a cell, a cell nucleus or a DNA molecule ($< cm^3$). At these scales, the concepts of microdosimetry become essential to predict the biological response induced by a non-homogeneous dose deposit, taking into account the type of particle, its energy, the nature of the irradiated medium as well as the location of the energy deposit in the cell. Indeed, between a homogeneous irradiation on the whole nucleus and an irradiation restricted to a localized region of the nucleus, the energy deposition will be different, in spite of an equivalent dose, and will thus lead to different radiation-induced biological effects.

The study of this radiation-induced damage has historically been focused on DNA damage in order to estimate the risk of mutation and tumor development despite the distribution of the deposited dose being random and not necessarily reaching the DNA. Therefore, the impact on the rest of the cell, such as on other biological macromolecules (proteins, lipids, RNAs, etc.) and the consequences on the proper functioning of the cell remain to be defined.

It is with this objective of controlled and precise energy deposition that the microbeam irradiation line of the AIFIRA facility was created. This microprobe is directly connected with a fluorescence microscope which allows to visualize and target within a micrometric precision, defined sub-cellular compartments such as the nucleus, the mitochondria. Studies have been performed *in vitro* to monitor *in situ* and in real time the cellular response to DNA damage via the use of GFP-tagged proteins[227] and also *in vivo* with the use *Caenorhabditis elegans* 2-cell stage (Torfeh *et al*). These experiments have allowed over the years to precisely optimize the microprobe in parallel with the development of Geant4 codes allowing the calculation of the dose deposit according to the target used, the type and the number of particles (microdosimetric studies).

In this context, the objective with this microprobe is now to be able to selectively irradiate a cell type within an organism with the ultimate goal of being able to study the radiation-induced cellular response *in vivo* at the single cell level. This objective is accompanied by constraints on several points: (i) how to immobilize the target organism to precisely target the desired cell

type, (ii) which cell type to irradiate, (iii) how to synchronize the different organisms to obtain a more homogeneous response, (iv) how to analyze this radiation-induced response.

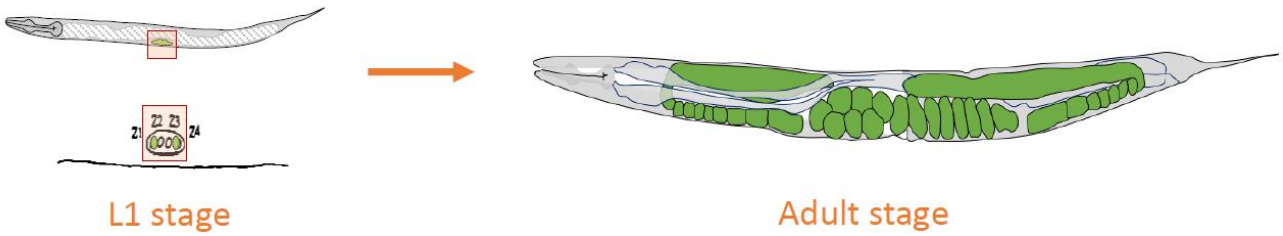
A protocol has recently been developed which aims to meet these criteria. For this purpose, adult *Caenorhabditis elegans* worms of a mutant line (GZ264) are used to reproducibly irradiate targeted cells in an *in vivo* organism with a controlled dose deposit. In this context of micro-irradiation, *C. elegans* offers many advantages to allow for a reproducible and specific irradiation of a specific region in an organism:

First, its small size and water-like density make that these worms do not have an important stopping power for 3 MeV protons (100 µm in water) thus avoiding the potential risk that a particle ends its path in the worm preceded by a Bragg peak. This means that all particles passing through the worm, although they do not all cross areas of equal density, deposit roughly similar doses of energy along their path thus minimizing the variability of dose deposit.

Second, the worms can be reversibly immobilized by placing them in a mixed medium of levamisole and poloxamer. Over a short period of time, this medium is not toxic for the worms and thus allows an efficient targeting of the microbeam on the targeted parts of the worms and thus avoids accidentally reaching non-target areas.

Third, through the use of a *C. elegans* mutant strain, GZ264, we have the capacity to precisely target the Z2-Z3 cells which are the precursor cells of the worm's reproductive system when it is at the L1 stage. This capacity comes from the GFP::PCN-1 transgene which is inserted in the GZ264 worms' genome, the PCN-1 gene is specific to the reproductive system and thus is only expressed in the precursor cells in young worms. Aside from offering a convenient way to target specific cells, this line allows the study of radiation-induced damage from several angles. The targeting of single cells allows the study of short-term effects (~3h post-irradiation) but also "pseudo-long-term" effects (because of the short development time of *C. elegans*) by studying the impact on the reproductive system of the worm once it reaches the adult stage (~3 days post-irradiation).

(A)



(B)

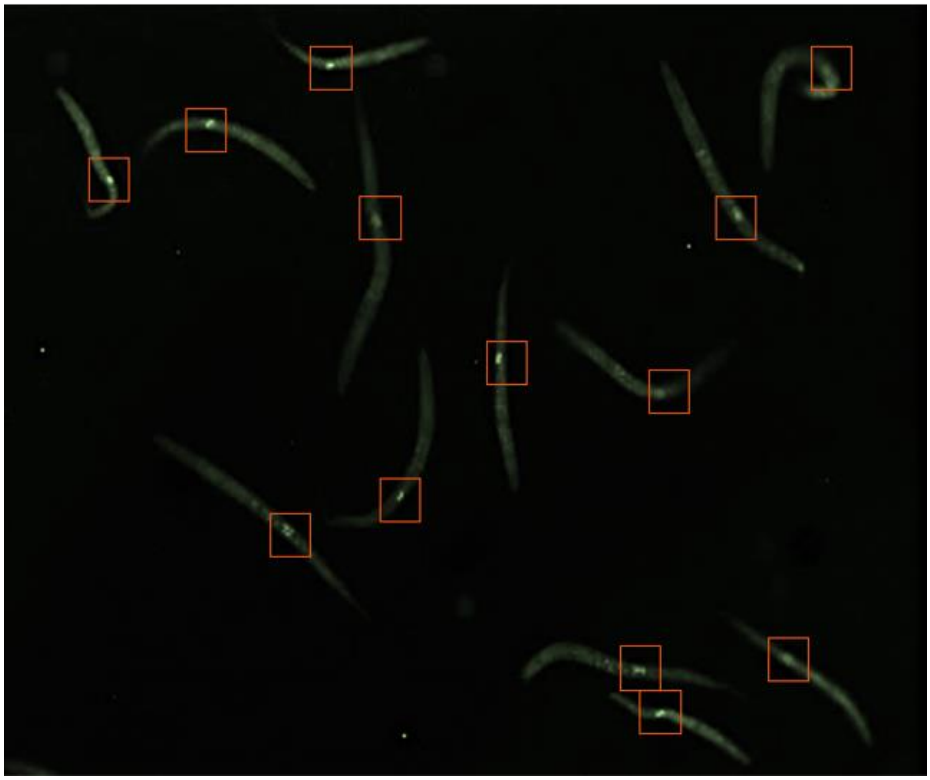


Figure 36. GZ264 mutant line (A) Schematic of the reproductive system at L1 and adult stages. (B) Worms as seen on the fluorescence microscope at the end of the microbeam line. Z2-Z3 cells targeted in red squares.

This experimental configuration allows us to study in comparison two types of irradiations: irradiation on the whole organism (macro-irradiation) versus irradiation on precise cells (micro-irradiation). Micro-irradiations performed at 300Gy on the Z2-Z3 allowed us to observe developmental abnormalities on the targeted region by microscopy on worms having reached the adult stage. By Hoechst^{33342} (DNA), GFP::PCN-1 and PhalloidinAF^{594} (Actin) labelling, we

were able to observe three phenotypes: a total absence of vulval development, an abnormal development resulting in non-functional cells and finally a vulval eversion characterized by cells behaving like tumor cells by their uncontrolled development. These results are similar to those obtained in the Seydoux et al[228] article which studies the development of the vulva and the anomalies caused by mutagenesis or laser ablation of precursor cells. In our case, however, our study by exposure to ionizing radiation allows us to study the damage in a dose-dependent manner and without directly killing the cells.

This result thus allows us to confirm our ability to produce observations of development anomalies of the vulva and gonads in an *in vivo* organism after targeted irradiation of the precursor cell of the reproductive system. By "bulk" sequencing of these macro- and micro-irradiated worms, we can thus evaluate the validity of the complete protocol (sample preparation, irradiation, sample recovery, RNA extraction, bioinformatics analysis) and compare the cellular response between these two types of irradiations at the organism scale. The long-term goal being to move towards "Single-cell" type technologies that allow sequencing to be performed on individual cells.

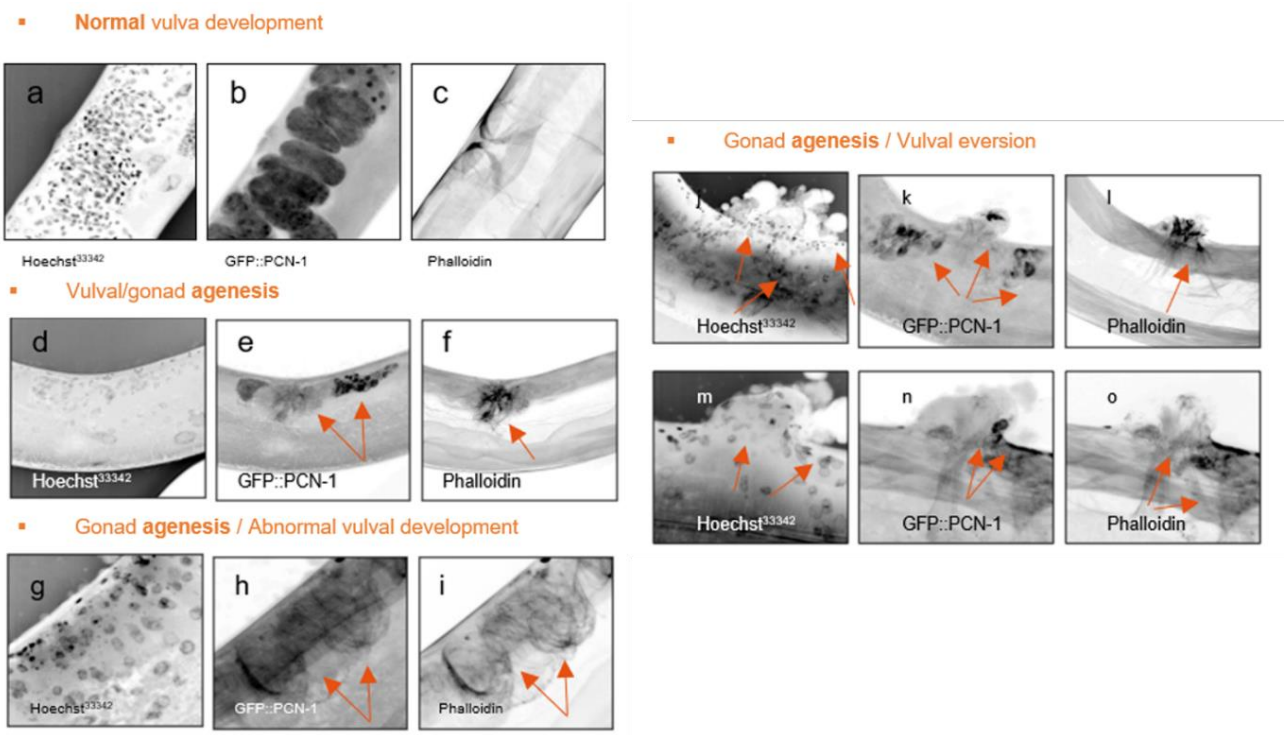


Figure 37. Phenotypes observed on L4 stage worms after irradiation at 300Gy of precursor cells of the L1 stage worm reproductive system. Produced by Hervé Seznec (iRiBio, LP2iB).

Materials and methods

1/ Worms strain and culture

C. elegans worm were maintained on nematode growth medium (NGM) agar plates and fed *ad libitum* with *Escherichia coli* strain OP50 at 19°C, according to standard protocols (Brenner, 1974). We used the transgenic GZ264 strain carrying an appropriate fluorescent marker (GZ264 *(isIs17[pGZ265:pie-1::GFP-pcn-1(W0D2.4)]).* The *Caenorhabditis* Genetics Centre (CGC, University of Minnesota) provided this *C. elegans* strains and the *E. Coli* OP50.

2/ Synchronization of large population of C.elegans (L1 stage)

The bleaching technique was used for synchronizing *C. elegans* cultures at the first larval stage (L1). The principle of the method lies in the fact that worms are sensitive to bleach while the egg shell protects the embryos from it. After treatment with an alkaline hypochlorite solution and rinsing, embryos were maintained on NGM agar plates without food, which allows hatching but prevents further development. Once hatched, the L1 larvae were kept onto NGM agar plates without food source until the irradiation time. Populations of young gravid hermaphrodites from standard, well-fed culture stocks were collected with M9 buffer (3 g/l KH_2PO_4, 6 g/l Na_2HPO_4, 5 g/l NaCl, 1 mM $MgSO_4$) and washed three times with sterile water to remove bacteria. Then, worms pelleted *via* centrifugation (2 min., 2000 rpms, room temperature) were treated with a freshly prepared alkaline hypochlorite solution (1.5 % (v/v) NaOCl, 1M NaOH). The suspension was swirled every 2 minutes with vortex-mixing (~6 min.). The released embryos were pelleted *via* centrifugation (2000 rpms; 2 min; 4°C) Supernatant was carefully discarded and embryos washed three times with M9 buffer followed by centrifugation. The pelleted embryos were suspended in 50 µL of fresh M9 and platted on an NGM agar plate without bacteria. The elapse time between hatching and irradiation was shortened in order to favor healthy conditions (> 24hours).

3/ Sample preparation and mounting for irradiation

We adapted the sample preparation conditions in order to use our custom-made support dish described previously by Muggiolu *et al* (2017*)* for micro-irradiation and live imaging. This sample holder provides a stable long-term environment for microscopic analysis and micro-

irradiation experiments. The day of irradiation, L1 Larvae were collected by centrifugation (2000 rpms; 2 min; 4°C) in cold M9 medium and resuspended in mounting medium (M9 supplemented with 0.25 mM Tetramisole hydrochloride (Sigma-Aldrich), 30% (v/v) Poloxamer-407 (Sigma-Aldrich). L1 were stored at 4°C to avoid their immobilization by the polymerization of the mounting medium. The number of worms is estimated to adjust the dilution volume in order to obtain ~100 worms per µl. 30 min before irradiation, an aliquot ~2 µl was directly deposited on a sterile 4-µm thick polypropylene (Goodfellow) and immediately covered with an afresh agar pad (3% (w/v) in M9) in order to maintain worm immobilized in a thin layer of medium. In order to prevent desiccation and contamination, the dish is closed with a glass side cover-slip. The elapse time between "mounting" and irradiation was shortened in order to favor healthy conditions (> 1hour).

4/ Irradiation

3 MeV protons (H^+, LET=12 keV.µm^{-1} in liquid water) were accelerated by a 3.5 MV electrostatic accelerator (Singletron, High Voltage Engineering Europa, The Netherlands) present in the AIFIRA facility. In order to target Z2-Z3 cells, the beam was strongly collimated to reduce the particle flux to a few thousand ions per second on target and focused using a triplet of magnetic quadrupoles to achieve a sub-micron resolution under vacuum. After extraction in air, the beam spot size is 1.5 µm. The delivered dose was controlled by counting the particles using a thin single-crystal chemical-vapor-deposited (scCVD) diamond membrane detector system dedicated for microbeam cell irradiation and fully compatible with micro-irradiation and online fluorescence time-lapse imaging.

5/RNA collection

After irradiation L1 worms were cultured during 3 hours on NGM plates seeding with *E. coli* OP50 strain, washed in RNase-free water and pelleted by centrifugation to remove bacteria. Briefly, worms were lysed by 20 cycles of freeze-cracking using a Dounce tissue homogenizer (Sigma Aldrich) and Total RNA were isolated with RNeasy Mini kit according to manufacturer's instructions (Qiagen). Total RNA integrity was assessed with the Agilent high-sensitivity RNA system for TapeStation.

6/ Sequencing

Whole-transcriptome cDNA libraries were first constructed from extracted mRNA using a PCR-cDNA barcoding kit (SQK-PCB109; Oxford Nanopore Technologies) following the standard associated protocol. Two libraries, each containing barcoded cDNAs from all the studied experimental conditions, were produced from different biological samples. The libraries then were sequenced on a Mk1C MinION using R9.4.1 flow cells with a min_qscore of 7 and live basecalling until the flow cell runs out of active pores.

7/ Bioinformatic analysis

The fastq files were merged and mapped to the WS283 *C. elegans* reference transcriptome using minimap2 with the option "–ax map-ont" and the alignment files were processed using samtools. Alignment results were converted into an expression matrix with an associated metadata table using a custom Python script. The differential expression analysis was then performed in R using the edgeR[229] and limma[230] libraries. The expression matrix was inserted in a DGEList object (edgeR package). Genes with a <1 CPM (counts per million) were removed and gene counts were then normalized to log2-CPM (functions calcNormFactors and voom). A linear model was fitted for each gene (function lmFit) and contrasts between experimental conditions were extracted (functions makeContrasts and constrasts.fit). The log odds of differential expression for each gene was then determined using an empirical Bayes test (function eBayes) and pvalues adjusted using the Bonferroni method[231] (function p.adjust) and differentially expressed genes obtained (function decideTests). Enrichment analysis was then performed using the gprofiler[232] g:GOSt functional profiling method with default settings. The codes used are available at:

https://github.com/pelotbdr/iribio_scripts/tree/main/bulk_transcriptome_analysis

Experimental results

Caenorhabditis elegans worms at L1 stage were irradiated under two different configurations: Macro-irradiation on the whole sample or Micro-irradiation on manually targeted Z2-Z3 cells (precursor cells of the worm reproductive system). These irradiations were performed at 3, 30 and 300 Gy in order to study the effects at different doses and RNA extractions are performed 3h after irradiation in order to try to detect the beginning of the response to radiation-induced damage. However, due to the time required to manually target the worms on the Micro-irradiation conditions, the number of worms in these samples is much lower than in the Macro-irradiation ones (around 250 in micro- samples and 2500 in macro- samples), in order to avoid a large time difference between the first irradiated worm and the last irradiated worm within the same sample.

1. Libraries quality

The difficulty to produce these irradiation series as well as the time needed to set up the protocols used mean that we currently have very little material. Furthermore, this study is performed on L1 worms which due to their smaller size and lower cell number contain less mRNA than more developed worms, resulting in significant difficulties in extracting sufficient amounts of mRNA to perform sequencing which requires a minimum of 500 ng for direct-RNA or 50 ng for RT-PCR. The first runs of direct-RNA sequencing did not yield analysable results, the sequenced reads had very low-quality scores and the yield was largely insufficient to study their expression. We therefore sought to test the quality of our mRNAs by using an Agilent chip on an Agilent Bioanalyzer 2100 to migrate 18S and 26S ribosomal RNAs extracted from the samples and observe the level of fragmentation.

On the macro-irradiated series, we found the RNAs to be totally degraded on the 30 and 300 Gy samples. This result is not surprising considering the doses used but it means that it is not possible to sequence these samples, the mRNAs being too damaged. The quality was good on the 0 Gy and 3 Gy samples with scores sufficient for sequencing. However, we had technical problems during our RNA storage which led to the loss of the 3 Gy sample. We were therefore only able to sequence the 0 Gy (labelled Control in next figures) sample from the entire Macro-irradiated series at the time of writing.

In the two Micro-irradiated series, the RNAs were of acceptable quality and thus usable for sequencing. The quality score distribution differs slightly however between the two micro-irradiated series with the series A RNA appearing as slightly more damaged which could indicate an eventual issue during the RNA extraction or a previous problem with the worm population.

In summary, RNAs from two batches of Micro-irradiated samples and one control sample from a macro-irradiated batch presented sufficient RNA quality to be extracted for sequencing. However, because of the small amount of RNA available from our samples, which complicates Direct-RNA sequencing, we decided to use an RT-PCR in order to start from smaller amounts of initial material. This was performed using a barcoding kit in order to sequence all the previously mentioned conditions on the same flow cell. 3 standard R9.4.1 flow cells were used until the pores on the library produced were exhausted.

Flow cell 1	Read count	Average size	Median size	Mapped %
Control	1 863 112	622.02	595	66.96
0Gy(A)	582 966	542.85	540	38.83
3Gy(A)	183 857	548.73	588	37.21
30Gy(A)	222 179	463.40	380	36.53
300Gy(A)	478 665	508.87	427	45.97
0Gy(B)	2 470 710	553.10	603	37.93
3Gy(B)	1 312 704	591.24	620	37.46
30Gy(B)	1 907 595	518.20	454	50.97
300Gy(B)	1 983 189	481.58	412	44.32

Flow cell 2	Read count	Average size	Median size	Mapped %
Control	2 655 204	623.30	595	66.94
0Gy(A)	917 443	538.27	527	39.05
3Gy(A)	303 392	550.06	588	37.97
30Gy(A)	364 649	461.86	378	36.74
300Gy(A)	770 846	506.65	423	45.96
0Gy(B)	2 507 552	550.05	598	37.86
3Gy(B)	1 435 338	591.49	621	37.96
30Gy(B)	1 931 771	515.89	447	50.23
300Gy(B)	2 143 462	479.01	408	43.66

Flow cell 3	Read count	Average size	Median size	Mapped %
Control	1 402 829	526.31	481	62.29
0Gy(A)	433 842	505.96	466	37.54
3Gy(A)	388 897	432.98	353	31.59
30Gy(A)	592 615	351.60	284	28.69
300Gy(A)	1 025 370	387.85	307	36.67
0Gy(B)	2 349 949	488.42	449	35.95
3Gy(B)	1 148 784	505.73	506	36.54
30Gy(B)	1 589 785	455.52	373	46.62
300Gy(B)	2 012 329	378.43	303	35.74

Table 3. Results of the sequencing runs of the micro-irradiated samples on the 3 flow cells used.

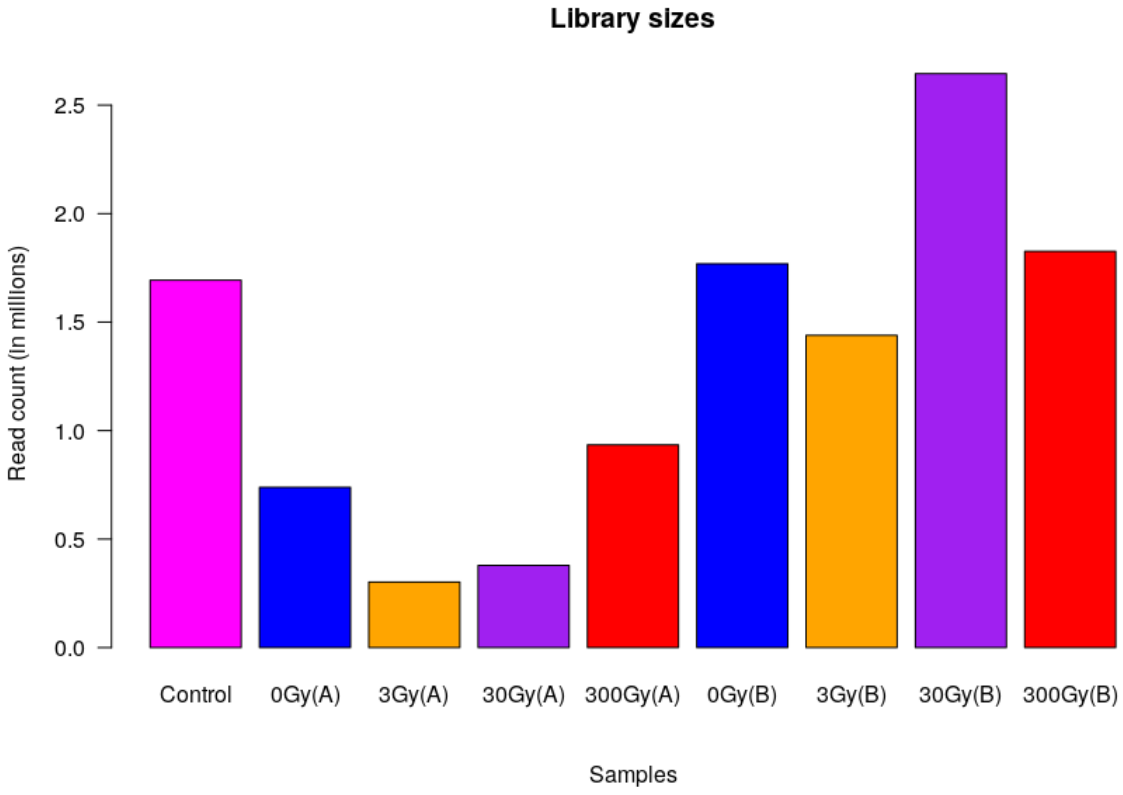


Figure 38. mRNA library sizes in Control (Macro 0Gy) and Micro-irradiated samples.

The number of reads obtained differs significantly between the different conditions (Figure 38). It can be observed that overall, the Micro-irradiated series A, which was the series with the

lowest RIN scores, is the one in which the least number of reads could be produced especially on the 3Gy and 30Gy conditions. It should be noted that out of all the libraries, on average over half of the reads produced could not be assigned to reference mRNAs, being either ribosomal RNAs or *E.coli* contaminants from the worms culture medium.

The low amount of starting material and large percentage of unusable reads results in disappointing library sizes on some conditions despite the use of 3 sequencing chips. The rest of the conditions produced more satisfactory quantities of reads indicating that the problem observed with the A series is correctable in future experiments.

2. Differential expression analysis

We then performed a PCA to observe the global distance between the different irradiation conditions but also between the Macro series control sample and the Micro series control samples. Indeed, the preparation of these samples differs slightly, the micro-irradiated being exposed to UV light for the duration of the targeting-irradiation manipulation, and the amount of initial material being very different, we seek to determine if these differences result in different cellular expressions.

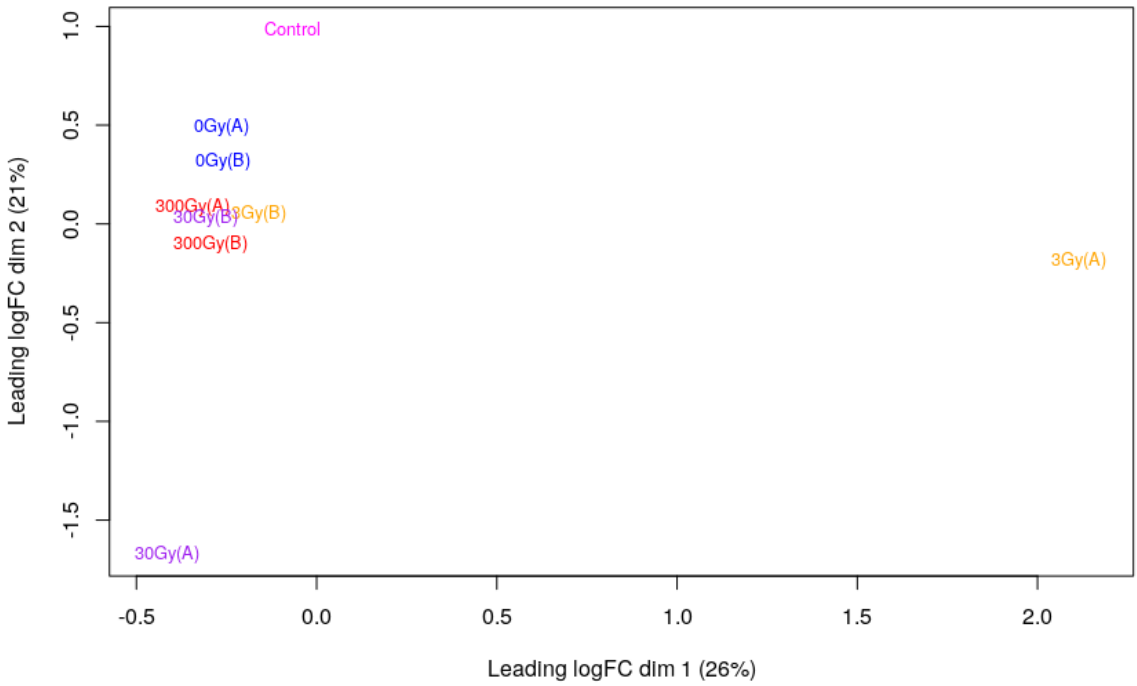


Figure 39. PCA of libraries of Control (Macro 0Gy) and Micro-irradiated samples using all expressed genes.

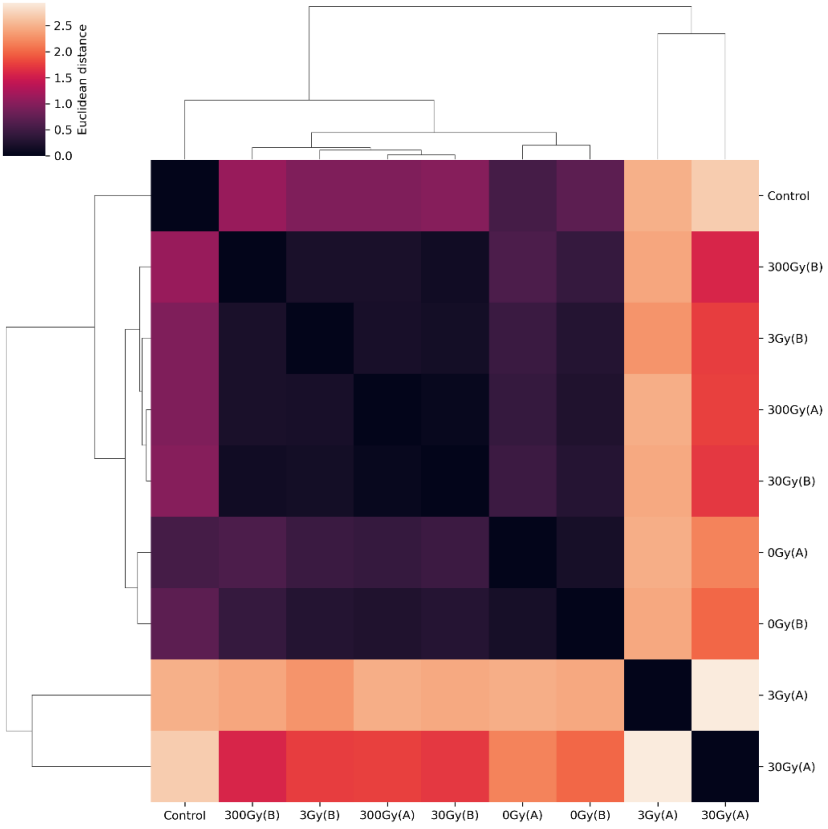


Figure 40. Clustermap of Euclidean distances extracted from the PCA plot.

It can be observed on this PCA first of all that two conditions, the 3Gy and 30Gy Micro-irradiated samples of the A series, are found apart from the other samples which are rather grouped (Figures 39-40). This distance, which normally reflects a difference in cell expression in the sample, can however be questioned in this situation. These are indeed the two samples with the lowest number of reads, and therefore considered at a lower degree of confidence, and it can also be seen that the samples of the same condition but for the B series do not show the same distance.

This distance seems all the stranger since the 300Gy Micro-irradiated sample of the A series is grouped with the other samples, which would indicate a significant cellular response at 3Gy and 30Gy but not at 300Gy.

On the Macro-irradiated series control sample, it can be seen that it is globally located with the Micro-irradiated control samples but that it sits somewhat apart from them. Normally no difference should be found between these samples, as their preparation is the same. The most

likely hypothesis for this difference seems to us at this stage to be the quantity of initial material which could potentially hinder the sequencing of certain weakly expressed genes.

To further investigate the differences between the different sequenced conditions, we then performed a differential expression analysis on all conditions compared to the 0 Gy condition of the Micro-irradiated series (Figure 41).

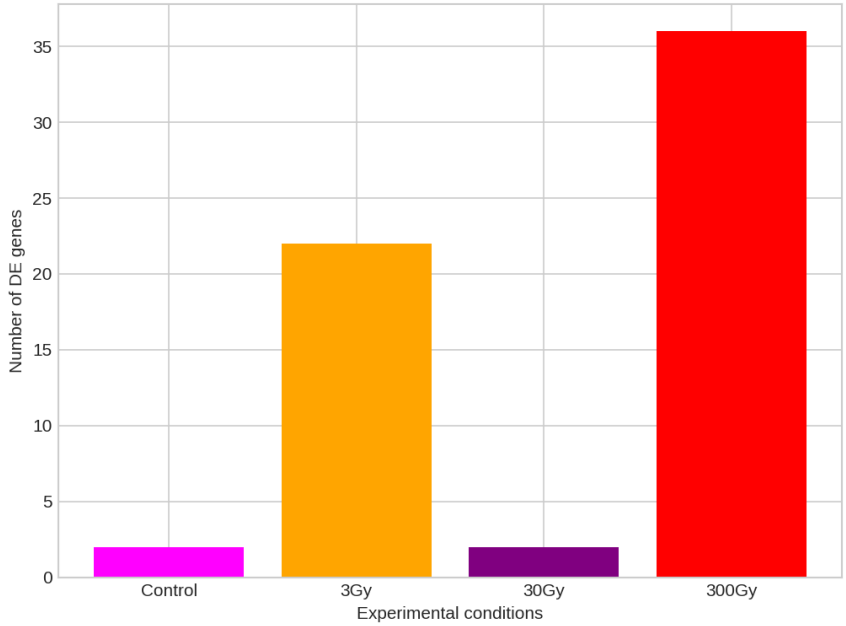


Figure 41. Number of DE genes per experimental condition sequenced compared to the Micro-0Gy condition.

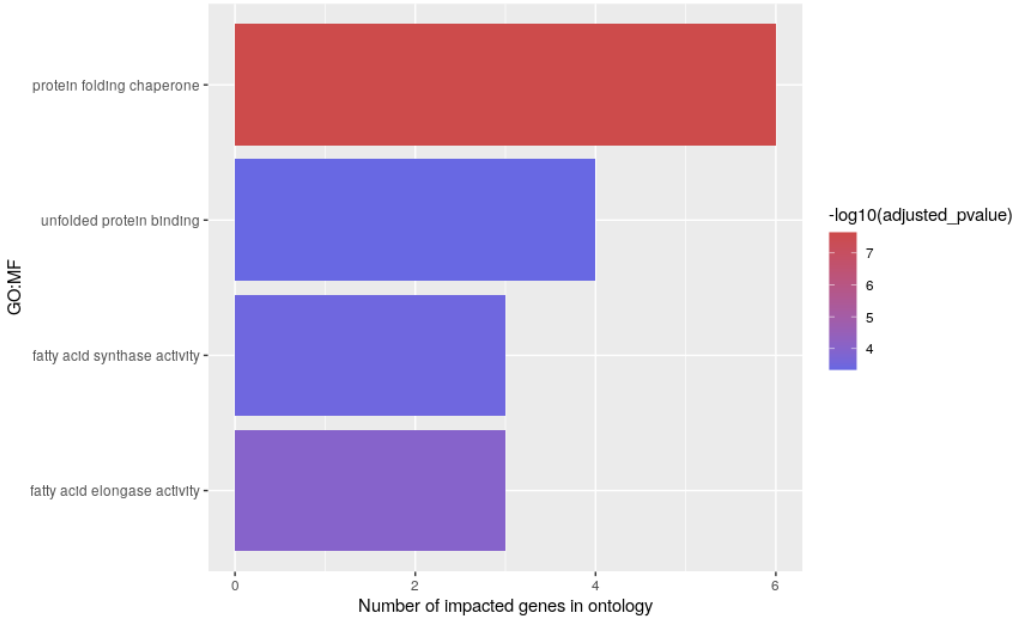
For all conditions, a low to very low number of DE genes is obtained, indicating at best a moderate difference of detectable cell expression with the controls of the Micro-irradiated series. In the Macro-irradiated series control we find 2 DE genes, which means that even if the global expression is similar between these two controls, slight differences can be identified, confirming the distance observed on the PCA. Although this difference is not substantial, it seems to indicate that, as the protocol stands, the amount of initial material can affect the measured cellular response. For the 3Gy, 30Gy and 300Gy Micro-irradiated samples, 22, 2 and 36 DE genes are obtained, respectively, which would indicate at first glance a stronger impact on cellular expression of a 3Gy dose than a 30Gy dose but must be put into perspective with the difference in expression observed on the PCA between the 3 and 30 Gy replicates which may affect the calculation of DE genes, as the samples are normalized and therefore have the same weight in the statistical analysis and an aberrant replicate could thus distort the results. For the 300Gy condition, the 36 DE genes appear to us to be more credible than previous doses, with samples from both runs behaving similarly on the PCA and the series A sample not

suffering as few reads as the other irradiated samples in the run. An interesting point about the DE genes found in the 3 Gy condition is that, although their validity may be questioned due to the quality of the series A replicate, several genes are found in common with the 300 Gy condition as well as some genes from similar gene families. We find 3 genes common to both conditions coding for heatshock proteins, T27E4.3, T27E4.8 and T27E4.9, as well as genes of the nlp (Neuropeptide-Like Protein) and ugt (UDP-GlucuronosylTransferase) families. These elements common to the two conditions could therefore give credibility to the DE genes identified in the 3 Gy condition as being authentic elements of a cellular response and not simply a result of aberrant samples. It cannot be excluded at the time being that these genes are part of a stress response not the irradiation itself but to the irradiation setup, although the absence of these DE genes for the 30 Gy condition prevents any certainty on the question. In the absence of additional sample batches, we cannot draw any conclusions at this time about whether these DE genes in the 3 Gy condition are indeed the product of a radiation-induced cellular response or a problem arising from sample production at some stage in the protocol.

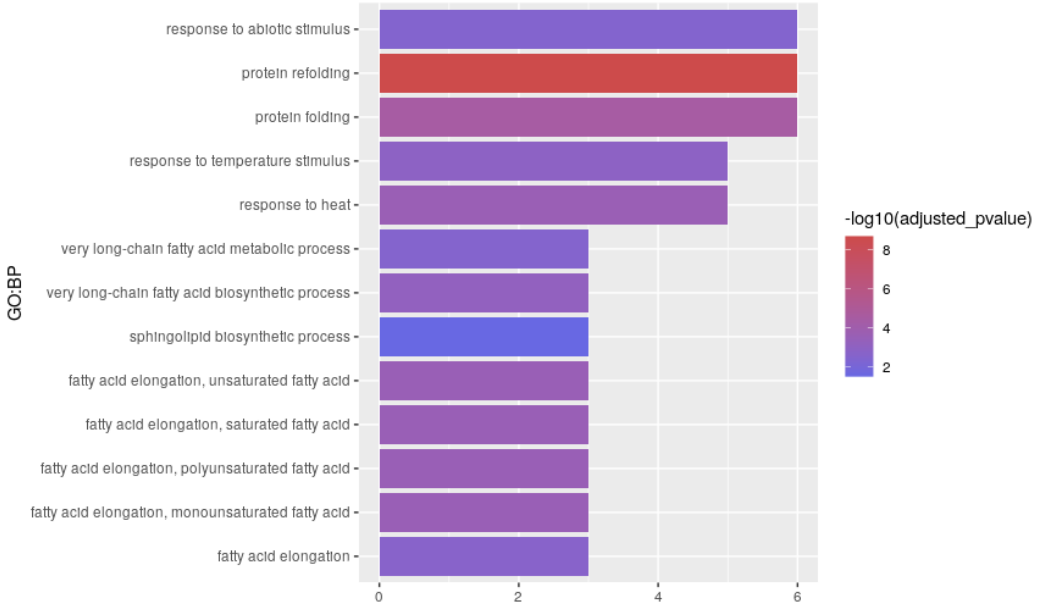
3. GO enrichment analysis

From the previously obtained DE genes, we looked for the cellular pathways corresponding to these impacted genes using the g:Gost function of gprofiler to obtain impacted GO terms. The DE genes in the 3 Gy condition were subject to a significant degree of doubt and the 30 Gy condition had only 2 DE genes, so we focused on the 300 Gy condition.

(A)



(B)



(C)

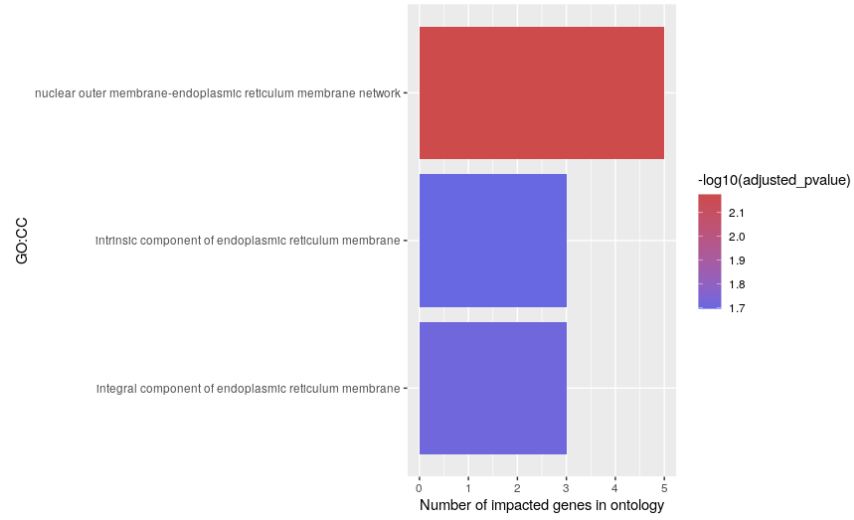


Figure 42. GO terms from (A) Molecular Factor (B) Biological Process and (C) Cellular Component significantly impacted (adjusted pvalue <0.05) in Micro-irradiated at 300Gy condition.

The ontologies significantly impacted in this condition can be summarized in two main pathways: protein folding and fatty acid metabolism (Figure 42). These cellular pathways have in common their location at the endoplasmic reticulum, as confirmed by the impacted CC terms, which manages protein folding and lipid metabolism in the cell.In the CC terms, we also find the plasma membrane which can be reasonably related to the disruption of lipid metabolism due to the structure of this membrane.

We can therefore identify in this cellular response of Micro-irradiated worms to 300Gy a possible stress of the endoplasmic reticulum characterized by a disturbance of the protein metabolism as well as of the lipid metabolism thus affecting the lipid bilayer of the plasma membrane.

Discussion

The analysis protocol of micro-irradiation of *Caenorhabditis elegans* worms having been validated from the preparation of the samples to be irradiated to their return to culture after irradiation in a controlled and targeted dose, the next objective was to extend this protocol to the RNA extraction and transcriptomic analysis of these worms. From the samples produced, we were able to obtain sufficient mRNA to perform sequencing with RT-PCR and subsequently identify a cellular response in samples micro-irradiated at 300Gy. This observed response contains coherent elements in the cellular response to ionizing radiation, the impact on lipid metabolism[233,234,235] and the unfolded protein response both linking to potential endoplasmic reticulum stress[236] which are elements of this response already identified in the scientific literature. It lacks however other cellular pathways expected for high dose irradiation such as the response to DNA damage or cell cycle regulation[237].

A question can be raised about the time applied between irradiation and RNA extraction. Indeed, the current delay being 3h, it is possible that the cellular response has not yet been fully implemented or that several cellular pathways are impacted at different times, which would potentially lead to different results if, for example, delays of 6h or 12h were used. We can also note that among the impacted cellular pathways we do not see any that concern the functioning of the reproductive system of the worm, despite the fact that the precursor cells of this system are targeted. This should be all the more visible as for the 300 Gy dose, the microscopic observations had revealed a significant impact on the development of the reproductive system of the worm once the L4 stage was reached, either by the total absence of development of this system or by a quasi-tumoral development of certain cells. One would therefore expect to see a difference between a control where the reproductive system develops normally and a system where the cells composing it have been deeply impacted. If again the use of bulk sequencing could explain the lack of response observed, it is also possible that our protocol as it stands is not optimal to study the impact on the reproductive system at this stage of development. It has been shown that during starvation of L1 worms, the development of the reproductive system is paused and only resumes once access to food is restored[238]. Our worms being exposed to a starvation that can last a few hours when they are placed on the irradiation support, it is possible that this starvation episode stops the development of the reproductive system and that the 3h during which the worms are put back in culture before the RNA extraction are not enough to restart the development. Again, the solution of choosing a different time between irradiation and RNA extraction could be considered to see if there is a difference emerges.

Technical problems around the macro-irradiated series, too much degradation of mRNAs for the 30Gy and 300Gy samples as well as a storage problem on the 3Gy sample, prevented the sequencing of this series which could have been used as an element of direct comparison with the micro-irradiated series. This step should be an important objective on the next series of samples produced to provide a basis for comparison between the two types of irradiations.

In summary, we have the ability to dose-control micro-irradiate worms on specific cells, return these worms to culture, extract the mRNAs and sequence them to seek to identify an eventual cellular response. While we did not detect all the expected components of a radio-induced cellular response, this first batch of sequencing samples confirms that a response is observable by sequencing a complete organism despite the fact that only a localized region has been irradiated.

This protocol can be still be improved by refining the RNA extraction protocol to increase throughput which would result in more robust samples and cellular response, with the added objective of performing Direct-RNA sequencing and going towards the detection of modified bases. New configurations could also be considered by changing the time between irradiation and RNA extraction to determine which timeframe best allows observation of the radiation-induced response. It could also be considered to use other *C.elegans* mutant lines, in which other cell types would contain a fluorescence causing transgene, to target other tissues that could be easily dissected (head neurons, intestine, etc.) in order to sequence only the irradiated tissues.

Overall, the limited observable response on all conditions on our current configuration, confirms the need to progress towards single-cell techniques in order to make the best use of the micro-irradiation tool and the developed protocol.

Part II.2: Study of the *in vivo* nano-induced cellular response combined with microscale detection and quantification of nanoparticle exposure

Introduction

Nanotechnologies are increasingly present in all areas of everyday life. Nanoparticles are used in the composition of many manufactured products from the food, cosmetics and medical industries. Nanoparticles are defined as objects with at least one of their dimensions smaller than 100nm. They have particular physicochemical properties such as a greater surface reactivity, a greater surface/volume ratio and multiple shapes, which makes them extremely interesting in the industry. Moreover, as their size approaches that of proteins, DNA and other biological entities, they are now used in a wide range of biological applications as therapeutic agents, antimicrobial agents, transfection vectors and as fluorescent agents. The toxic effects of nanoparticles during their internalization in cells is a demonstrated phenomenon, the extent and nature of this damage strongly depending on many factors: size, shape, surface charge and surface area, hydrophilicity, ability to form aggregates, solubility, etc[239]. However, this toxicity is not uniform in a given organism, as the bioavailability of nanoparticles dictates their distribution in a given tissue or in given cells and thus their capacity to cause cellular damage.

It is in this context that chemical imaging techniques can be used to monitor nanoparticles and their possible internalization in cells. Nuclear microprobe analysis including PIXE (Particle-Induced X-ray Emission), RBS (Rutherford Backscattering Spectrometry) and STIM (Scanning Transmission Ion Microscopy) were thus initially carried out on human keratinocytes and made it possible to detect and quantify the internalization of TiO_2 nanoparticles (Muggiolu *et al*). These methods were thus put in common with biological analyses, such as qPCR, revealing stress of the endoplasmic reticulum and a rupture of the calcic homeostasis during the internalization of this same type of nanoparticles[240]. However, a weakness of *in vitro* experiments is the ability to faithfully reproduce constant cellular conditions over time (homeostasis) and to control the individual level of exposure at the cellular level in a sample, leading to risks of dose heterogeneity, particularly at low doses. The absence of cellular communication phenomena also limits the scope of the observed biological response.

The extension of these studies thus focused on carrying these assays to an *in vivo* model: *Caenorhabditis elegans* to try to detect a more complex biological response that would allow us to understand the molecular and cellular mechanisms of the biological effects induced by exposure to NPs. The aim with this model is not only to determine the toxicity of TiO_2 NPs on the *C. elegans* development but also to relate biological observations to the NPs physiochemical

features (morphology, state of agglomeration). TiO_2 NPs were synthetized with controlled shapes, sizes, crystallinity and surface chemistry to test the toxicological impact of these parameters on TiO_2 NPs and compare with commercial TiO_2 NPs. Two types of TiO_2 nanoparticles were used for this study: P25 beads and Titanate scrolled Nanosheets (TNs) which were the ones causing the highest toxicity on the *in vitro* model. These nanoparticles were added to the normal culture medium of the worms for a period of 24 hours on worms at different stages of development

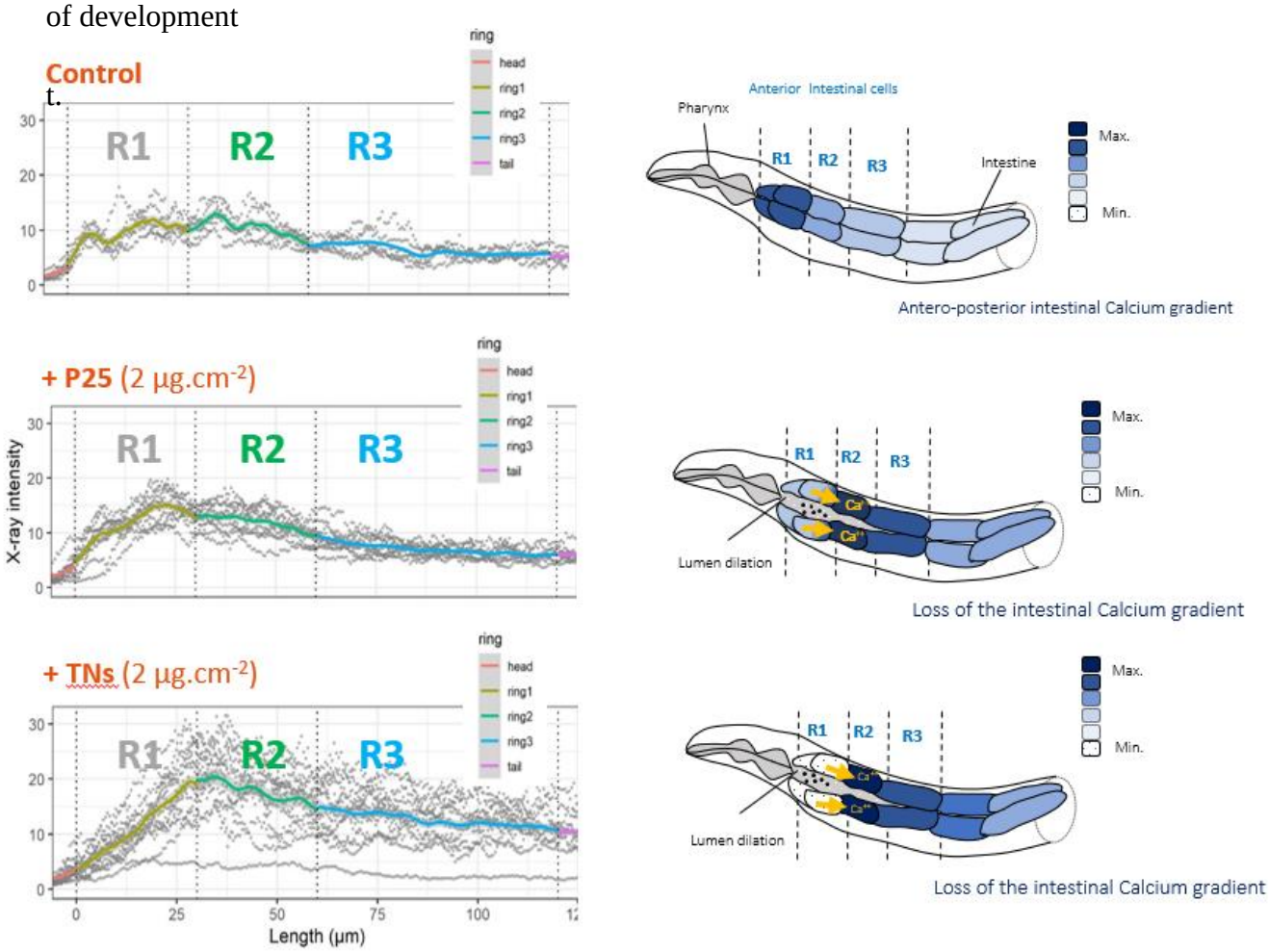


Figure 43. Calcium levels obtained by PIXE chemical quantification in rings R1, R2 and R3 of the intestine of *C.elegans* exposed to P25 and TNs nanoparticles. Produced by Guillaume Devès (iRiBio, LP2iB).

Flow cytometry studies, allowing to measure the size of the worms, revealed that the addition of these nanoparticles caused a delay in the growth of the worms and further chemical imaging revealed a consequent impact on calcium and potassium distribution in the intestinal cells.

However, the detection and quantification of nanoparticles by chemical imaging revealed that if they are well ingested by the worms, they are not internalized in the intestinal cells of the worm but are traced along the intestine, sometimes forming slow-moving clogs of nanoparticles. At first glance, it would seem that these nanoparticles cause toxicity at the organism level, leading to developmental delays despite the absence of internalization, a phenomenon normally required for nano-induced toxicity. Although, the developmental delays observed could also be caused by an unplanned starvation of the worms, as the nanoparticles are not nutritious and their presence may physically prevent the ingestion of the worms' food.

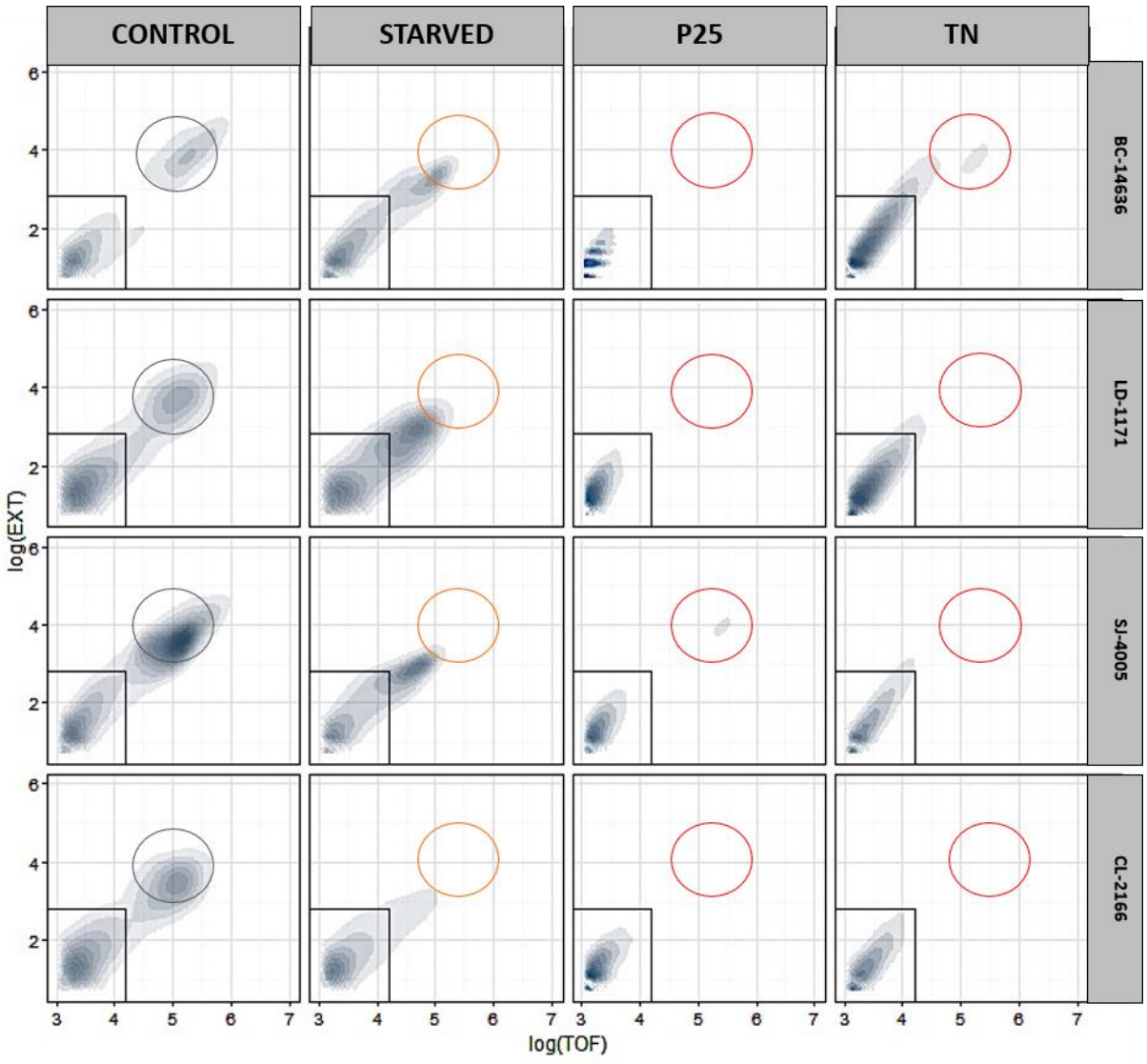


Figure 44. Flow cytometry results of L1 stage *C.elegans* exposed to P25 or TNs nanoparticles. TOF: Time of Flight (~worm length), EXT: Optical Extinction (~worm density). Produced by Guillaume Devès (iRiBio, LP2iB).

The methods used so far in the project do not allow to answer these questions, the study of the cellular pathways impacted in the different conditions could therefore provide new information to clarify these results. Transcriptome sequencing was therefore used in this project with 3 objectives: (i) To compare the transcriptomic expression of worms exposed to nanoparticles with starved worms in order to estimate their proximity, (ii) to detect if the presence of nanoparticles and their toxicity leads to a cellular response from the organism and (iii) identify the cellular pathways possibly impacted by nanoparticles.

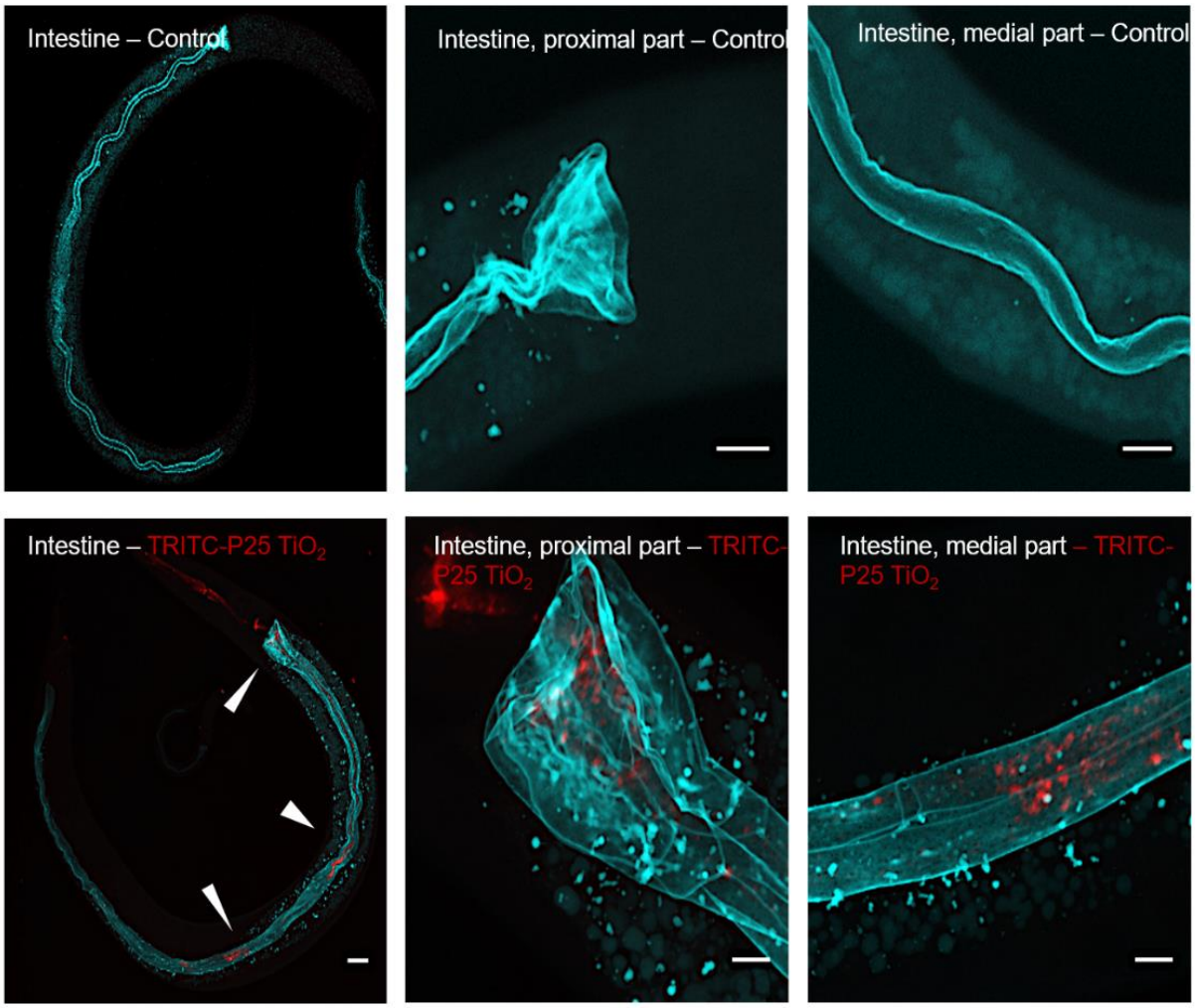


Figure 45. Distribution of P25 TiO_2 nanoparticles (red) in a *Caenorhabditis elegans* worm gut from a fluorescent gut mutant line. Image produced by PIXE imagery by Guillaume Devès.

Materials and Methods

1/ Worm strain and culture

C. elegans worms of the N2 Bristol strain were maintained on nematode growth medium (NGM) agar plates and fed *ad libitum* with *Escherichia coli* strain OP50 at 19°C, according to standard protocols (Brenner, 1974). The *Caenorhabditis* Genetics Centre (CGC, University of Minnesota) provided this *C. elegans* strain and the *E. Coli* OP50.

2/ TiO_2 NPs synthesis and characterization

P25 nanoparticles (P25, AEROXIDE) were kindly provided by Degussa/Evonik and used both for biological tests and as precursor for all the synthesis performed in this study. Titanate scrolled nanosheets (TNs) were produced via hydrothermal process described by Kasuga et al. (ref) Briefly, 2 g of P25 were introduced in a 50 mL Teflon lined autoclave with 28 mL of 10 M sodium hydroxide solution, sealed and heated at 130 °C for 20 h. The white precipitate was washed with nitric acid (0.1 M) and water for neutralization and identified as titanate scrolled nanosheets (TNs). All the synthesized NPs were kept in aqueous solution avoiding aggregation issues. Mass concentrations were measured by drying a known volume of solution and weighting the extracted powder. Suspensions with concentration of 1 $mg.mL^{-1}$ were finally produced, sonicated and kept in the dark. The surface modification was performed in two major steps (Simon et al., 2011). Briefly, 600 mg of TiO_2-NPs were mixed with 3 mL of ammonium hydroxide 25 % (v/v), 100 µL 3-aminopropyltriethoxysilane and absolute ethanol and stirred at room temperature for 48 h. The suspension was then heated at 100 °C for 2 h under reflux. The white powder was then washed 5 times with absolute ethanol. The powder was then added to a 30 mL of Na_2CO_3 (0.01 M) aqueous solution containing 2.5 mg of tetramethylrhodamine-isothiocyanate (TRITC) and stirred for 48 h. NPs were finally washed several times with Na_2CO_3 (0.01 M) solution and then with milliQ water. The TRITC-TiO_2 NPs were dried for 24 h under vacuum at 40 °C and kept in dark.

3/ TiO_2 NP exposure

TiO_2 NPs were resuspended in ultrapure water, sonicated before each experiment and immediately spread over NGM plates with *E. Coli* OP50. Age-synchronized cultures were isolated from gravid adults, treated with *bleaching mixture* (NaOH 10N, Hypochlorite solution 5%, (v/v)) (Sigma Aldrich). Resultant eggs hatched and gave an age synchronous population

and L4 stage worms were then exposed to two types of TiO_2 NPs (P25, TNs) at a 2 µg.cm^{-2} concentration during 24h.

4/ Sequencing

Whole-transcriptome RNA libraries were first constructed from extracted mRNA (from 5-10 000 worms) using a Direct-RNA kit (SQK-RNA002; Oxford Nanopore Technologies) following the standard associated protocol. Two libraries were produced for each experimental conditions from different biological samples. The libraries then were sequenced on a Mk1C MinION using R9.4.1 flow cells with a min_qscore of 7 and live basecalling for > 24h.

5/ Bioinformatic analysis

The fastq files were merged and mapped to the WS283 *C. elegans* reference transcriptome using minimap2 with the option "–ax map-ont" and the alignment files were processed using samtools. Alignment results were converted into an expression matrix with an associated metadata table using a custom Python script. The differential expression analysis was then performed in R using the edgeR and limma libraries. The expression matrix was inserted in a DGEList object (edgeR package). Genes with a <1 CPM (counts per million) were removed and gene counts were then normalized to log2-CPM (functions calcNormFactors and voom). A linear model was fitted for each gene (function lmFit) and contrasts between experimental conditions were extracted (functions makeContrasts and constrasts.fit). The log odds of differential expression for each gene was then determined using an empirical Bayes test (function eBayes) and pvalues adjusted using the Bonferroni method (function p.adjust) and differentially expressed genes obtained (function decideTests). Enrichment analysis was then performed using the gprofiler g:GOSt functional profiling method with default settings. The codes used are available at:

https://github.com/pelotbdr/iribio_scripts/tree/main/bulk_transcriptome_analysis

RNA base modifications were analyzed using the standard Nanocompore[241] protocol which makes use of Nanopolish[242] to realign raw fast5 data with fastq reads and then extract events of interest in the signal.

Experimental results

In the prolongation of the studies on the toxicity of nanoparticles within the iRiBio team and carried out until now on *in vitro* models, experiments were carried out on *Caenorhabditis elegans* worms cultured on media containing TiO2 nanoparticles in order to study the presence or not of induced toxicity on this model. If these experiments were carried out on different lines and at different stages of development, the study of the transcriptomic response was limited to the N2 line on worms at the L4 stage and exposed to nanoparticles in their culture medium for 24 hours before extracting the RNAs.

Two types of differently shaped nanoparticles were used in this study, P25 beads and TNs nanosheets. The transcriptome of starved worms was also extracted in addition to those of the exposure conditions to nanoparticles. The libraries produced were sequenced on standard R9.4.1 chips and each condition was realized in biological duplicate.

1. Differential expression analysis

A PCA was first performed on all the sequenced samples in order to observe the resulting distance between the conditions and if any clustering can be observed.

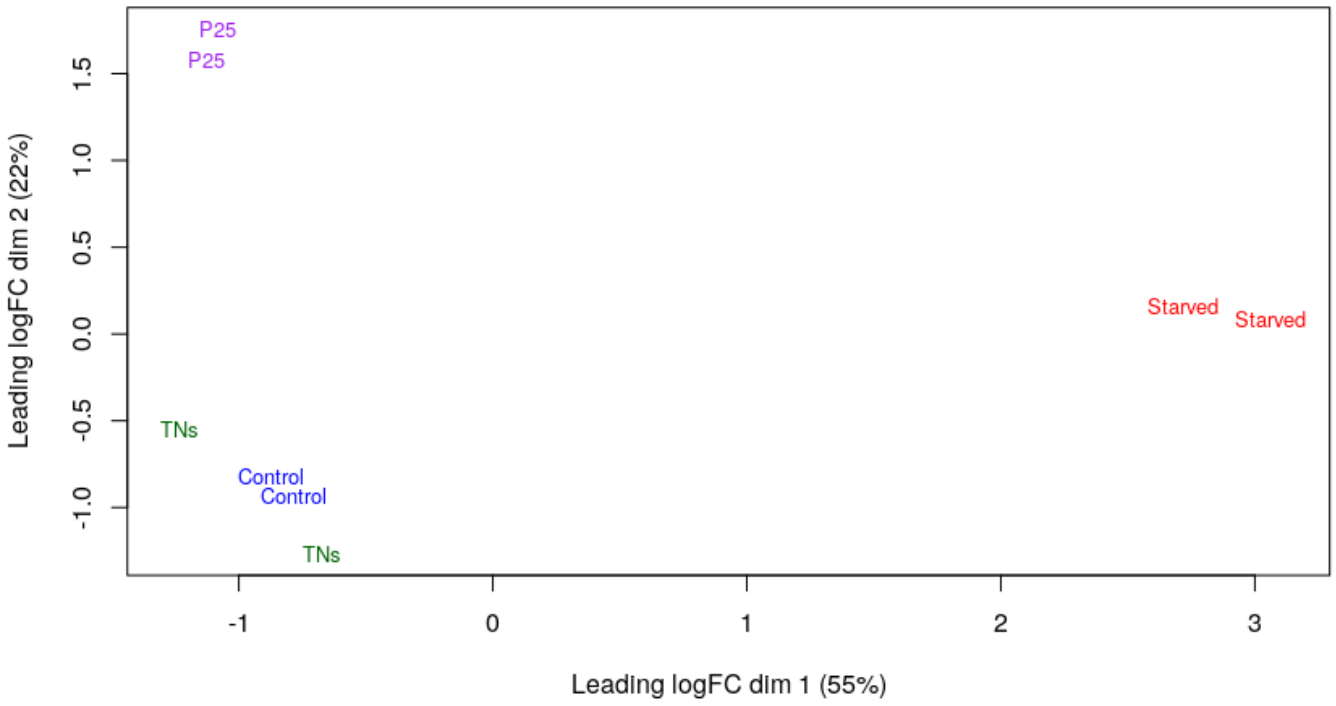


Figure 46. PCA of libraries of all NPs exposure conditions using all expressed genes.

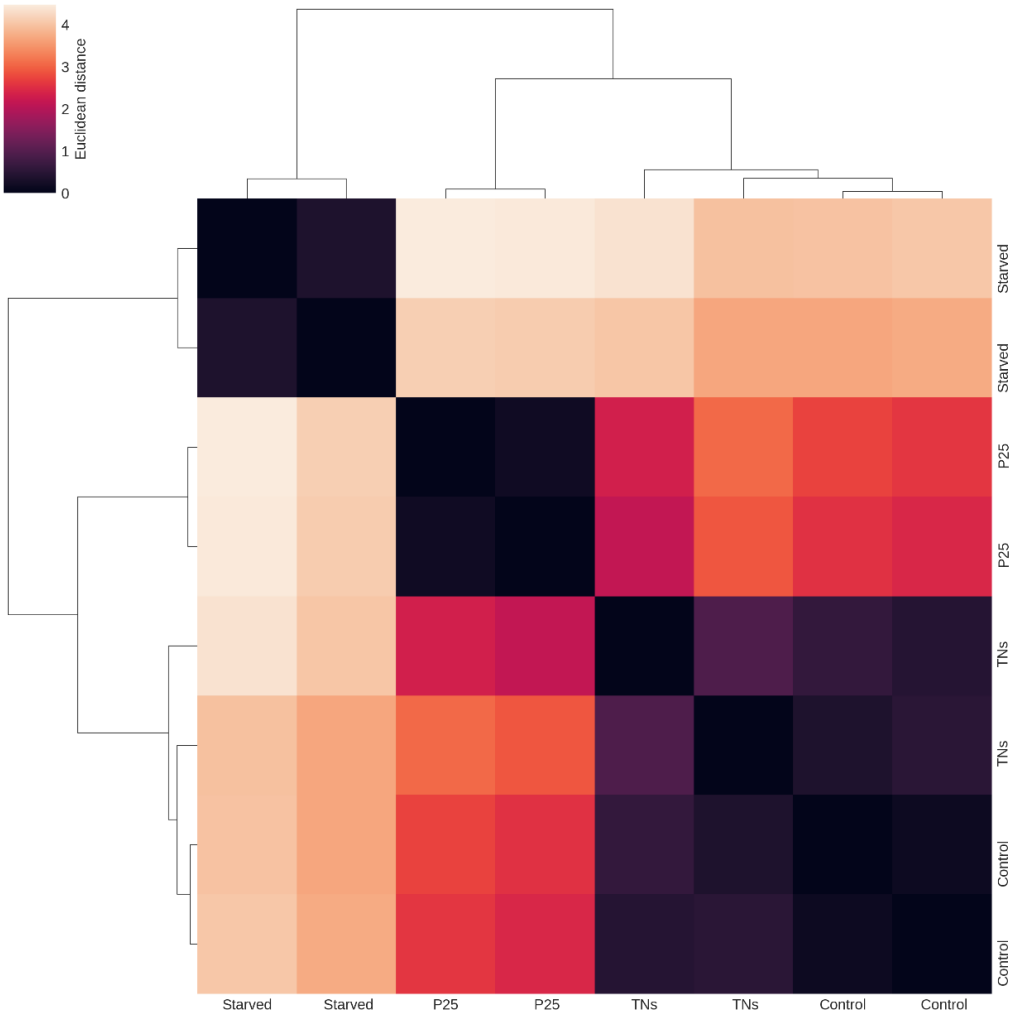


Figure 47. Clustermap of Euclidean distances extracted from the PCA plot.

It can be observed from the PCA that the condition replicates cluster well, with just a slight distance between the two TNs replicates, suggesting good reproducibility across samples. We also observe that the P25 and Starved conditions are distinctly at a distance indicating a significant difference in expression. The TNs condition on the other hand shows a small distance to the control indicating an overall similar expression (Figures 46-47).

To further investigate the difference between these datasets and the control, we continued with a differential expression analysis to search for the impacted genes.

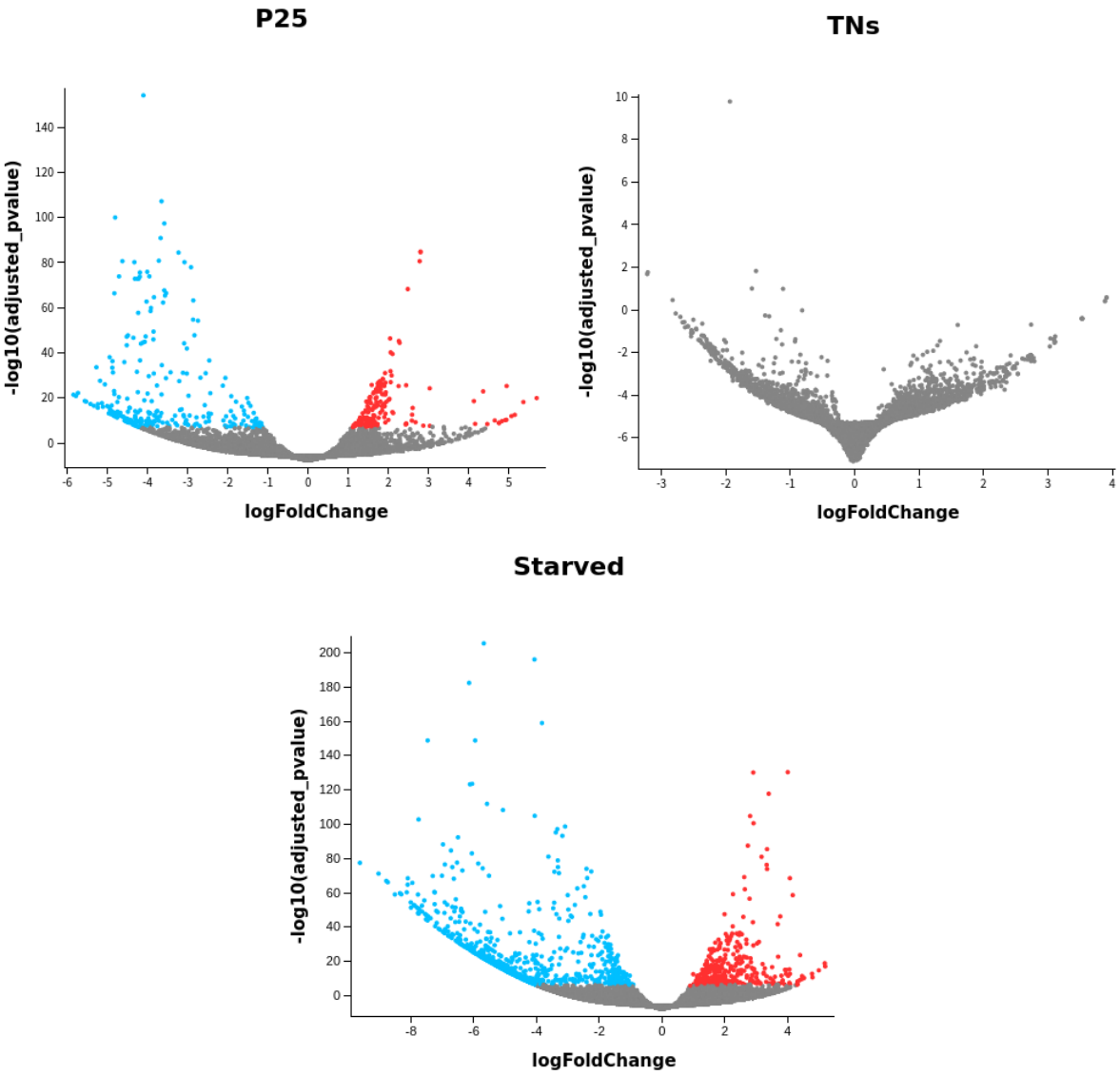


Figure 48. Volcano plots of DE genes of P25, TNs and Starved conditions compared to the control. Blue: under-expressed, Red: over-expressed.

419 DE genes were found in the P25 condition of which 259 were under-expressed and 160 were over-expressed, confirming a significant difference in expression with the control. For the Starved condition, 1191 DE genes were found of which 863 were under-expressed and 328 over-expressed, apparently indicating a stronger cellular response than in the P25 condition (Figure 48). For the TNs condition, in accordance with what was observed via PCA, no DE genes could be identified, thus showing an expression similar to the control worms.

2. Comparison between Starved and P25 exposure cellular response

Based on this finding that the P25 and Starved conditions have a different cellular expression than the control, we seek to determine if this modification is similar in nature between these two conditions. To do so, we compare them in two ways: from the DE gene lists against the control (Figure 49) and by performing a differential expression analysis between the two conditions directly (Figure 50).

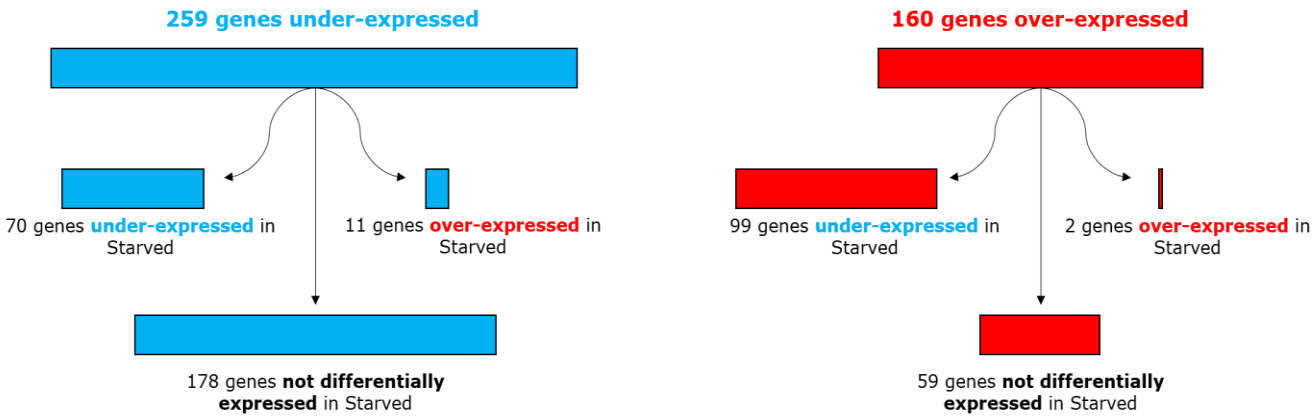


Figure 49. Comparison of DE genes from P25 condition compared to control and Starved condition compared to control.

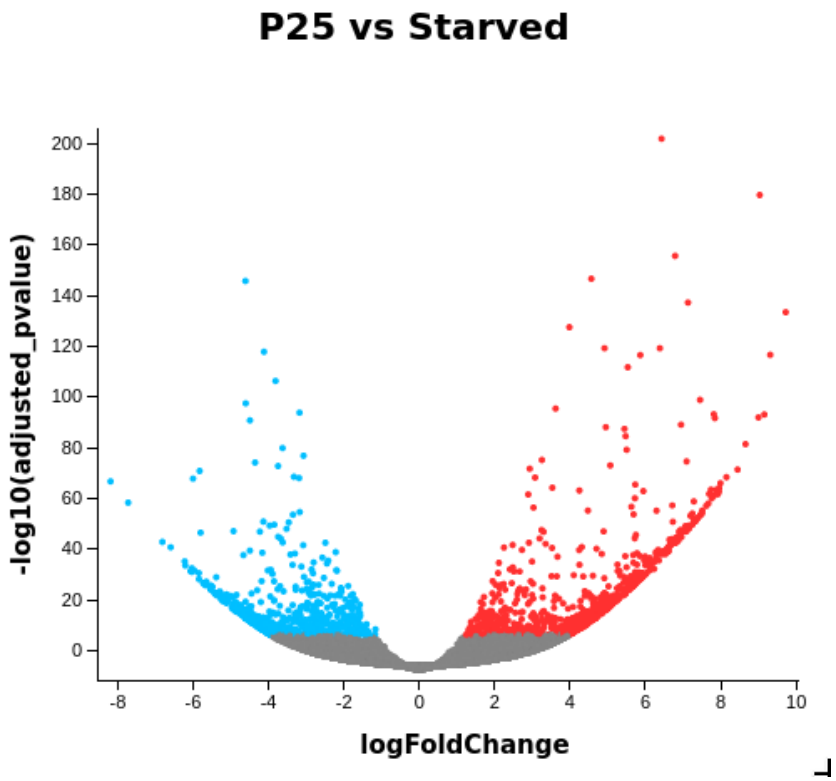


Figure 50. Volcano plot of DE genes from P25 condition compared with Starved Condition. Blue: under-expressed, Red: over-expressed.

Comparison of DE genes to control between the two conditions reveals a significant number of genes in common but that most of these genes are impacted differently by condition. Thus, we find 72 similarly impacted genes and 110 inversely impacted genes while 237 DE genes found in the P25 condition are not in the Starved condition, i.e. more than the majority of them. This first comparison thus seems to indicate that part of the cellular response is indeed common to both conditions but that it differs significantly. Differential expression analysis between P25 and Starved conditions also results in a significant difference with 1363 DE genes of which 737 are overexpressed and 626 are under-expressed. This number of DE genes is even larger than between the Starved condition and the control indicating a strong difference in cellular expression.

In summary, if part of the cellular response seems to be common between the P25 and Starved conditions with nearly a hundred commonly impacted DE genes, it remains however strongly different between these two conditions.

3. GO enrichment analysis

Following this finding, we then looked at the impacted cellular pathways from the DE gene list of the P25 condition by obtaining the significantly impacted GOs with the g:Gost function from gprofiler.

(A)

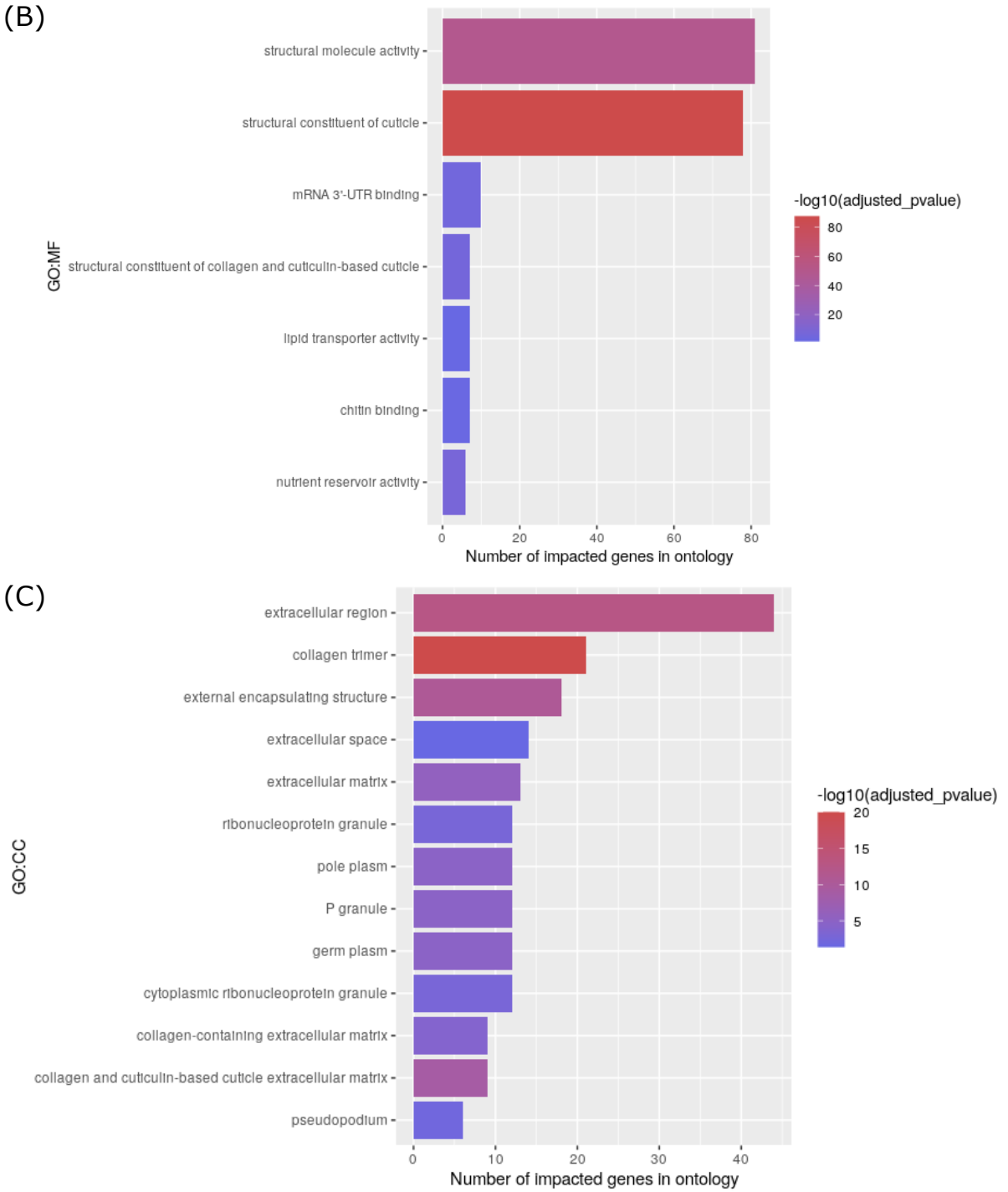


Figure 51. GO terms from (A) Biological Process (B) Molecular Factor and (C) Cellular Component significantly impacted (adjusted pvalue <0.05) in P25 condition.

Considering the impacted GO terms, we can observe that the cellular response to P25 nanoparticles seems to be articulated around 4 main pathways: worm development, reproduction, molting cycle and the response to unfolded proteins (Figure 51). Several terms related to collagen and lipids are also present and potentially related to the molting cycle and

the cuticle which is composed of a complex collagen structure and a lipid layer. The molting cycle itself can also be grouped with the cellular pathways of development, with molting occurring at each stage of worm development.

In summary, the cellular response to nanoparticles in the P25 form is clearly distinct from the cellular response to starvation and is characterized by cellular stress that strongly disrupts normal worm function at the sexual reproduction and developmental levels. Exposure to TNs, on the other hand, does not appear to lead to a global disruption of the worm's cellular expression.

4. RNA base modifications detection

Although several RNA modifier detection software programs have been developed by users of the Oxford Nanopore Technologies sequencer, the lack of a formal tool complicates the application of this technology to our study. We tried to use several of these tools (xPore, Yanocomp, nanoRMS, Tombo) but difficulties in installing/using the software finally led us to use nanocompore. Unlike the other tools mentioned above, it does not allow to determine the presence or not of a specific modified base but rather to identify positions on which the presence of a modified base is probable with regard to the intensity of the electrical signal and the duration of the time lapse before the passage to the next position. This software is built on the basis of the nanopolish tool which processes the electrical signal and detects events that may correspond to baseline changes, the comparison between two conditions is then performed by nanocompore.

The comparison of reads from the Control and P25 conditions allowed us to evaluate the characteristics of 370 312 positions on 500 genes. Out of all these positions, only 1 could be identified as significantly different and is found on the B0041.4.1 (rpl-4) gene. The absence of more modified bases is not surprising because although nanoparticles can cause oxidative stress, and thus base modifications, their limited presence in the intestinal lumen and the use of bulk sequencing severely limit the amount of observable modified bases. The use of this tool has at least allowed us to become familiar with this type of data and to define the procedure for future analyses of this type of data.

Discussion

Sequencing of worms exposed to nanoparticles in their culture medium revealed the presence of a significant cellular response to this external stimulus in the case of P25 beads, but an absence of observable response in the case of TNs. These results are to be put in context with the distribution of these nanoparticles in the studied organism, the latter not being internalized in the intestinal cells of the worms when they are ingested but rather staying in the intestinal lumen. The difference in response between P25 and TNs could therefore be explained by a difference in distribution due to physical factors. P25 beads have a conformation that is easier for worms to ingest and present little or no risk of blockage in the intestinal lumen. The TN nanosheets, on the other hand, have a more complex shape which could influence their distribution in two ways: blockage of the nanoparticles in the pharynx due to congestion or disinterest of the worms which manage to identify them as non-nutritive elements. Chemical imaging by PIXE to see the distribution of these nanoparticles allowed to observe cases of worms having ingested TNs but this technique cannot realistically be carried out on the whole sample in a reasonable time frame and thus does not permit to quantify on the totality of the worms which percentage has ingested or not the TNs. It is therefore possible that this type of nanoparticle causes cellular stress when ingested but that only a small proportion of worms have ingested them which at the scale of the entire sample make their cellular response invisible.

In any case, in the case of the P25 beads, we were able to identify a theorized but as yet unproven phenomenon of the presence of a cellular response of an organism exposed to TiO_2 nanoparticles without the latter being internalized[243], as confirmed by PIXE chemical imagery, and excluded the possibility of it being due to a starvation phenomenon, going against the theory that internalization is necessary to cause cellular damage.

If we were able to observe this response on the scale of the organism, it is probable that it is stronger on the cells in direct proximity with the nanoparticles, the intestine and the pharynx. It could therefore be considered for future studies to move towards single cell methods or dissection of worms to extract the intestine and perform sequencing only on these cells in order to obtain a higher level of definition of the cellular response and potentially also to study it more easily in the case of TNs in case of cell blockage.

Part III. Evaluation of single-cell RNA-Seq applicability in a low yield and high complexity experiment

Introduction

In the context of our study of radio- and nano-biological damage, we aim to characterize the transcriptomic response of specifically targeted cells within an otherwise non-irradiated organism. The interest behind this objective lies in the innate heterogeneity of ionizing radiation and metal oxide nanoparticles deposition on cells. The micrometric scale capabilities of the microbeam lines of the AIFIRA facility allow us to either irradiate specific cells or to quantify their exposure to nanoparticles. We therefore seek to combine this capability with a transcriptomic analysis of the cellular response. While recent advances in sequencing techniques have provided access to increasingly detailed analysis of transcriptomic expression, the measurement of this expression is limited by the mixing of mRNAs from multiple cells during an RNA extraction performed on a sample. As a result, the expression measured is an average of individual expressions, called "bulk" sequencing[244]. There are methods to limit the variance of measured expression between cells by performing RNA extractions only on tissues of interest but there may be significant differences even between cells of the same tissue, masking the individual expression of these cells.

To overcome the technical limitations of bulk sequencing, single-cell sequencing methods have been developed to obtain the transcriptome expression of individual cells in a given sample. Methods already existed to measure the expression of certain genes at the single-cell level, such as single-cell qPCR which were first used in the early 1990s[245,246]. However, single-cell sequencing is the first method that makes it theoretically possible to obtain the expression of all the genes expressed within a cell and therefore to measure all the range of variations according to their situation (stress, gene knock-out, etc.)[247].

The first single cell sequencing was published in 2009 in order to study rare cell types in mouse blastomeres by separating the cells into individual tubes[248]. The goal was also to evaluate the performance of this technique in comparison with microarray techniques. This first method suffered from a low yield with few cells studied, the major progresses in this field are linked to technological advances on two levels: the development of techniques for separating and sequencing large groups of cells simultaneously and the progress of sequencing techniques allowing for a better sequencing depth[249]. Several popular protocols were developed to produce multiplexed Single-Cell RNA-Seq sequencing libraries from separated cells with various

modalities of amplification technology and transcript coverage thus providing different options depending on the type of study performed.

The main methods that were used in the first iterations of Single-Cell RNA-Seq are described here:

- STRT-Seq: Cells are separated in individual wells, lysed and an oligo-dT primer is used to begin synthesis of cDNA the addition of 3-6 untemplated cytosines at the 5' end of only the full-length synthetized strands. A helper oligonucleotide hybridizing to the cytosines overhang then promotes template switching and introduces the barcode into the newly synthetized strand of cDNA which is then amplified by single-primer PCR[250].

- CEL-Seq: Cells are separated in individual wells, lysed and a barcode is added to the RNA sequence 3' end before undergoing Reverse Transcription (RT). The cDNAs from all reactions are then pooled and PCR-amplified.[251]

- MARS-Seq: Cells are isolated in individual wells, lysed and a T7 promoter containing a barcode is annealed at the RNA 3' end to generate the first cDNA strand. An exonuclease removes the leftover RT primers and the cellular lysates are pooled and single-strand cDNA are converted to double-stranded cDNA. These cDNA are then transcribed to RNA and treated with DNase to remove leftover DNA templates. The RNA strands are fragmented, annealed to sequencing adapters and undergo RT to generated cDNA barcoded libraries[252].

- SMART-Seq: Cells are isolated in individual wells, lysed and oligo-dT primer is used to begin cDNA synthesis with a few untemplated cytosines nucleotides added at the 5' end of only the full-length synthetized strands. An oligonucleotide primer linking to the cytosines overhang is used to synthetize the second strand. The double-stranded cDNAs are then PCR-amplified and purified for sequencing[253].

- SMART-Seq2: This protocol improves on the SMART-Seq method and follows the same steps up to the second strand cDNA synthesis. Instead of a classic oligonucleotide primer, a template-switching oligo carrying 2 riboguanosines and a modified guanosine is linked to the cytosines overhang to produce a locked nucleic acid as the last base of the 3' end of the synthetized cDNA strand. The double-stranded cDNAs are then PCR-

amplified and tagmentation(cleavage and tag) is used to construct sequencing libraries[254].

Improvements to these techniques were subsequently developed such as pico-wells, *in situ* barcoding or nano-droplets which allowed for the sequencing of tens of thousands of cells in parallel[6]:

- Picowells: This technology upgrades on the microwell technology by reducing the wells volume to the picoliter scale, thus allowing for the study of tens of thousands to hundreds of thousands of cells per experiment[255].

- *in situ* barcoding: Multiple fixed and permeabilized cells are distributed in wells and a first molecular index, containing a well-specific barcode and a polyT, is introduced to the mRNAs with *in situ* reverse transcription. The cells are pooled and redistributed by fluorescence-activated cell sorting (FACS) in wells in limiting numbers (10-100) with a DAPI staining step to discriminate single cells from doublets during sorting. The second strand synthesis, tagmentation with Tn5 transposase, lysis and PCR amplification are performed while incorporating a second well-specific barcode. The resulting amplicons are then pooled and ready to sequence. This method greatly increases the number of cells sequenced in experiments using wells plates[256].

- Droplet barcoding: Microfluidic method of cell capture in hydrogel microspheres in which the lysis, mRNA capture, barcode primer hybridization and reverse transcription reactions are automatically performed upon cell capture. This method offers the capacity to capture up to hundreds of thousands of cells per run[257].

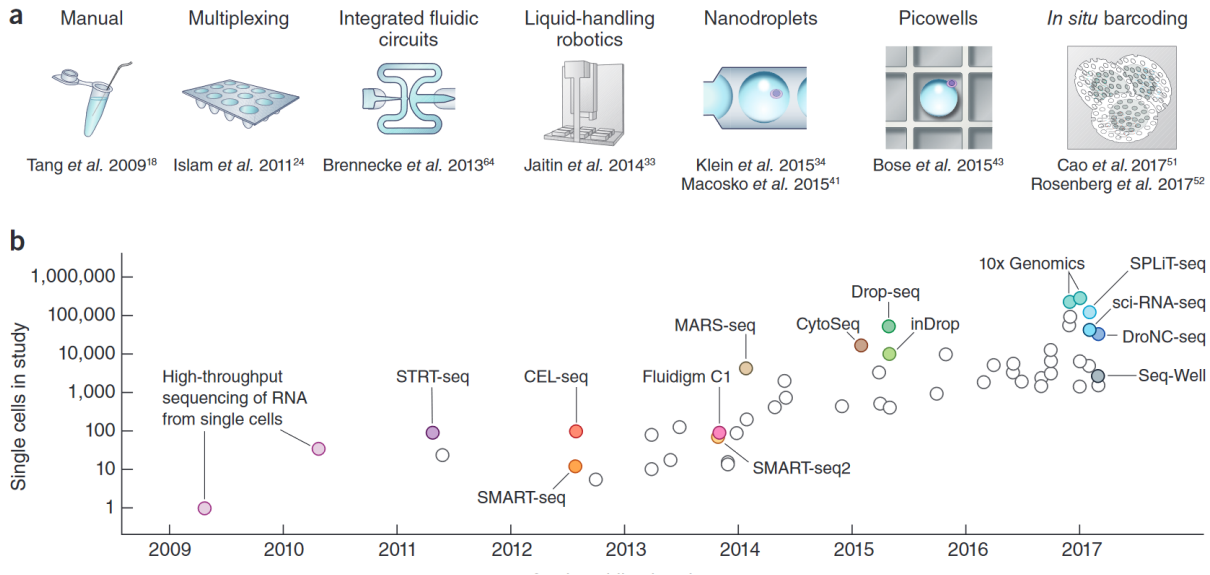
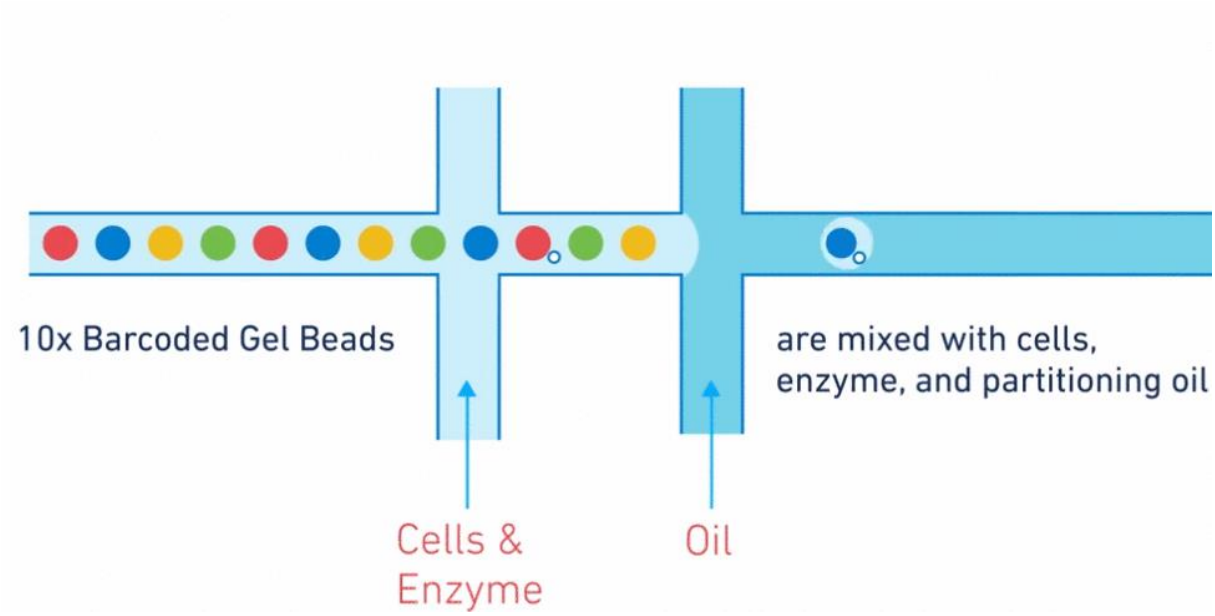


Figure 52. Scaling of scRNA-seq experiments. (a) Key technologies used in scRNA-seq. (b) Cell numbers reported in publications annotated by the type of technology used[258].

The development of bioinformatics analysis methods for this type of data has followed rapidly alongside the production of the first single-cell transcriptomics datasets[259,260] because of the need to process the produced reads and identify the cell types of the sequenced cells, as the information is lost during tissue dissociation and the inadequacy of conventional "bulk" sequencing pipelines[261]. The current Single-cell RNA-Seq library preparation protocols use small nucleotide sequences to label the molecules before sequencing, a barcode (identifies the cell of origin, 1 barcode = 1 cell) and a UMI (Unique Molecular Identifier, specific to each RNA molecule used to identify the duplicates caused by PCR amplification) are grafted onto each RNA molecule[262]. The subsequent cell type identification is performed by separating cells by gene expression to create clusters of similar expression which can then be assigned according to the tissue-specific genes expressed (“markers” genes) in each cluster[263,264].

This technology has been used on multiple biological models, including *Caenorhabditis elegans*, but there are still some differences of opinion in the scientific community on the relevance of the results obtained, in particular on the use of UMAP and t-SNE clustering techniques deemed too untrustworthy by some and on the low yield per cell offered by the current methods which create bias in the cellular type assignment[265,266,267].

(A)



(B)

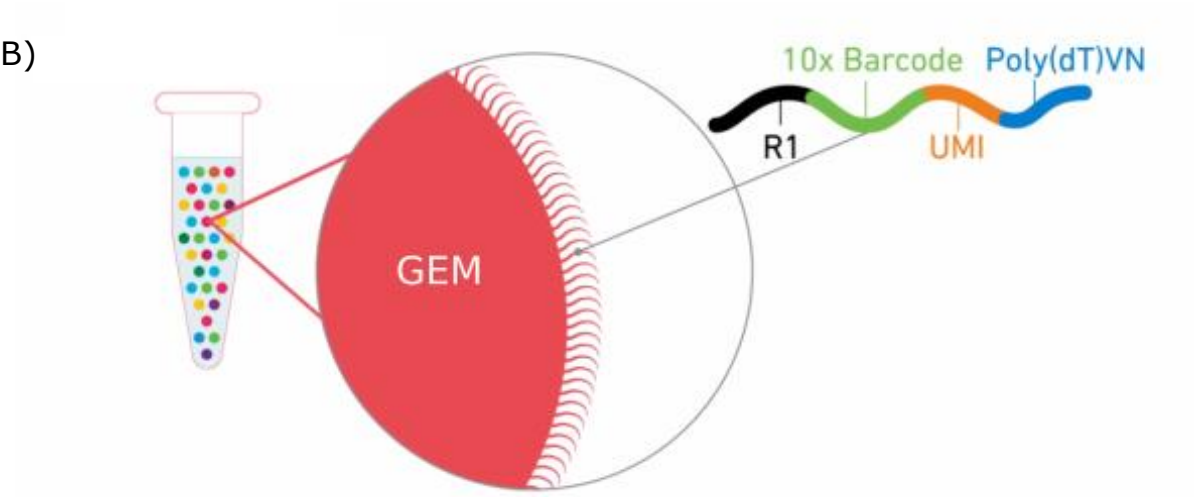


Figure 53. (A) Example of a microfluidic cell capture process using GEMs (Gel bead in EMulsion) (B) Zoom on the content of a GEM and the format used for adding barcodes and UMIs to the transcript sequences. @10xGenomics

Questions addressed in this work

Our micro-irradiation experiments on *Caenorhabditis elegans* at the L1 stage so far have been conducted on limited numbers of worms due to technical reasons (manual targeting, low amount of worms for better precision and to limit time difference between first and last worm). The small number of worms irradiated has already led us to use PCR-cDNA sequencing kits rather than direct-RNA because of the small quantities of mRNA obtained, thus preventing us from potentially detecting base modifications. This issue could become even more critical when it comes to Single-Cell sequencing because of the low number of cells available for sequencing, not accounting for eventual waste during extraction and library preparation protocols.

It therefore seems important to us to evaluate the relevance of this technique to the special circumstances of our study before committing to potentially fruitless experiments. This evaluation of Single-Cell RNA-seq for our case study is based on the analysis of a reference publication making use of this method on *Caenorhabditis elegans*: "A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution."[268].

In the context of this study, my work has been focused on:

- Downloading and processing of raw data from a *Caenorhabditis elegans* Single-Cell RNA-Seq article.

- Use of cell assignment methods commonly used in the literature (UMAP, Louvain) and comparison with the more classical method of linear regression to determine the accuracy of these methods

- Evaluation of the content of sequenced cells to determine the amount of extractable information in order to study differences between different experimental conditions

Materials and Methods

The raw fastq files were downloaded from Sequence Read Archive (SRP186643) using the SRA toolkit (https://github.com/ncbi/sra-tools/) with the commands prefetch and fastq-dump. Each condition was stored in a different SRA dataset, each dataset containing itself multiple subfiles. The reads sequences were separated in 3 files: barcode+UMI (1), transcript sequence (2) and Illumina sample index (3). The barcodes sequences were first extracted and processed to regroup reads belonging to the same barcode and the transcript sequences were mapped to the reference *C. elegans* transcriptome from Wormbase (WS275) using TopHat v2.1.1. The UMI tags were then used to eliminate duplicate reads, the choice for the remining read after filtering was based on the mapping status of the read.

Most of the data analysis was performed using custom Python 3.8 scripts with the following packages: pandas, numpy, scipy, seaborn, upsetplot, regex, re, csv, statistics, random.

The Louvain/UMAP clustering was perfomed in R with the following packages: Seurat, ggplot2, plotly, dplyr, matrix. The cells expression was organized in a matrix, normalized (log normalization) and scaled before performing a PCA (PC=100). The results of this PCA were used as input for the UMAP clustering (n_neighbors=20, metric="cosine",min.dist=0.1).

Since the files used in this analysis are very large, some of the scripts used require a lot of computing power and time to complete and were thus performed on the Curta cluster from the University of Bordeaux (https://redmine.mcia.fr/projects/cluster-curta).

The scripts used are available at:
https://github.com/pelotbdr/iribio_scripts/tree/main/single_cell_analysis

Experimental results

1. Re-processing of the article raw data

The data in this analysis was published in 2019 in Science by Packer et al[268]. The authors used Single-Cell RNA-Seq to study embryonic cell and follow their fate using their transcriptomic profile. This paper initially attracted our attention because of the major advance that a complete characterization of the transcriptomic profiles of *C.elegans* cells at different developmental stages would represent and the prospect of performing a re-analysis of the article data by focusing on cell types of interest (e.g. cells of the reproductive system for our micro-irradiation experiments).

We have therefore downloaded the raw data from this article on the Sequence Read Archive (SRA) from the NCBI. These data although considered raw were however already pre-processed with each read divided into 3 parts: 1st file with Barcode/UMI, 2nd file with mRNA sequence, 3rd file with Illumina sample index. This pre-processing is due to the authors' use of the 10X Genomics protocol associated with their Cell Ranger bioinformatics pipeline which automatically outputs the data in this format[269].

These ‘raw’ data were processed to regroup the reads by barcode, eliminate the duplicated UMIs and map the mRNA sequences on the reference transcriptome. During this step, it is common practice to eliminate cells: (i) considered to be too small, by establishing a minimum threshold of UMI count, (ii), cells with high levels of spike-in RNAs and (iii) cells suspected of being the result of a two-cell droplet. We did not perform this cleaning step for two reasons:

- The criteria used in the paper were not described and were most likely pre-defined parameters from the Cell Ranger pipeline which is neither open-source nor free meaning we couldn’t access them
- One of our major concerns with this technology is the yield due to our experimental setup of micro-irradiation. For this reason, we want to limit the loss of data as much as possible, which includes cells that may have been excluded by mistake during this cleaning step. Moreover, since the list of barcodes used by the authors is available, we are able to compare them with the excluded barcodes and judge the validity of this filtration step.

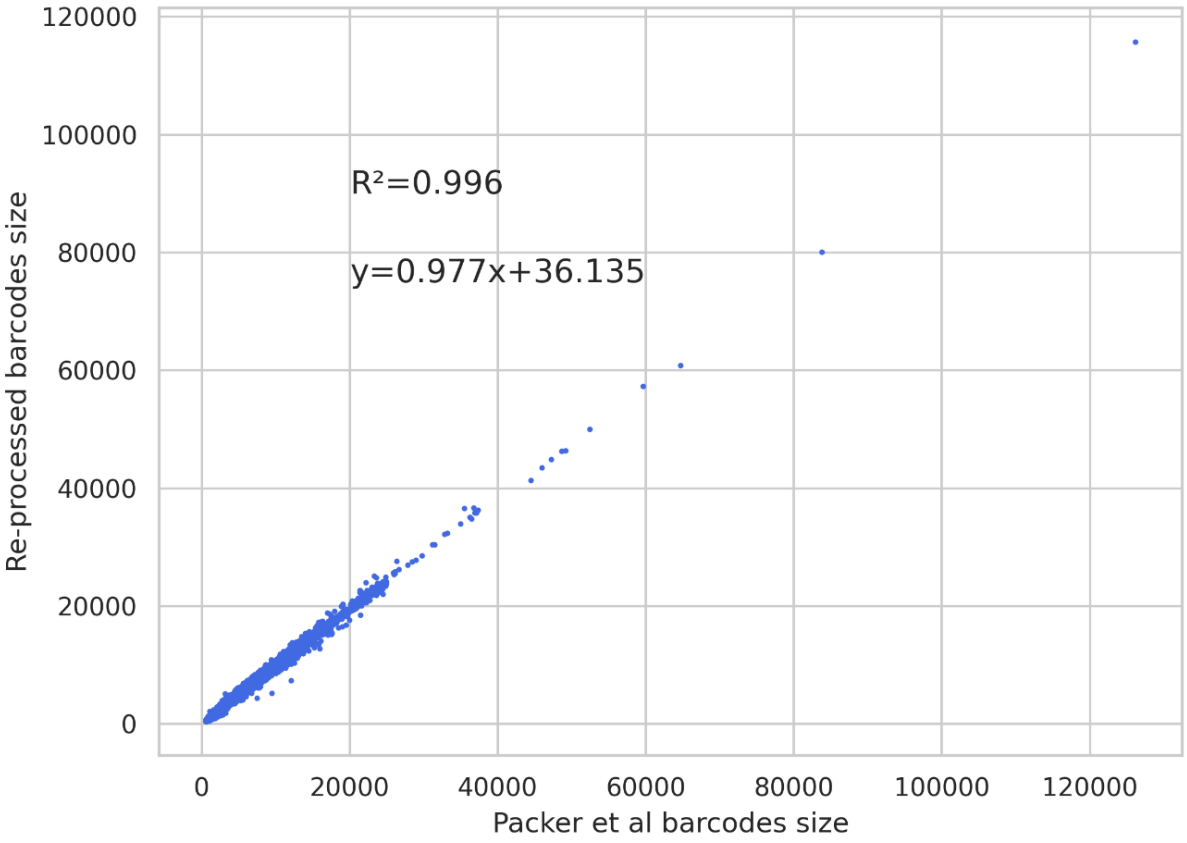


Figure 54. Correlation of barcodes size between authors data and re-processed raw data.

We checked that our re-processing of the raw files resulted in similar data by comparing the UMI count per barcode in the paper (information available in the article supplementary materials) with our UMI count per barcode (Figure 54). The results obtained seem to confirm the validity of our data processing as the cell sizes are highly correlated. A small difference can nonetheless be observed, as shown in the linear regression equation with a slope slightly different from 1 and a intercept of 36, but it can be attributed to a difference of method in the alignment of transcript sequences. The authors modified their alignment reference file by performing 3'UTR extensions of the genes to increase their alignment rate which can explain the overall minimal decrease in barcode size for our own cells. We did not perform a similar step as we remain circumspect about it both for its validity and its usefulness. To also investigate whether the transcriptomic profile of our retreated cells was consistent with those presented in the study, we separated the cells on the basis of their expression by UMAP size reduction using the same parameters as the authors.

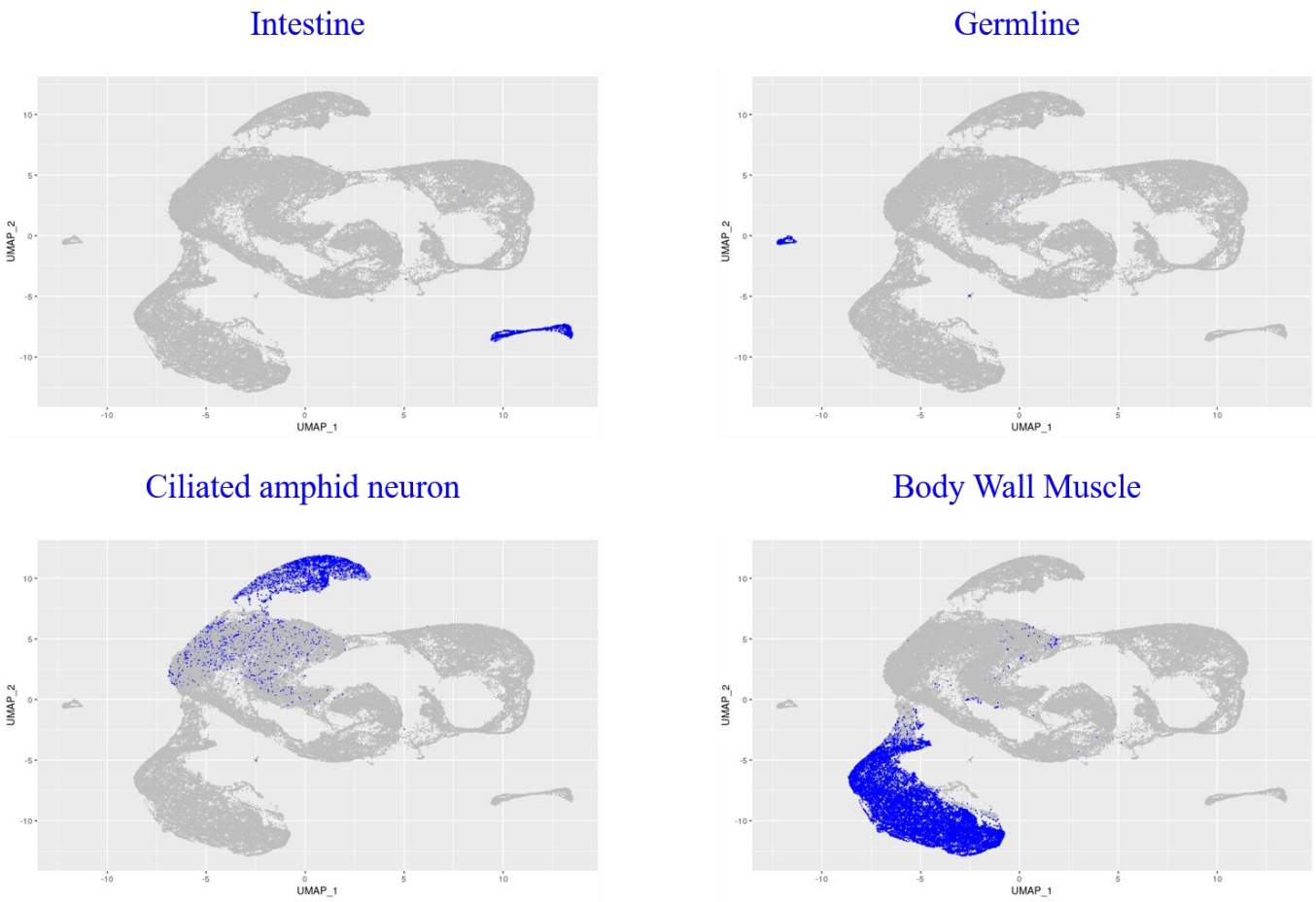


Figure 55. UMAP visualization of barcode clustering by gene expression with coloring for 4 given barcodes types based on authors assignation.

By this visualization method we can observe that the various cell types such as Germline, Intestine and Body Wall Muscle are well grouped together (according to the assignment made by the authors on the basis of their gene expression). For more specific cell types such as ciliated amphid neurons, the distinction between clusters is not as clear-cut, with significant overlap between the "regions" of these clusters (Figure 55). This closeness can be explained by the fact that the authors performed several iterations of UMAP on selections of barcodes in order to obtain more accurate assignments, especially neurons. Indeed, since the allocation of cells is performed on the set of expressed genes, limiting to cells that are anatomically, and therefore at least minimally close from a gene expression standpoint, allows for the elimination of the "weight" from other cell types on the UMAP dimension reduction and thus help to distinguish the cell types more clearly.

It should be noted that the distribution of cells obtained by UMAP, while preserving the content of the global clusters, differs from those of the authors in terms of the shape of the clusters obtained. While the UMAP parameters used were the same as those used in the article, the authors developed their own visualization tool which combined with the high sensitivity of the UMAP technique (VisoCello) and the slight differences of mapping results due to the use of their modified reference can help explain the differences even if cellular types globally remain together.

2. Analysis of UMAP clusters

Since the clusters formed by UMAP are based on complex and fuzzy parameters which are hard to interpret, we sought to assess their biological relevance. To this end, we performed barcode-to-barcode comparisons by using their expression for linear regression. We started with a presumed homogenous cell type, the Germline, which has not been divided into subgroups corresponding to more specific cell types in the authors assignments and appears very distinctly in the UMAP as separate from the rest of the data and should therefore have a very specific expression.

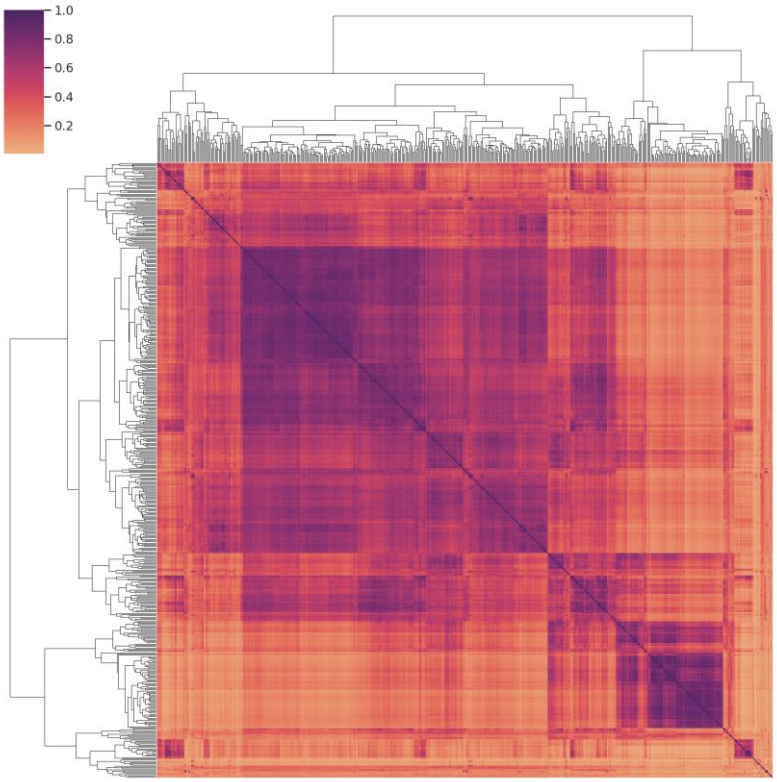


Figure 56. Clustermap of R^2 scores from linear regression on barcode-to-barcode comparison of all barcodes assigned to the Germline cluster.

The R^2 scores obtained on these comparisons allow us to observe globally strong proximities of expression within the cluster but which divide slightly into subgroups (Figure 56). We distinguish a majority group of highly correlated barcodes and a second smaller group with a strong internal correlation. At the very bottom of the cluster map, there are also several barcodes that do not seem to share a particularly high correlation with any other barcode in the cluster which could be erroneous assignments in view of their low scores. From this overall proximity between the barcodes in the cluster, we wanted to compare their average expression with barcodes from other cellular assignments to determine whether the distinction in expression between these two categories is clearly distinguishable. For this purpose, we produced a "metacell" of the Germline cluster which consists of the aggregation of the 150 barcodes of the cluster in order to produce an average expression. We then compared by linear regression the 157 remaining barcodes of the cluster to the "metacell" as well as 5,000 barcodes randomly selected from the rest of the data.

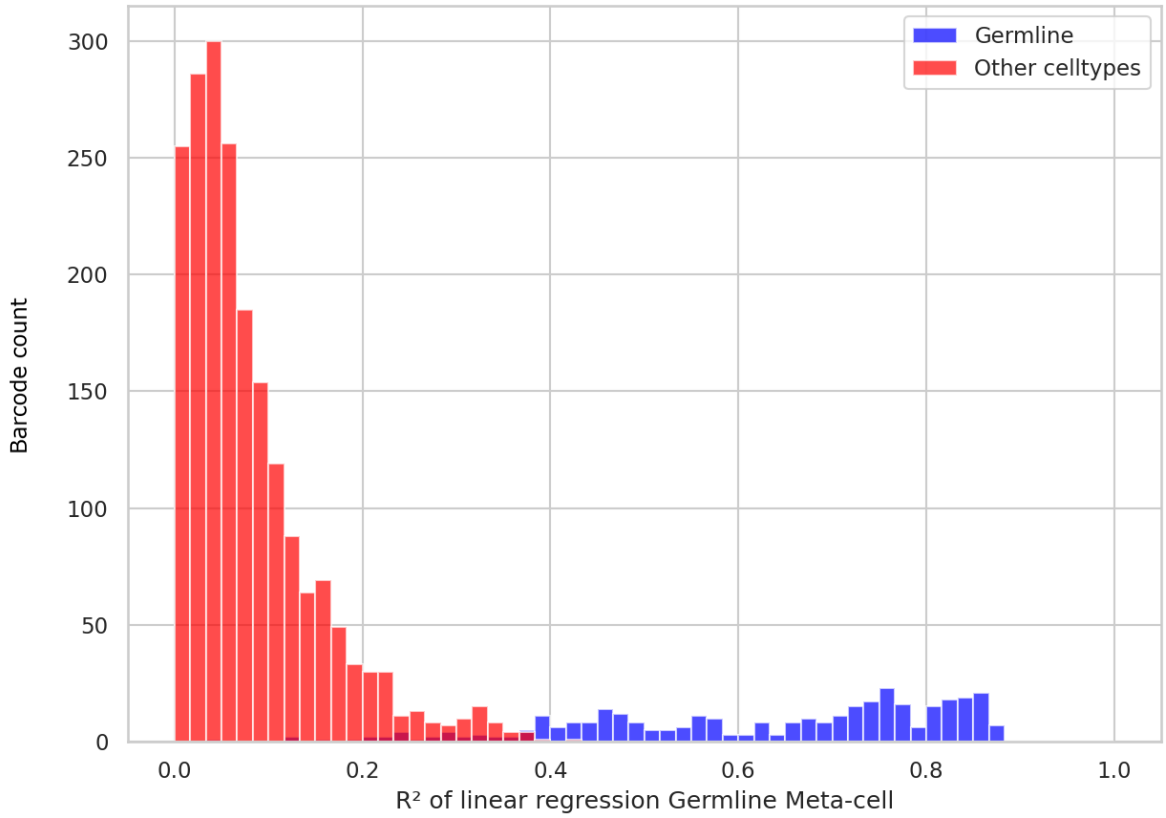


Figure 57. Distribution of R^2 scores for linear regression between Germline “metacell” and other Germline-assigned barcodes or other celltype-assigned barcodes.

We observe that most Germline-assigned barcodes have overall strong R^2 scores over 0.6 with the meta-cell, with a few barcodes getting scores around 0.5 and some rare exceptions falling

below 0.4 which could correspond to the bottom of the clustermap observed previously. The majority of barcodes assigned to other cell types have however a score ranging from 0 to 0.2 thus confirming the difference in expression, even if slight overlap is observable with some of these barcodes obtaining scores up to 0.4 (Figure 57). In summary, the global expression linked to this cell type thus seems to be quite distinct from the rest of the dataset, either by the UMAP method or by linear regression. A small proportion of the barcodes assigned in this cluster by the UMAP method, however, appear to share little or no expression when compared by linear regression.

We then performed a similar analysis but this time on the barcodes assigned to the Intestine, which were classified by the authors into three subcategories: anterior intestine, middle/posterior intestine and undefined intestine (barcodes that could not be assigned to one of the two previous categories but were nonetheless considered as sufficiently similar to be tagged as Intestine). In addition to these subgroups, this cell type shares with the Germline cell type the advantage of being clearly distinct from other clusters on the UMAP representation, although to a lesser degree as the cluster is more dispersed, which should therefore represent significant proximity in transcriptomic expression between the barcodes.

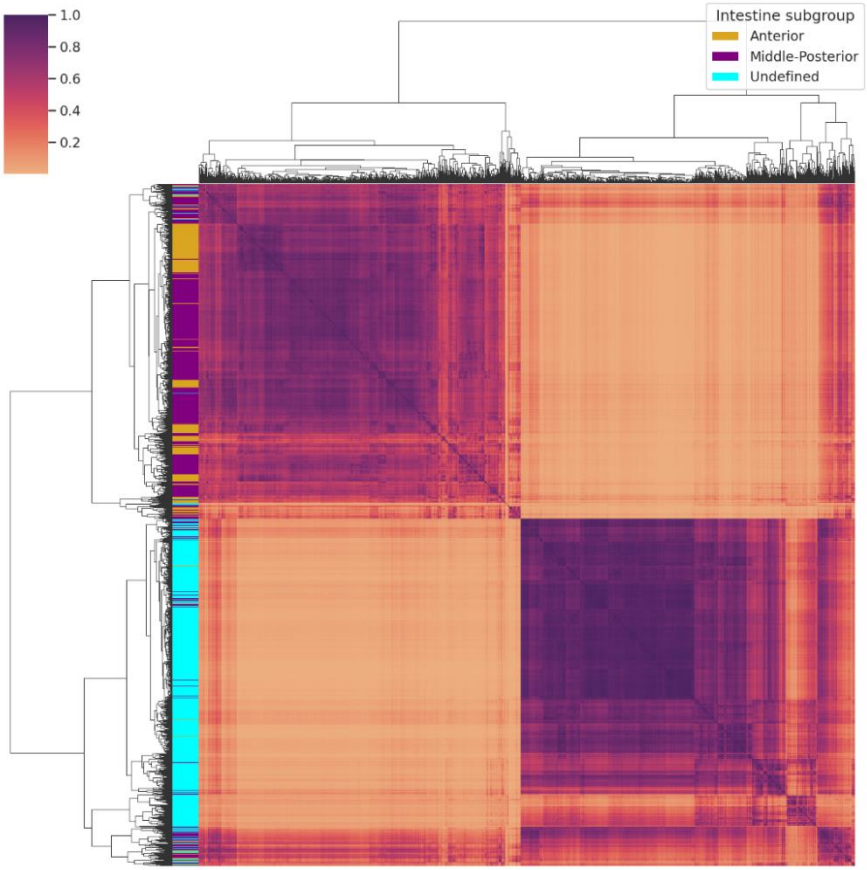


Figure 58. Clustermap of R^2 scores from linear regression on barcode-to-barcode comparison of all barcodes assigned to the Intestine cluster and row colored by Intestinal cellular subtype.

In contrast to the Germline cluster, an important schism is observable here with barcodes that were assigned to an Intestine cell subtype on one side and undefined Intestine barcodes on the other side (Figure 58). The difference between these two categories is very clear with R^2 scores mostly below 0.2 when comparing two barcodes from different categories, *i.e.* similar scores to those obtained when comparing the Germline "metacell" and barcodes from other cell assignments.

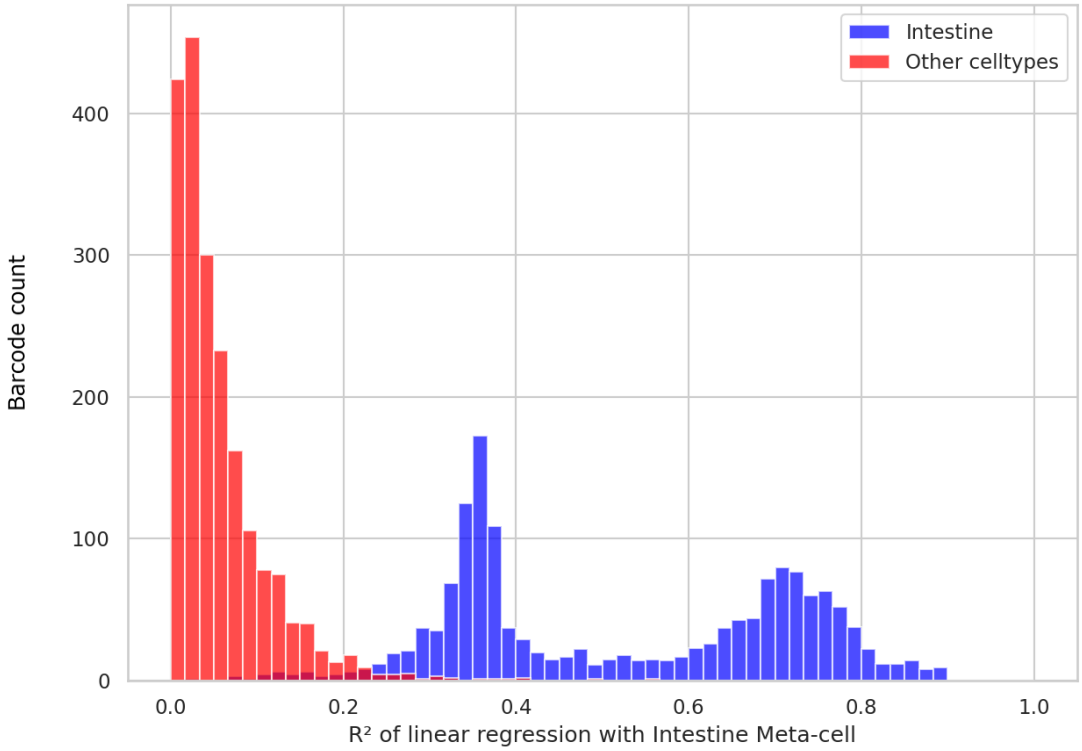


Figure 59. Distribution of R^2 scores for linear regression between an Intestine meta-cell and other Intestine-assigned barcodes or other celltype-assigned barcodes.

Comparison of the average expression of the cluster using a "metacell" with 5000 barcodes of other cell types still results in a clear demarcation with the barcodes of the cluster, but this result is biased by the fact that it is an average expression of the 2 distinct groups found in the cluster and not of these groups individually (Figure 59). We therefore made two "metacells" from the expression of barcodes of both Intestine cell subtypes identified in the cluster in order to compare them to the barcodes of undefined Intestine and the barcodes from the other subtype.

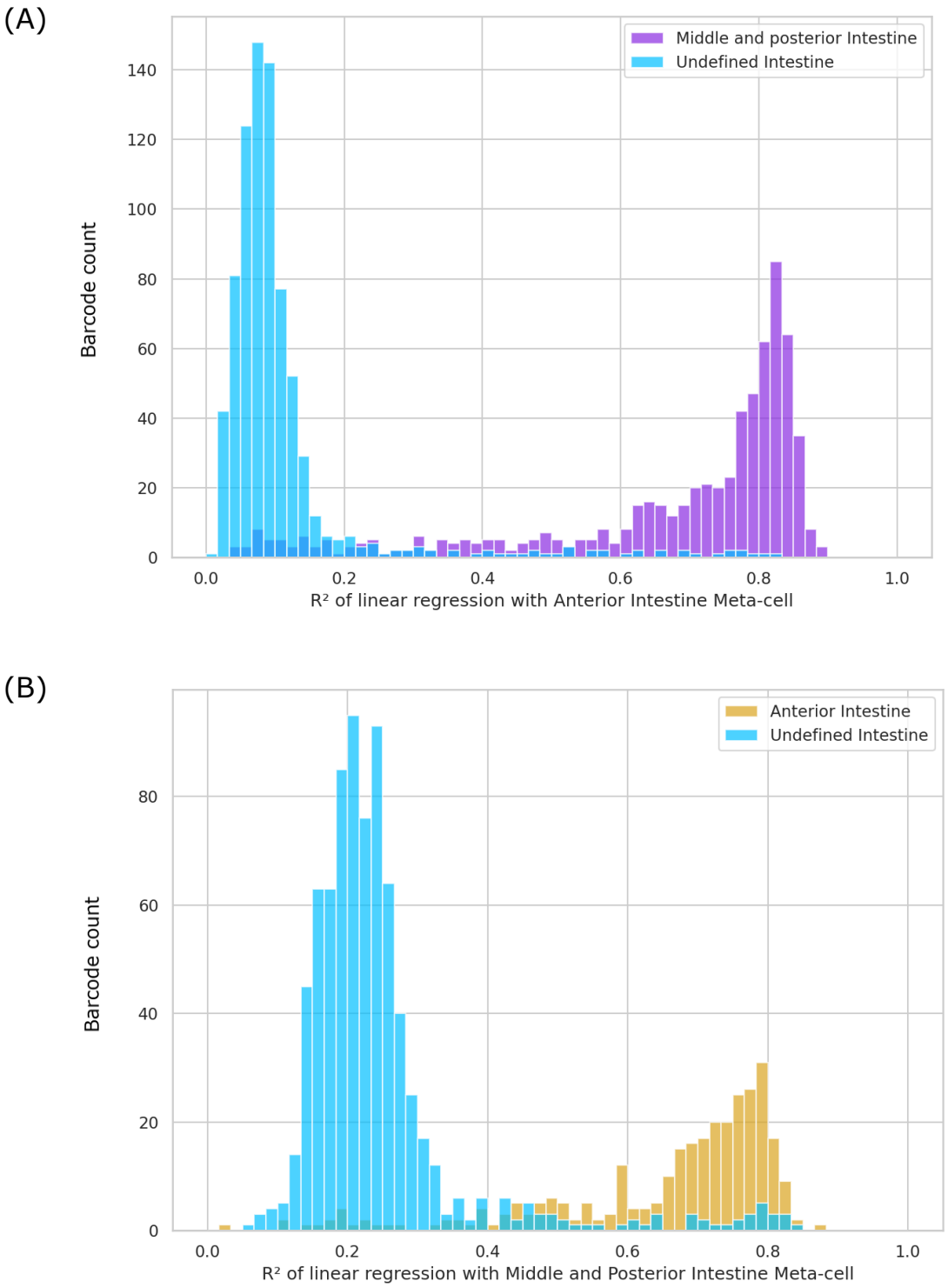


Figure 60. Distribution of R^2 scores for linear regression between a (A) anterior Intestine "metacell" or (B) middle/posterior Intestine "metacell" and undefined intestine-assigned barcodes or anterior intestine-assigned barcodes.

We find important scores, around 0.7-0.8, with these "metacells" when comparing the two defined celltypes of the Intestine cluster while the undefined Intestine barcodes showcase scores

around 0.1 or 0.2 depending on the “metacell” although some rare undefined barcodes get scores up to 0.8 (Figure 60). By these results, the difference in expression between the anterior and middle-posterior Intestine and the undefined Intestine can be clearly distinguished, particularly with the anterior Intestine "metacell" for which the R^2 scores are almost comparable to those obtained by comparing the average expression of the cluster and barcodes of other cell types. These scores are still slightly above those of the barcodes from other cell types which indicate at the very least a small degree of proximity but shouldn’t be sufficient for them being grouped together in the same cluster.

Within the Intestine-assigned barcodes, there is thus a significant schism between the barcodes whose expression allowed them to be categorized into cellular subtypes and those for which this was not possible. The question then arises as to why these two groups of cells were grouped together by the UMAP method in the first place despite showing such low correlation between each other.

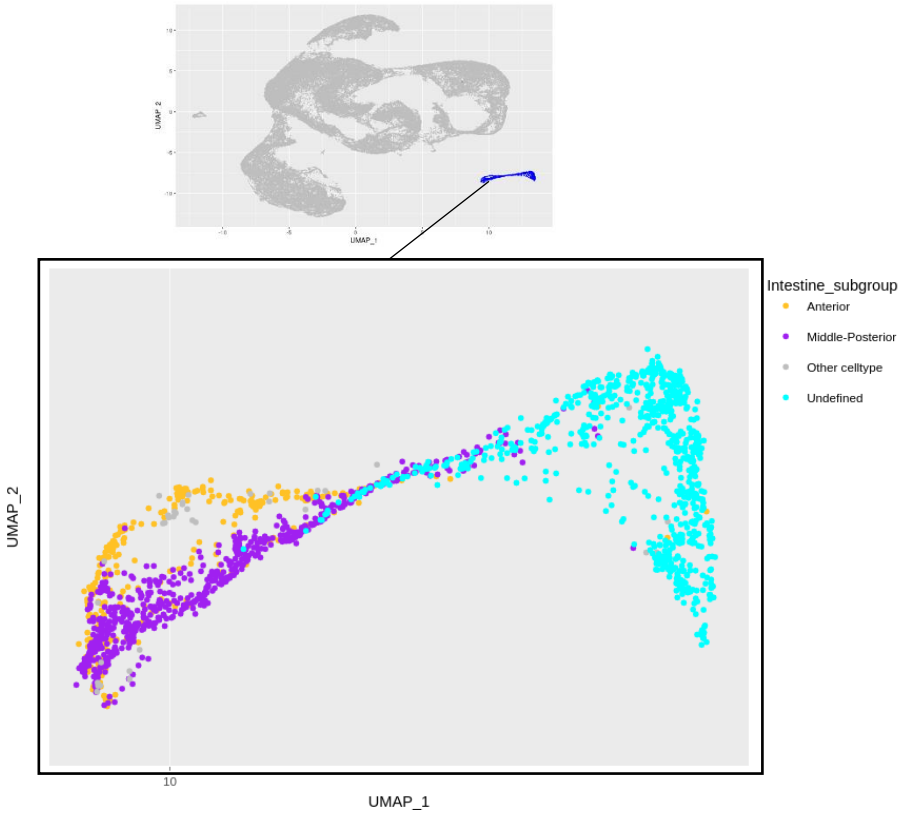


Figure 61. Focus on the UMAP cluster of Intestine-assigned barcodes colored according to their assigned subtype.

We observe on the UMAP that this distinction between the two groups is well preserved with on one side the anterior and middle-posterior barcodes and on the other side the undefined

barcodes (Figure 61). A link is present between these two clusters, consisting of a few barcodes from both sides, which is also visible at the bottom of the clustermap and could explain the few barcodes with important R^2 scores with the anterior and middle-posterior Intestine “metacell”. The overall distance between these two clusters being very low, it is possible that this link is at the origin of their grouping. However, in this case, it would mean that the undefined barcodes were considered as belonging to the Intestine based on the expression of only a fraction of the whole cluster.

Another hypothesis that could explain the grouping of these two clusters lies in the expression of marker genes. Indeed, the authors of the study have established a list of genes whose global expression within a cluster makes it possible to link it to a cell type or a cell subtype. As these genes should be specific to their cellular origin, they could theoretically justify the grouping of two clusters sharing a low proximity if these two clusters were the only ones to express one or several of those specific genes.

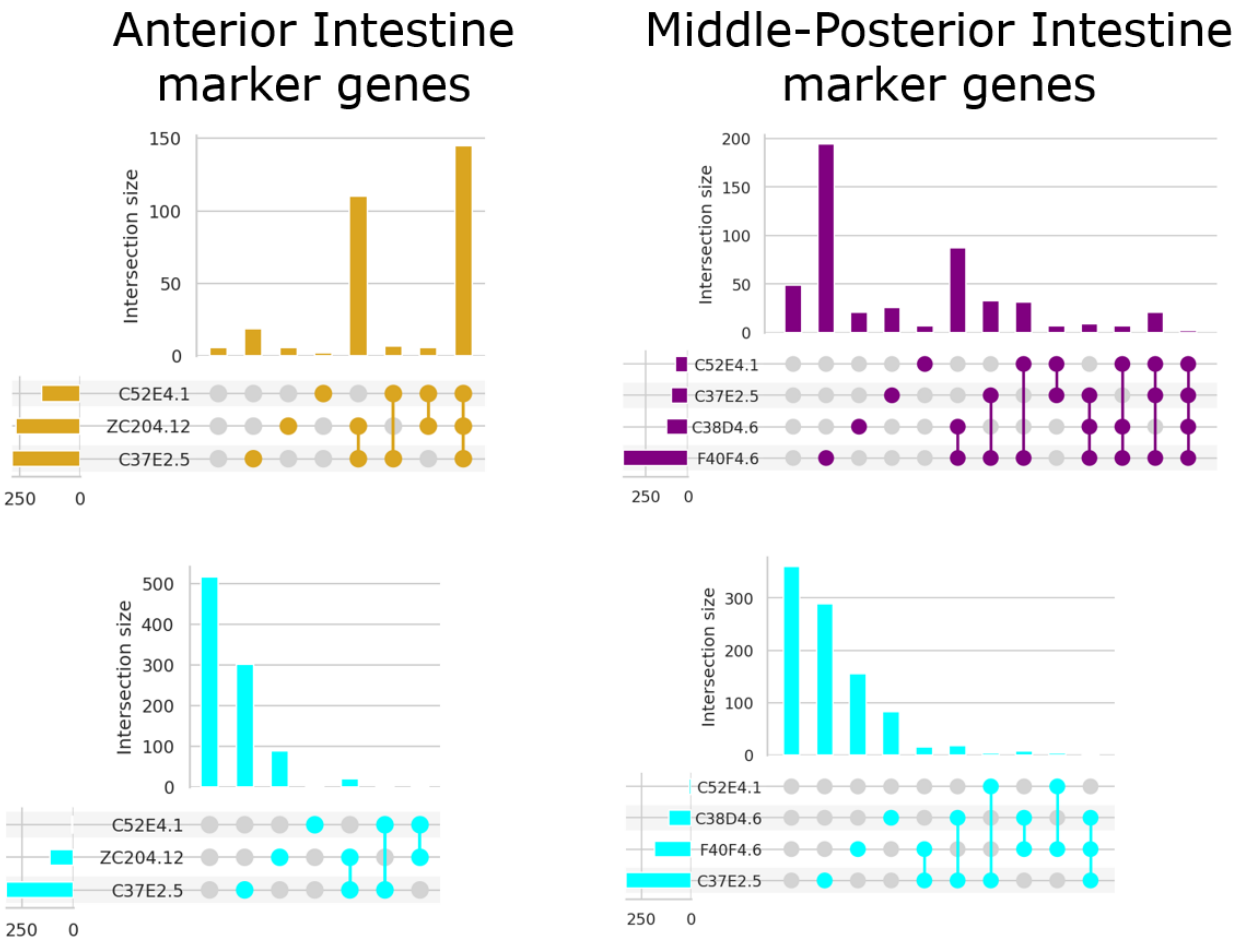


Figure 62. UpsetPlot of Intestine markers genes found at least once in barcodes of the defined and undefined Intestine clusters. Yellow: Anterior Intestine, Purple: Middle-Posterior Intestine, Blue: Undefined Intestine.

From the measurement of marker gene expression for the anterior and middle-posterior Intestine assignments, we can notice that 2 marker genes are common to both assignments, C37E2.5 and C52E4.1, and that two different expression profiles emerge on the two clusters in question (Figure 62). On the anterior Intestine barcodes, the majority of the barcodes express two or three of the three total marker genes, whereas on the middle-posterior Intestine barcodes only one or two of the four marker genes are expressed, F40F4.6 being the most frequent marker gene in this assignment. Overall, we find very few barcodes in these two assignments that do not express any of the marker genes. For the undefined Intestine barcodes, we find a fairly similar marker gene expression profile for the two marker sets, with about half of the barcodes containing no marker genes and the other half expressing one, mostly C37E2.5 or F40F4.6, and very few barcodes expressing multiple markers genes.

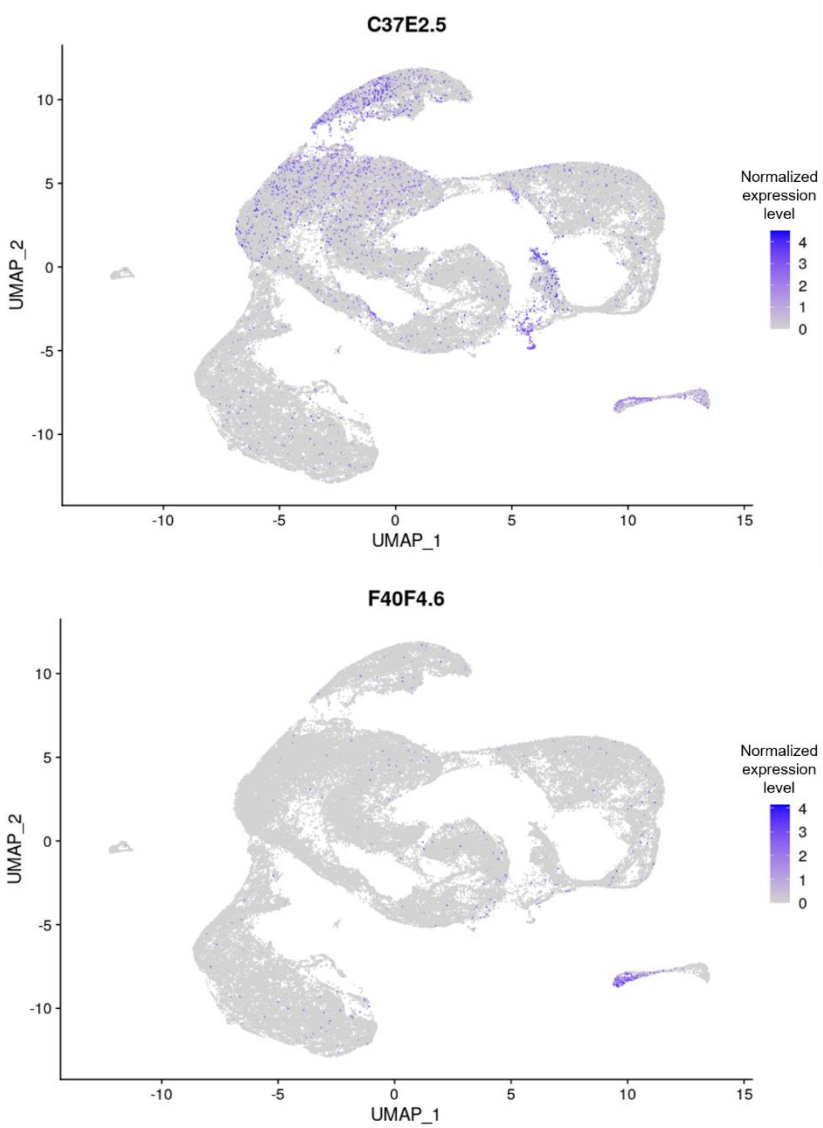


Figure 63. Normalized expression level of markers genes C37E3.5 and F40F4.6 highlighted on the UMAP visualization.

While the expression of C37E2.5 shows low specificity to the Intestine cluster, F40F4.6 seems however to be very specific to the Intestine barcodes with only a few barcodes from other clusters expressing it. However, the expression of F40F4.6 appears more frequent in middle-posterior Intestine barcodes than in undefined Intestine barcodes (Figure 63). The expression of this gene in barcodes of the undefined intestine could therefore be an important factor for their clustering.

In summary, the most likely reason behind the clustering of defined and undefined Intestine barcodes revolves around the proximity of expression between some of the barcodes from both categories and the common expression of genes, including potentially genes other than marker genes, which are rarely found in the rest of the dataset.We then investigated whether a similar phenomenon of the presence of distinct groups within the same UMAP cluster was also found on other cell types. We looked in the barcodes assigned to the Body Wall Muscle and to the Ciliated amphid neuron as both of these cellular types represent distinct regions of UMAP clustering and contain a significant number of barcodes (17,000 and 6,000, respectively) and thus are likely to exhibit more complex internal behaviors in terms of expression proximity.

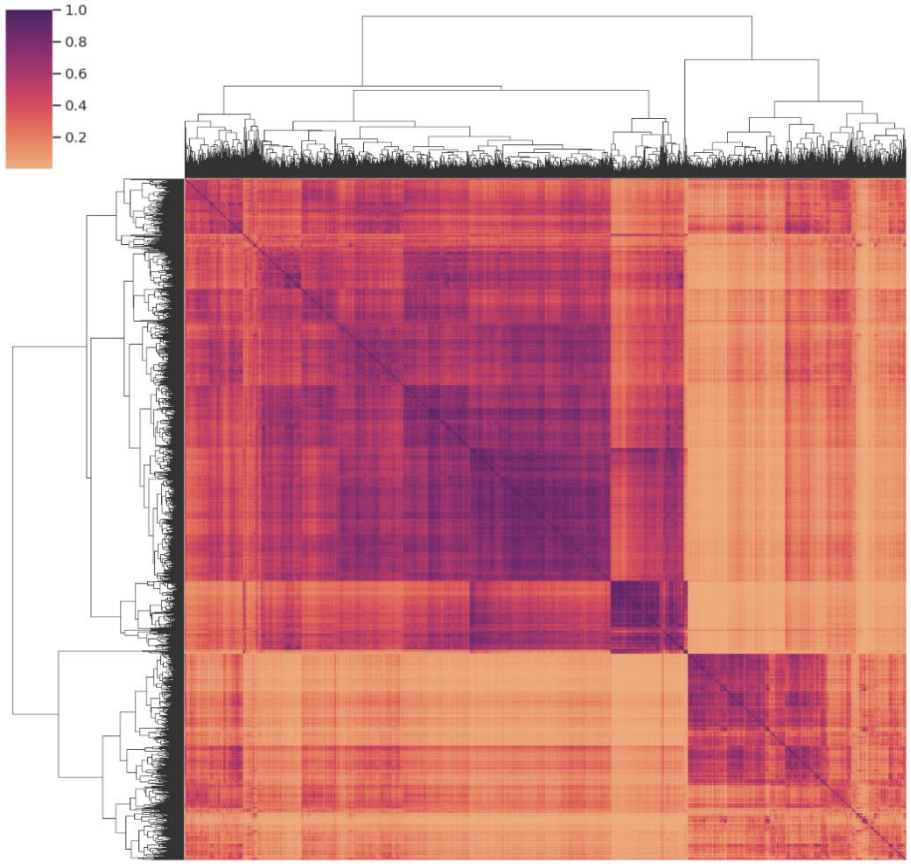


Figure 64. Clustermap of R^2 scores between barcodes in Body Wall Muscle (5 000 barcodes randomly sampled out of 17 000).

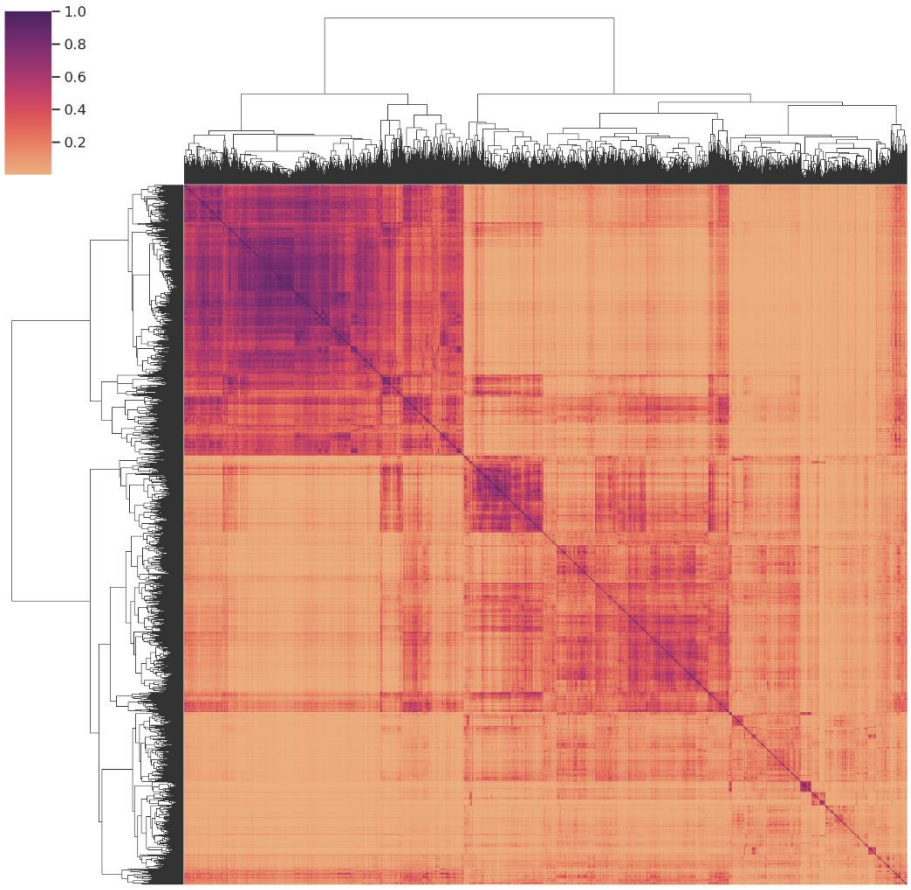


Figure 65. Clustermap of R^2 scores between barcodes in Ciliated amphid neuron (all 6 000 barcodes of this category).

For the Body Wall Muscle, we also observed two distinct groups showing a high internal score and a low score between the groups. Subgroups within these two groups can also be identified but they keep a consequent degree of proximity to their original group (Figure 64).

For the Ciliated amphid neuron, we have a slightly different case with a high internal score group representing only a fraction of the total barcodes while the rest of the barcodes show almost no proximity to the rest, some small groups of low proximity being nevertheless identifiable (Figure 65). The absence of a large group in this cell type is not completely surprising, as this cell type encompasses many different neurons. The question remains, however, why these barcodes were grouped together in the same cluster despite the near total lack of similarity for many of them. Even if successive rounds of clustering can better distinguish cluster formation via UMAP, this does not explain why barcodes with such differences end up together.

The clusters formed by the UMAP method thus seem to group together barcodes sharing an important proximity of transcriptomic expression, but also include barcodes based on a weak proximity and therefore whose attachment to a cell type seems more uncertain. This seems to translate either by the presence of two main groups of different expression within the same cluster or by a single main group and "satellite" barcodes attached on the basis of weak proximity. However, such a linkage would mean that not all barcodes have the same relevance from a biological point of view when it comes to extracting an average gene expression from the cluster. The results of single cell clustering should therefore be considered with caution due to the high heterogeneity of barcode expression.

3. Large barcodes content analysis

Since the goal of our micro-irradiation studies is to be able to analyze a cellular response at the single cell level, it is necessary that the content of the sequenced barcodes is sufficient both to determine the original cell type and to distinguish a difference in expression between an irradiated and a non-irradiated condition. However, in this paper, barcodes containing as few as 500 UMIs, in an organism with nearly 20,000 genes, are kept despite the fact that they can necessarily only represent a fraction of the expression of the cell of origin, even if this expression is distinct enough to place them in UMAP clusters.

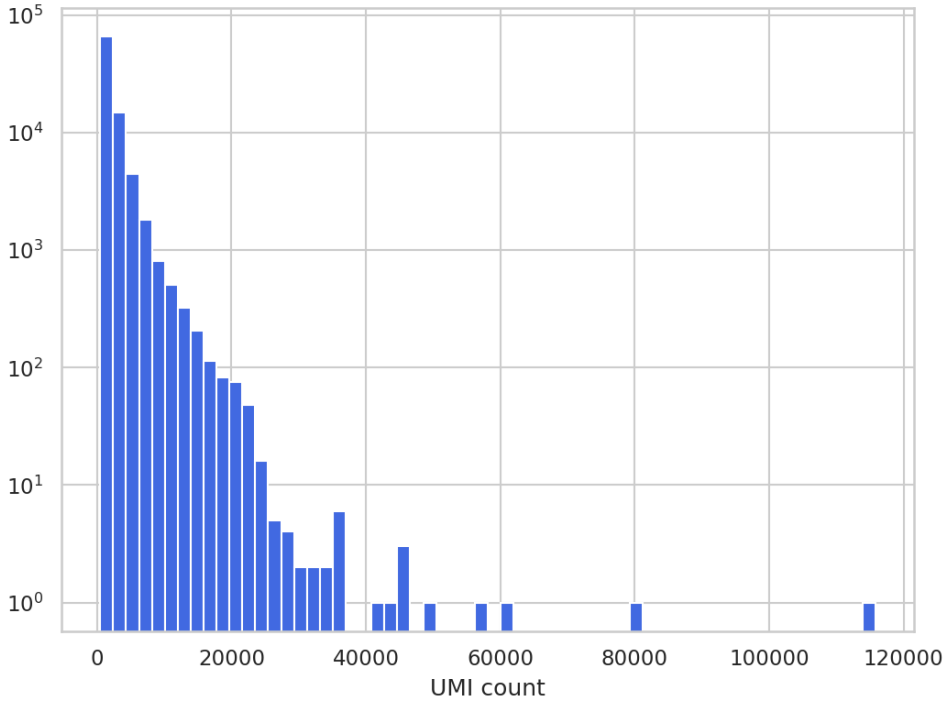


Figure 66. Size distribution of barcodes used in the article.

The distribution of all barcodes in the dataset allows us to observe that almost all barcodes contain less than 3000 UMIs and that the most frequent size is about 1000 UMIs (Figure 66). However, a small part of the barcodes stands out from the rest of the dataset because of their size. We identified 1 443 barcodes containing more than 10 000 UMIs out of the 89 701 total barcodes included in the study, which we've dubbed High Content Cells (HCC). Their profile is interesting because of their larger size, the "weight" behind their expression is more important and they offer the potential of studying a more snapshot of the expression of their cell of origin than the smaller barcodes. However, these High Content Cells are rare and we cannot expect to obtain them with certainty if we were to produce our own single-cell sequencing runs. We therefore also looked at the content of small barcodes to determine whether, since they represent only a fraction of cellular expression, grouping several of them into a single "metacell" would result in expression similar to that of a large barcode. Through this comparison of large and small barcodes, we sought to determine whether the efficient analysis of a cell type necessarily required the use of large barcodes or whether the combined expression of the large number of small barcodes could compensate for their low content.

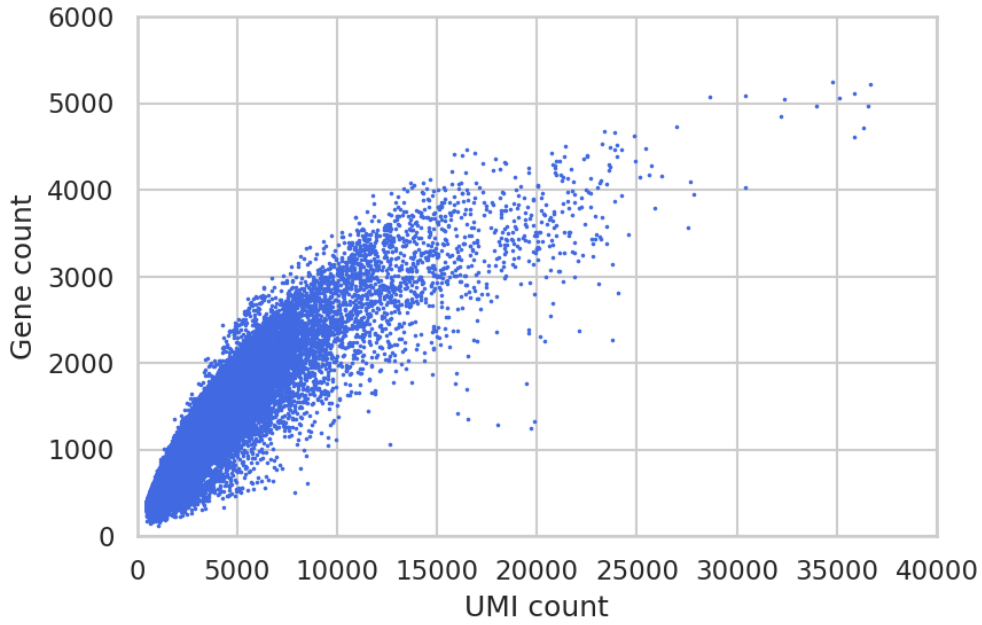


Figure 67. UMI count and Gene count per unique barcode used in the article with a cutoff at 40 000 UMIs for better visualization.

The distribution of the number of genes and UMIs per cell unsurprisingly shows that the increase in the number of UMIs in a barcode is related to an increase in the number of unique genes with the majority of barcodes containing between ~200 and ~2000 unique genes.

However, this increase appears to reach a plateau at ~5000 unique genes in a single barcode (Figure 67).

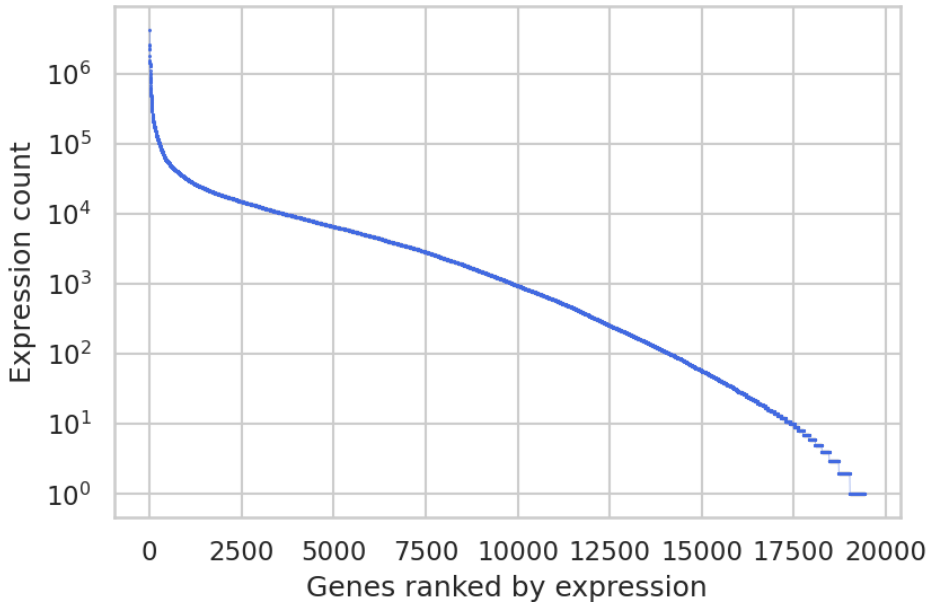


Figure 68. Expression levels of genes ranked in decreasing order

Looking at the total expression of the dataset, however, we see that about 1000 genes represent the vast majority of the expression (Figure 68). We therefore compared the proportion of barcode expression being composed by these 1000 genes, considering that genes specific to particular cell types and therefore indicative of cell expression are generally low expressed genes.

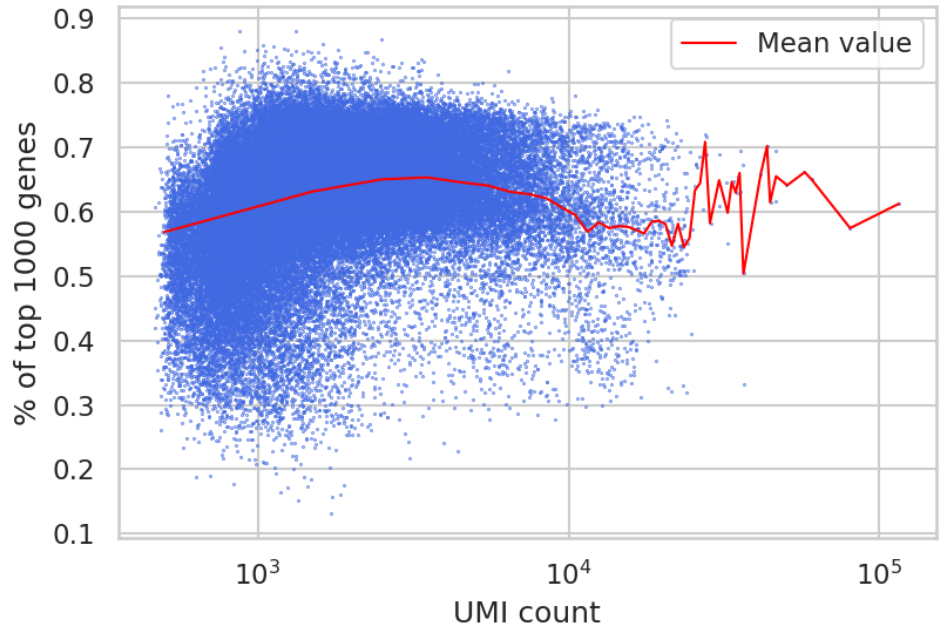


Figure 69. Percentage of expression comprised of the top 1000 expressed genes in barcodes categorized by size.

We observe that the proportion of the expression of barcodes composed by the top 1000 genes can vary greatly. particularly for small barcodes (Figure 69). The mean value of this proportions is however relatively constant across the different sizes of barcodes which allows us to establish two points:

-The small barcodes not being composed only of the most common expressed genes means that the rest of their expression may contain genes of lesser frequency. These genes are expected to be more informative of cell type specificity theoretically making them easier to cluster. The expression of these barcodes could therefore be highly specific of a given cell type and be a central element driving the global expression of a cluster.

-The similar profile between HCC and the rest of the dataset (~60-65% most expressed genes) suggests that despite the significant size differences found in the dataset, the sequenced material remains essentially the same and only the sequencing depth varies. HCCs or small barcodes would therefore not be the result of artifacts of the library preparation protocol or sequencing.

We then studied the behavior of these HCCs within the different cell-type clusters to identify whether these barcodes are particularly distinct from the rest of the dataset due to their far broader content.

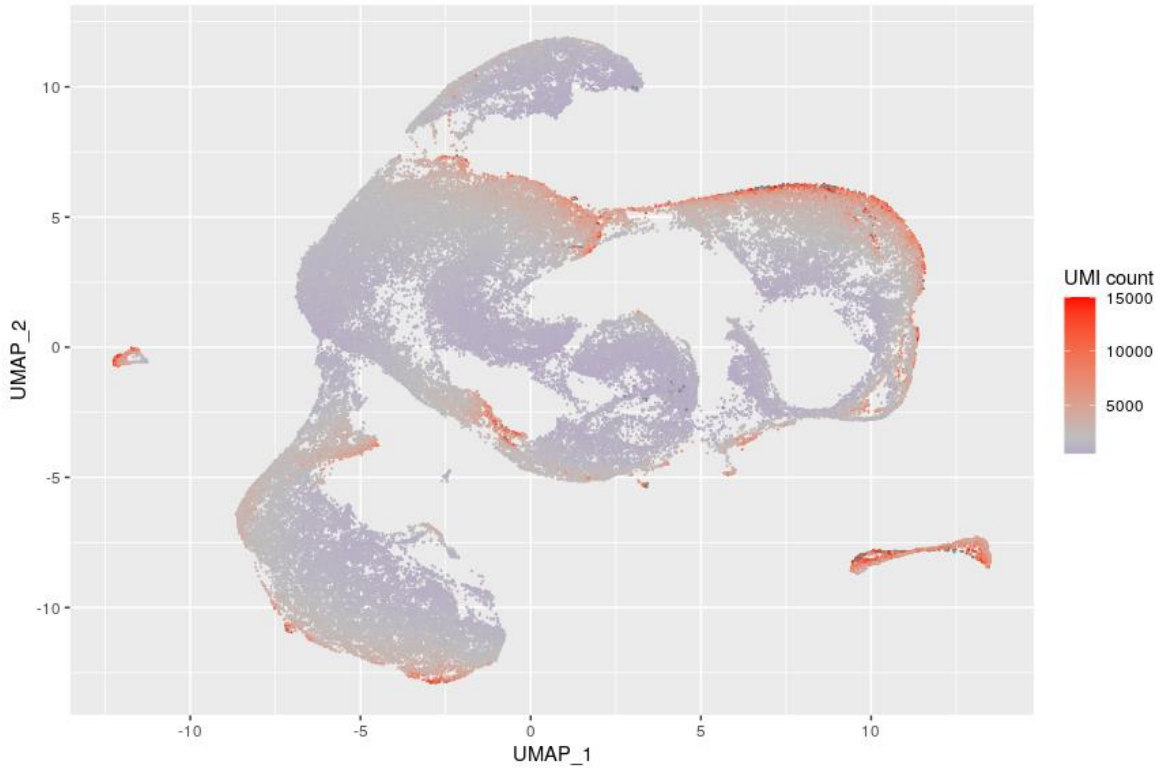


Figure 70. UMAP visualization of cell clustering with color based on barcode size. Maximum color threshold set at 15 000 for better visibility.

We observed on the UMAP that on the whole, the HCCs seem to be grouped on different areas but in general on the periphery of the clusters (Figure 70). Their distribution does not seem to be equitable between the different clusters, we can distinguish the Germline and Intestine clusters which are composed of a large proportion of HCCs (38.9% and 27.1% respectively) while clusters like the Body Wall Muscle contain only a small number (~0.3%).Their low quantity per cluster as well as their "external" position on the cluster layout seems to indicate that the HCCs are not the main components defining the expression of most clusters around which would be grafted the smaller barcodes with a less defined expression.

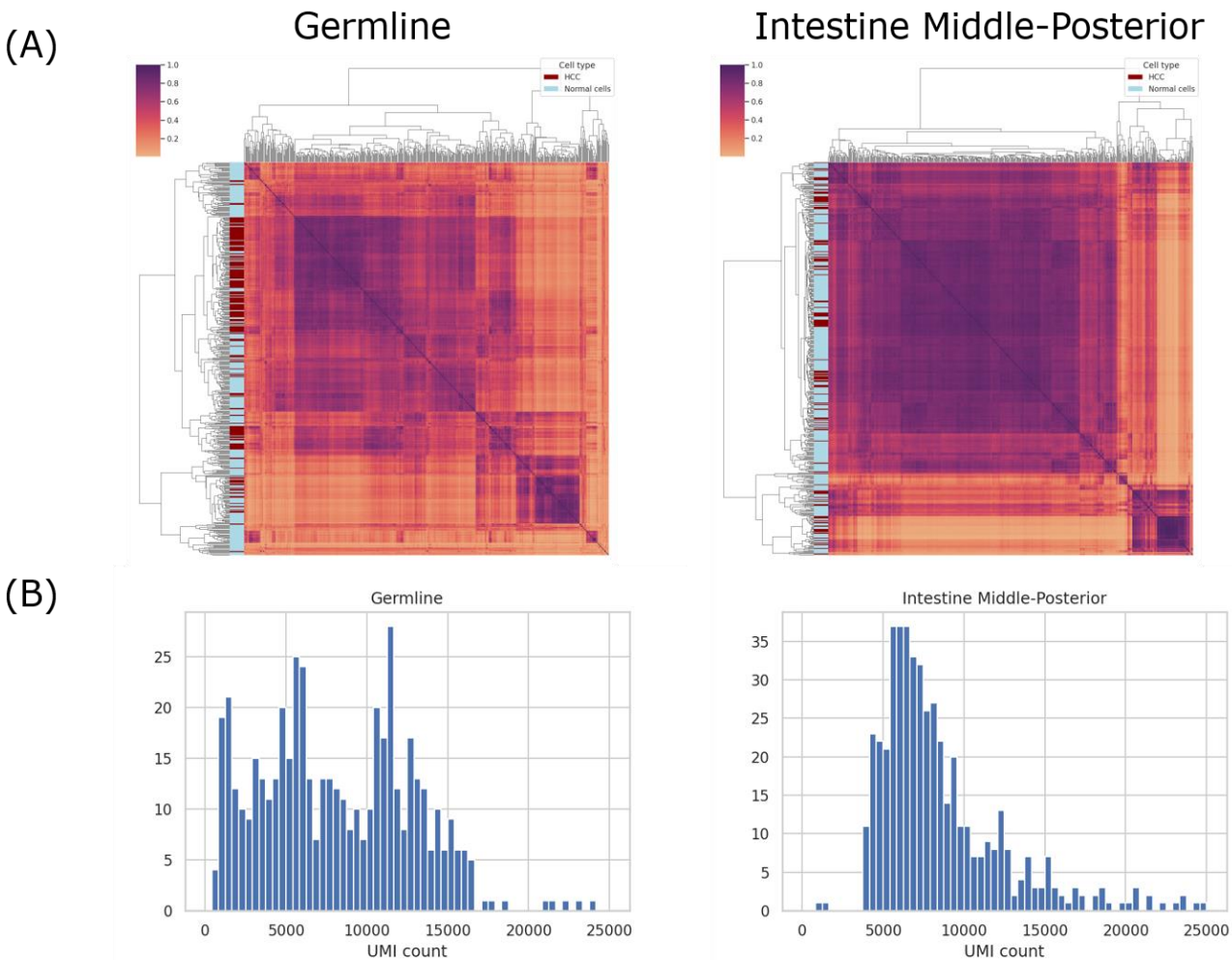


Figure 71. (A) Clustermap of R^2 scores, row colored by size category (Red: HCC, Blue: Normal cell), and (B) Distribution of barcode sizes in Germline and Intestine Middle-Posterior clusters.

In clusters containing a significant proportion of HCC (Intestine and Germline), the behavior of these barcodes differs between the two cell types (Figure 71). In Germline, we can observe that HCCs form a large cluster indicating a strong proximity of expression by linear regression which seems to constitute the "core" of this cell type expression, with smaller barcodes

displaying a weaker correlation on the outskirts of this cluster of HCCs. In the middle-posterior Intestine, on the other hand, if some HCCs do group together, they do not form a distinct cluster. The expression of the cell type which is here rather homogeneous between the barcodes of all sizes indicates a very distinct transcriptomic expression. The reason behind this difference could potentially lie in the average barcode size of these two cell types. Indeed, in the Intestine, most barcodes have a relatively constant size around 6,000 UMIs and therefore have a relatively well-defined expression, which probably explains the high R^2 scores observed and the absence of clustering of HCCs. In the Germline on the other hand the size distribution is more heterogeneous with many barcodes containing less than 5000 UMIs and the number of HCCs being much higher. The HCCs being the barcodes with the most defined expression in this population, they constitute the "core" expression of the cluster. Thus, even in clusters containing a high proportion of HCCs, the latter are not necessarily the main factors determining the "core" expression of the cluster, as smaller cells may also fulfill this role depending on the cell type.

However, the barcode size distribution observed in these two cell types is not the norm in the rest of the dataset, with the rest of the cell types generally having only a few HCCs within them. In these situations, the HCCs are therefore necessarily no longer the "core" of the cluster expression due to their too low number and the average expression is therefore determined from the expression of small barcodes.

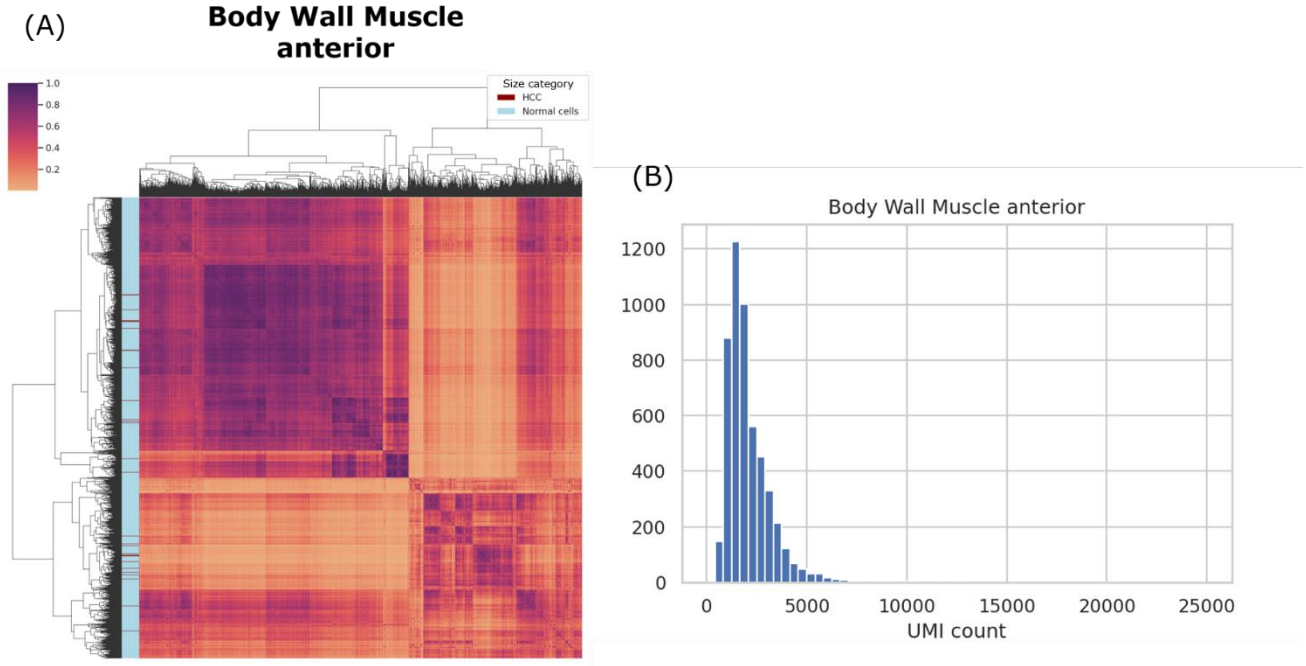
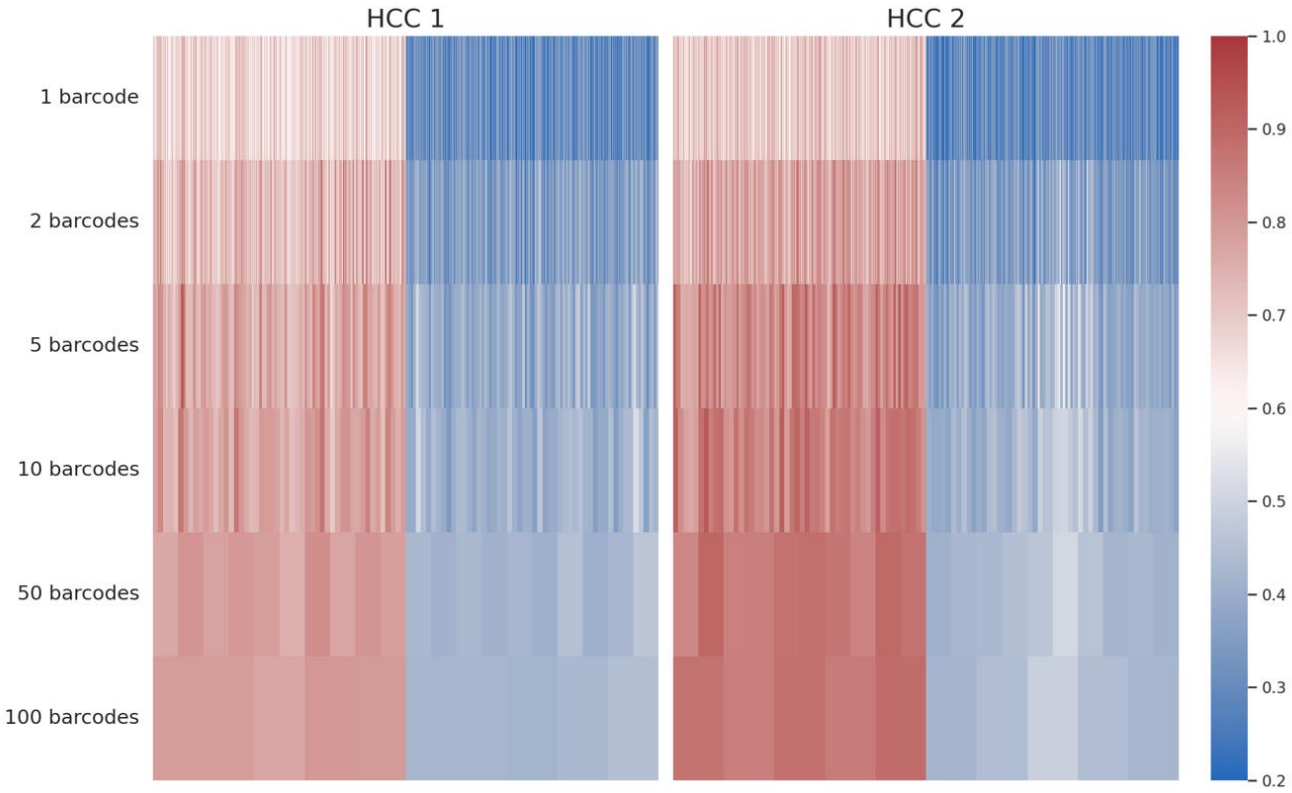


Figure 72. (A) Clustermap of R^2 scores, row colored by barcode category (Red: HCC), 2 000 barcodes sampled out of 5 191. (B) Distribution of barcode sizes in Body Wall Muscle anterior cluster.

In the Body Wall Muscle anterior cluster, we find 17 HCCs that are distributed without any apparent structure on the clustermap confirming that their expression does not define the "core" expression of their cluster (Figure 72). Since these HCCs are much larger than the rest of the barcodes in the cluster, we sought to determine whether the grouping of several small barcodes resulted in a "metacell" whose expression would be close to that of an HCC and thus more highly correlated than the expression of the individual barcodes composing it. Through this comparison, we sought to determine whether HCCs are effective representatives of their cluster and effectively being the closest thing to the complete expression of the cell of origin. We grouped two different types of barcodes: barcodes with high expression proximity to the HCC and barcodes with low expression proximity to the HCC in order to observe whether the initial expression gap between the two categories can be bridged by including more barcodes in the "metacell".

To make this comparison, we randomly selected 4 HCCs among the 17 in the cluster and for each of these HCCs we sampled barcodes whose linear regression R^2 score with the HCC is between 0.2-0.4 (low expression proximity) or 0.6-0.8 (high expression proximity). 500 barcodes were sampled for each category and these barcodes are grouped into "meta-cells" of different sizes and re-compared to the HCC.



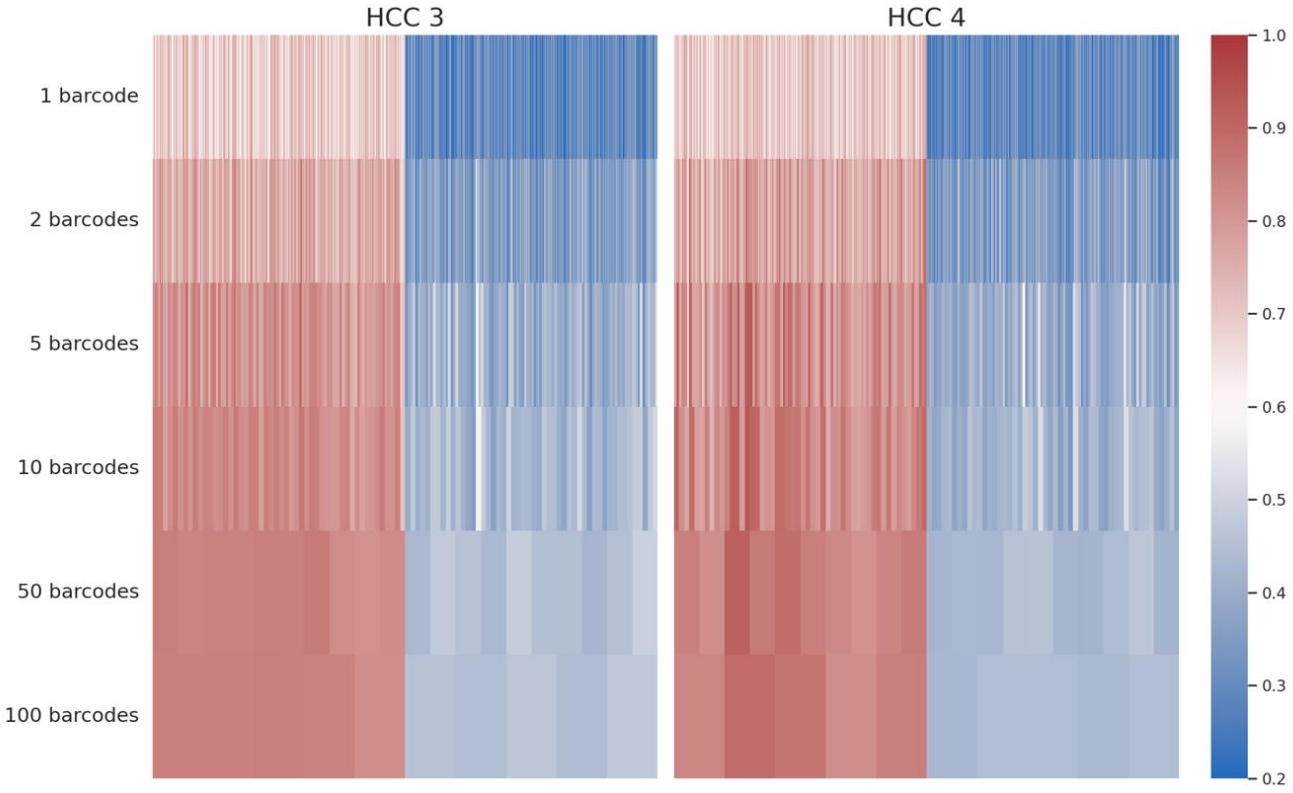


Figure 73. Heatmaps of HCC vs “metacells” R^2 scores for 4 HCCs from the Body Wall Muscle anterior cluster. First row represents individual barcodes which are subsequently grouped in "metacells" in following rows. Barcodes on the left are barcodes with strong initial correlation (0.6-0.8) and barcodes on the right are barcodes with low initial correlation (0.2-0.4)

We tested with several sizes of "metacells" to distinguish whether size was an important factor for the obtained R^2 score, with a maximum limit of 100 beyond which some HCCs had no more small barcodes matching the available selection criteria (Figure 73). What we observe globally is that the grouping of barcodes into "metacells" makes it possible to obtain an R^2 score higher than the average of the barcodes constituting the "metacell", this score sometimes going up to scores close to 1, indicating a very strong proximity of expression. As for the "metacells" containing barcodes with low proximity of expression, the gain in score is approximately of the same order but does not allow us to reach high scores, thus indicating that if the grouping of these barcodes makes it possible to refine their expression, they remain significantly different in terms of expression with respect to the HCC to which they are compared confirming as well the presence of multiple expression profiles within the originally defined cluster.

The size of the "metacells" has globally a positive impact on the scores obtained, with on average higher scores for bigger "metacells". However, there are a few cases where "metacells"

have a higher score with fewer barcodes, for example among the 10-barcode "metacells" of HCC 1, and this higher score is diluted in other "metacells".

Overall, the HCCs are seemingly barcodes containing a more complete expression than the rest of the dataset and thus are not artefact data. The small barcodes, whose expressions necessarily represent a fraction of the expression from their cell of origin, can thus be regrouped with other small barcodes which share a sufficiently close expression to simulate the expression their assigned cell type. Thus, the use of HCCs represents the possibility of having easier access to the expression of the original cell type and should be favored because of the lower number of cells required which minimize the eventual bias of regrouping multiple small cells. However, these barcodes are not necessarily required to study a cell type expression, the accretion of small barcodes into “metacells” seems amenable to approximate the expression profiles captured by HCCs. Our analysis indicate that this is very sensitive to clustering errors and accretion of cells with distinct patterns into “metacells” could give an inaccurate representation of the original cell type expression. In this perspective, it is thus more desirable to obtain and analyze HCCs rather than large number of low content cells.

4. Recovering unused data

Having previously confirmed that our processing of the Barcode/UMIs resulted in similar results to the author use of Cell Ranger, we also sought to understand why plenty of cells were found in the raw data that did not appear in the list of cells used in the authors analysis (Figure 74).

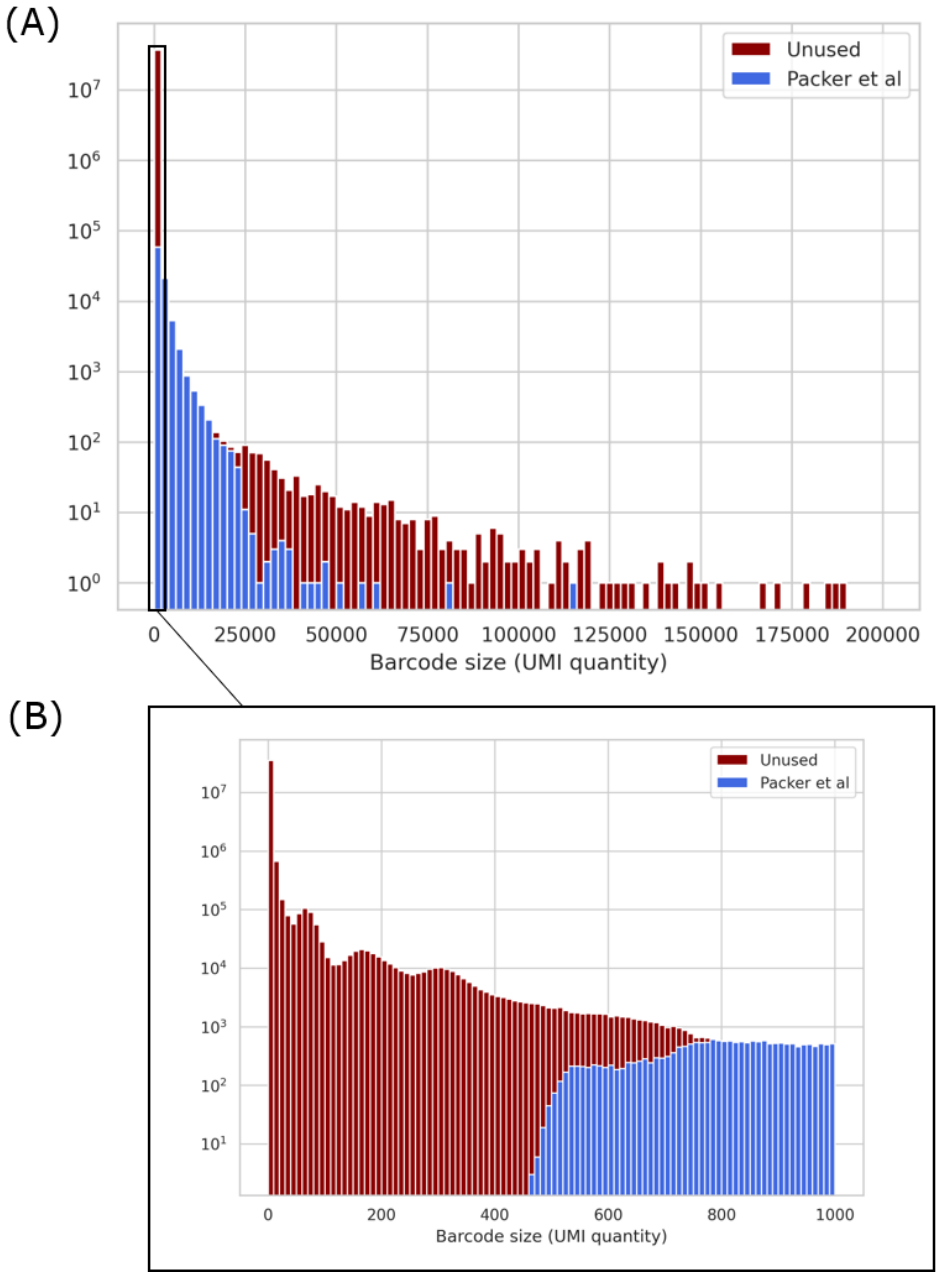


Figure 74. Size distribution of barcodes sequenced colored according to their use or not by the article's authors for (A) 0:200 000 range and (B) 0:1000 range.

Starting from the raw data, we found an incongruous number of theoretical cells with a total of 38 366 447 barcodes, a number out of all proportion to the initial number of cells sequenced. The explanation can be found in the barcodes containing less than 500 UMIs, which represent 38 204 946 of these barcodes. In order to filter out these outliers, the authors established an initial threshold of 1100 UMIs per barcode which they then reduced to 500 UMIs because of certain cell types that were more complex to sequence. This minimum threshold is a common practice in single-cell RNA-Seq studies to eliminate what are considered "empty droplets" in which contaminating or floating RNA would have been captured accidentally but this explanation appears to us as unconvincing. Indeed, all mRNA sequences linked to these UMIs map on the *Caenorhabditis elegans* reference transcriptome and therefore do not correspond to contaminating RNA (the UMIs linked to mRNA sequences unmappable on the reference have already been eliminated at this stage). Moreover, the quantity of RNA found in some of these cells seems inconsistent with the explanation of "empty droplets", if it seems possible for barcodes with 1 or 2 UMIs, it does not explain the presence of cells containing several hundred UMIs.

Our initial hypothesis was that these UMIs and barcodes most likely originated either from sequencing errors in the barcode sequence, from potential contamination or from errors during gel beads manufacturing leading to defective barcodes.

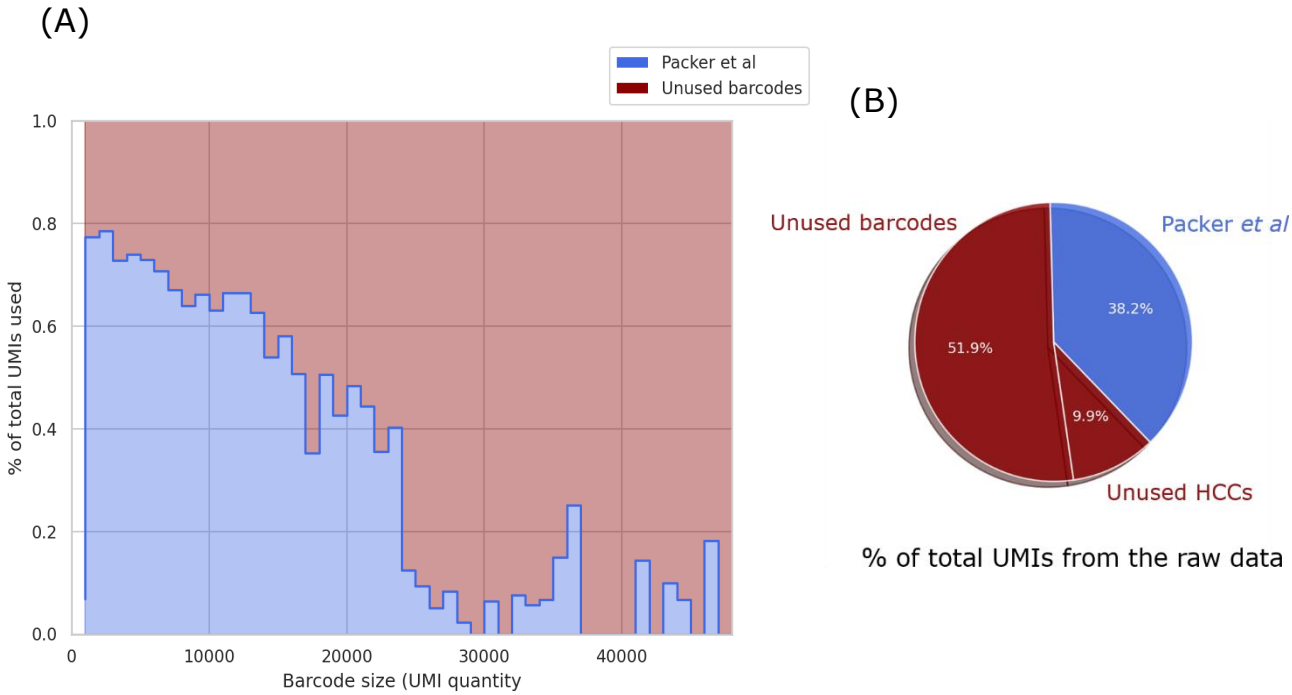


Figure 75. (A) Percentage of UMIs used in the Packer et al article per barcode size (cutoff at 50 000 for better visualization) and (B) total percentage of UMIs used in the Packer et al article and unused.

Almost two thirds of all the reads found in the raw data and mapped on the *C.elegans* transcriptome are not used in the article (Figure 75). We compared the sequence of those unused barcodes to the used ones to determine if they were the product of faulty barcode production or sequencing. Our aim was both to try to explain the amount of unused barcodes and determine if some UMIs were potentially recoverable by being re-integrated into already used barcodes. We searched for barcodes which could be degenerated (1 error allowed) from a used barcode and separated the results from unused barcodes of size >500 UMIs and < 500 UMIs as 500 UMIs was set in the article as the lower threshold acceptable for a barcode to be used.

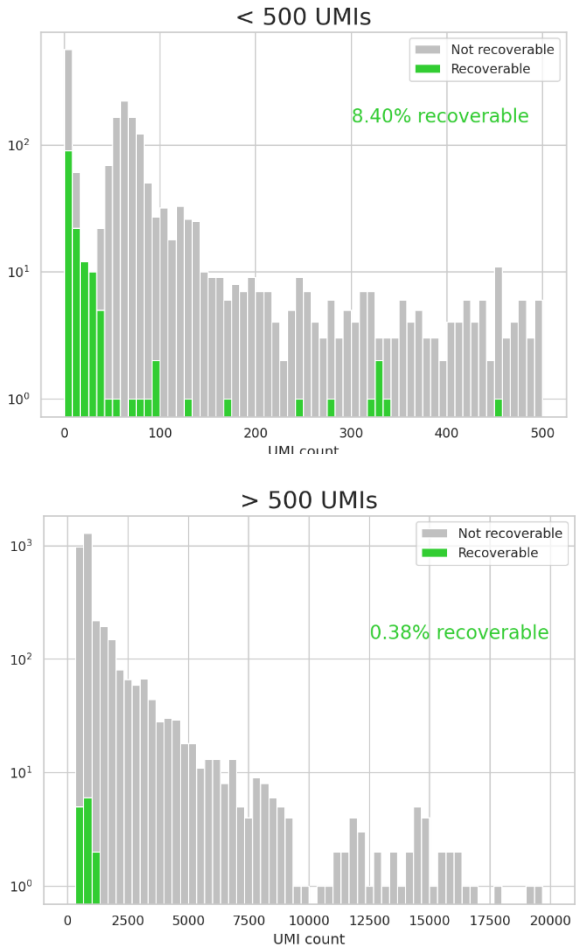


Figure 76. Distribution of barcodes size colored by recoverability status on the 300mn post_bleach dataset. 2000 barcodes were randomly sampled for the < 500 UMIs category and all barcodes were used for the > 500 UMIs.

We were able to find that a small part of the unused barcodes can be theoretically linked to barcodes used in the article by relying only on the barcode sequence (Figure 76). However, while the proportion of recoverable barcodes is much higher for barcodes < 500 UMIs than

for barcodes > 500 UMIs (8.4% instead of 0.38%), the vast majority of unused barcodes do not show any evidence of being derivative of accepted barcodes.

We also noted that a significant proportion of the barcodes > 500 UMIs are not used by the authors, particularly for the barcodes with the larger amounts of cells, representing a second important loss of data. If not using the data from cells < 500 UMIs makes sense because of the individual low content, the non-use of barcodes of higher size is a less tolerable problem especially in situations of low initial quantity of material such as ours in a micro-irradiation framework.

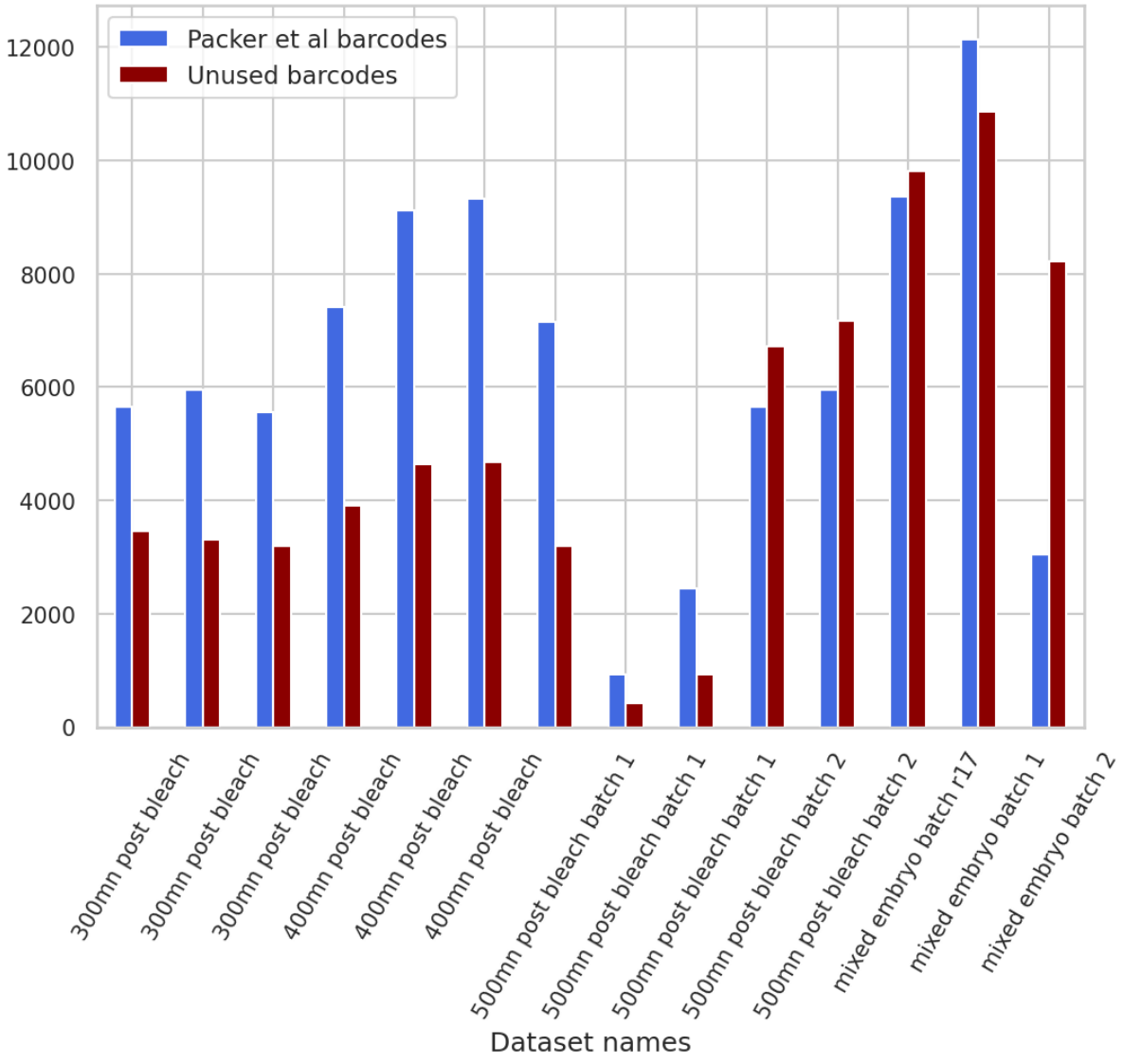


Figure 77. Number of used and > 500 UMIs unused barcodes per condition.

This situation seems even more inconsistent for HCCs, a large part of which are noticeably left off the study, despite the fact that they contain far more information to characterize the full

expression of their cell of origin as opposed to smaller cells which need to be grouped to simulate a similar result (Figure 77). The exclusion of these HCCs is not due to their size since some of the barcodes used in the study are sufficiently big to be considered HCC, the largest of which contains 126 000 UMIs. The variations in the quantity of unused cells per condition also suggests that these cells are not "protocol waste", the result of a known defect in the library preparation method that would be corrected by excluding artefact cells, but rather actual cells that were excluded on the basis of an arbitrary factor.

Of the 89,700 cells retained in the study, 27% could not be assigned to a cell type and 11.5% were loosely linked to one or multiple cell types without being able to be assigned with certainty (*e.g.* “Parent of X cell type” or “type X or type Y”) which means the cells excluded from the study were not excluded for lacking a recognizable cell type (Figure 78). We could find no explanation of the exclusion of this large body of data and indeed no evidence that the authors are aware of their existence.

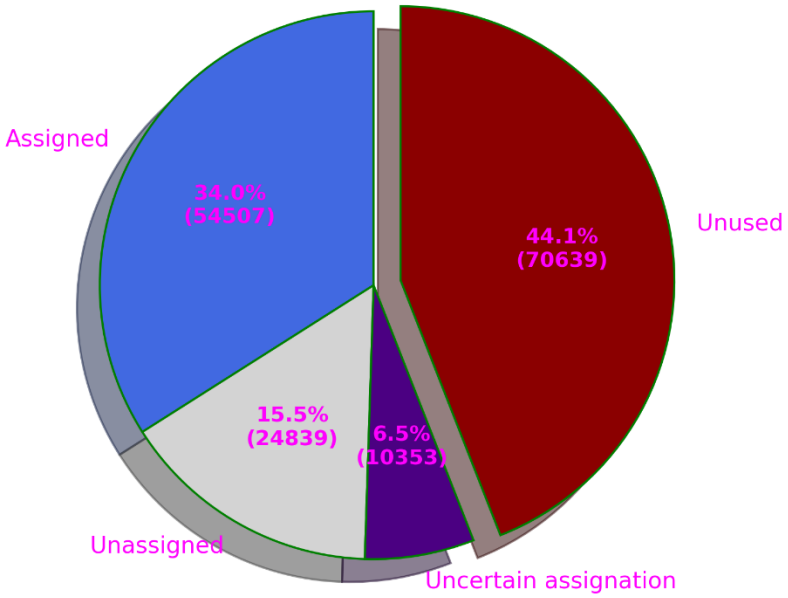


Figure 78. Assignment status of the barcodes used by the authors. Uncertain assignation corresponds to assignations such as “Parent of X”, “Type X or type Y”.

We therefore sought to analyze these unused barcodes by comparing them to the cells used that would have passed the filter in question. We first sought to compare the expression between the 2 categories of cells based on the most expressed genes and the UMI/gene ratio which, even if they are not very detailed factors, provide an overall view of the "expression behavior" of the

cells. The second comparison factor we used is the quality score of the reads contained in the unused cells. Although these reads were of sufficient quality to be aligned with the *C. elegans* transcriptome, it is possible that a minimum threshold was set and thus eliminated many reads and by cell association.

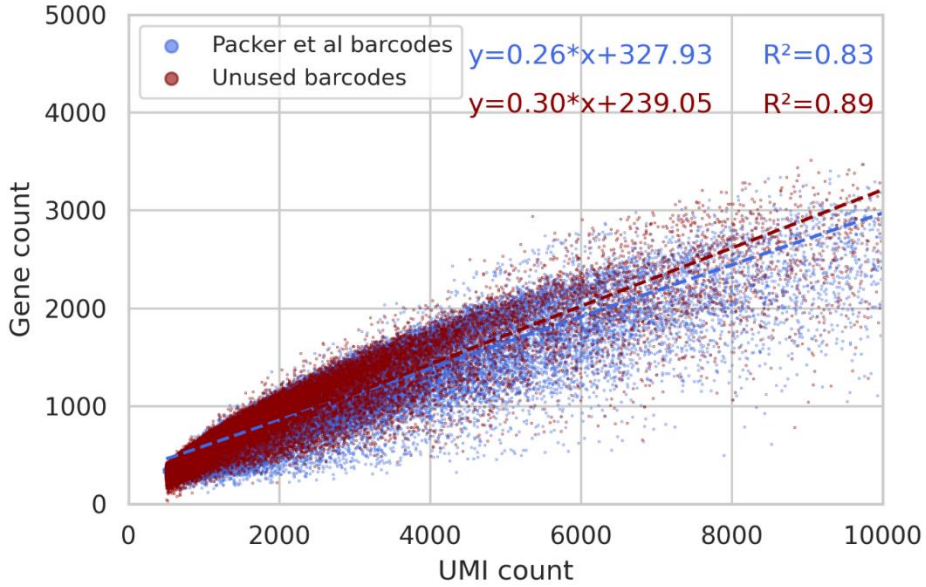


Figure 79. Correlation between the number of UMIs and the number of genes found per unique barcode with > 500 UMIs unused barcodes colored in dark red and cells kept by the authors in blue.

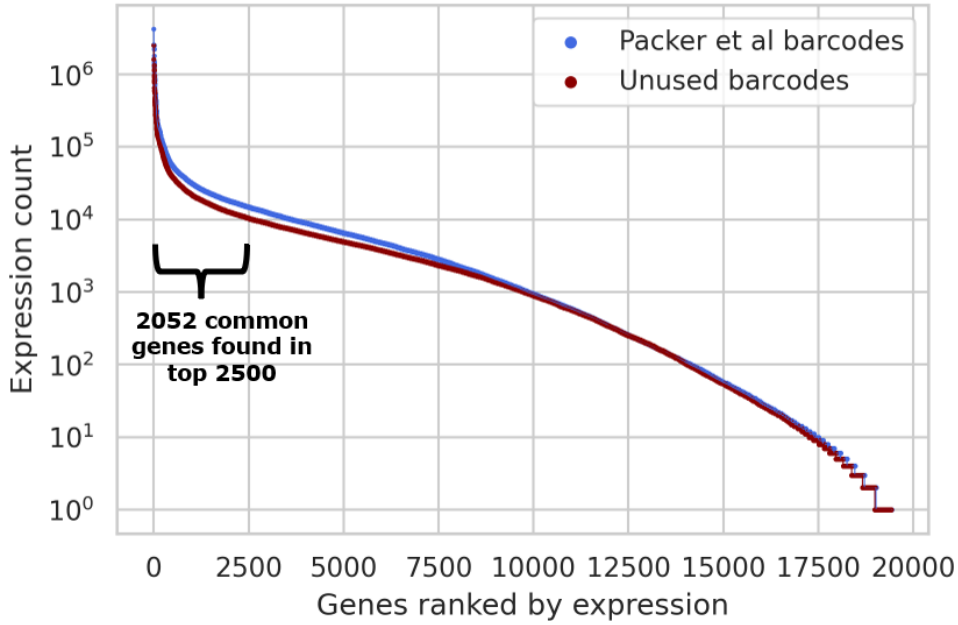


Figure 80. Expression levels of genes ranked in decreasing order

No significant difference can be observed between the UMI/gene count ratio distribution of unused cells and cells kept by the authors as illustrated by the linear regression curves

illustrating the similarity in behavior, and not even outlier values can be found among the unused cells that would explain the exclusion of a part of these cells (Figure 79).

As for the most expressed genes, we observe a similar profile between the unused barcodes and the barcodes used by the authors with a peak of genes much more expressed than the average of the genes (Figure 80). These most expressed genes are globally similar between the two categories with 2052 genes in common among the 2500 most expressed genes and this despite the fact that the barcodes do not come in equal proportions from all extraction conditions which adds to the variance between the two categories.

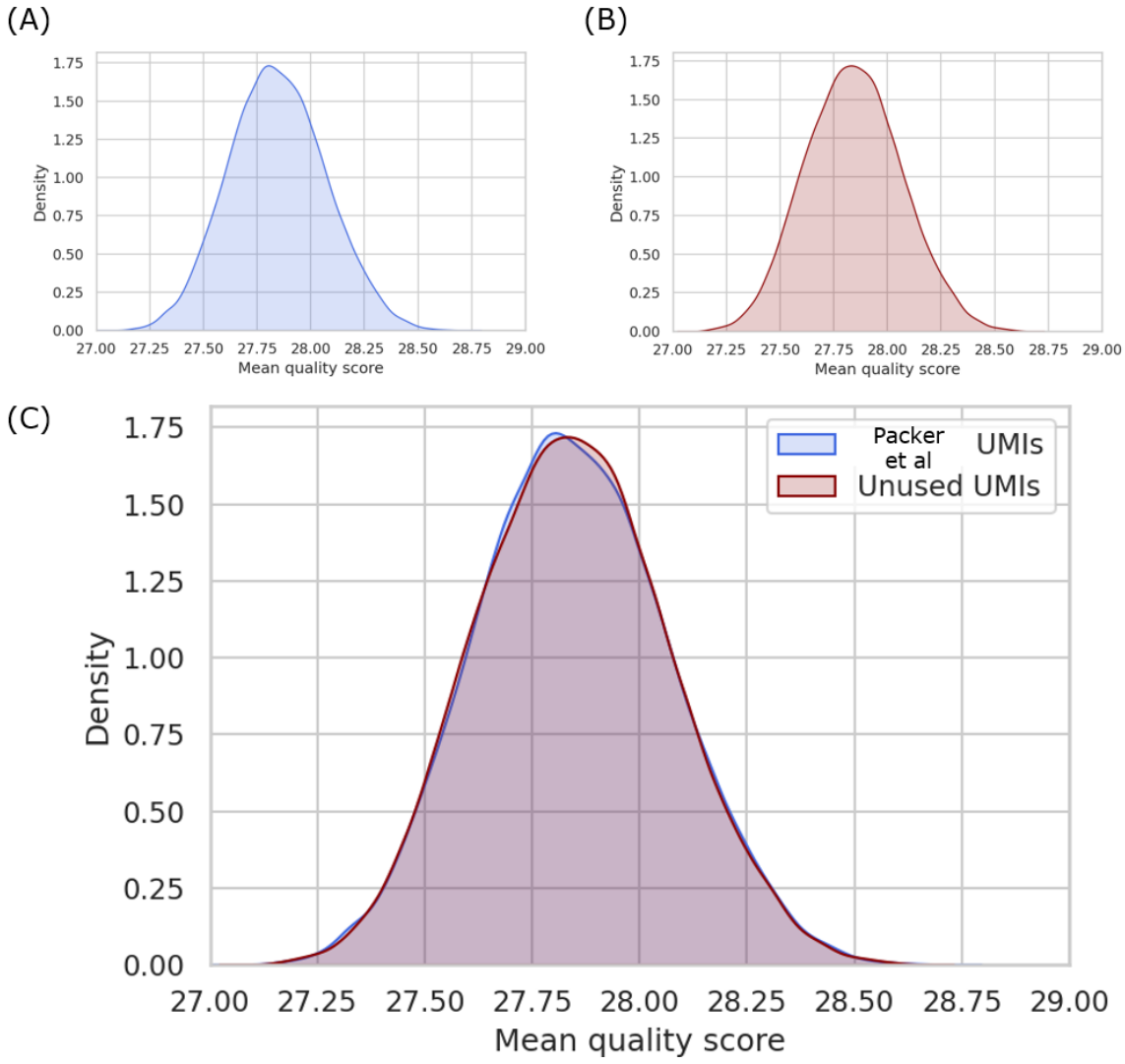


Figure 81. Mean quality score for reads from (A) authors barcodes, (B) unused barcodes and (C) both. 25 000 reads were randomly sampled from the cells in question with a limit of 50 reads per cell.

The quality scores also show no illustrate any meaningful difference between barcodes used by the authors and unused barcodes, with the density curve of the quality measure of these 2 categories almost completely overlapping (Figure 81). No outlier value was found when studying these quality scores which seems to indicate that these reads had already passed a filter on the quality score and that the removal of the barcodes occurred afterwards.

In summary, the barcodes not used in the study have similar enough characteristics that we cannot distinguish them from the rest of the cells at this stage of data processing. Another possible theory would be that the Cell Ranger pipeline uses a barcode “whitelist”, as the barcodes used are normally known prior to the microfluidic and library preparation step. If this step would bring a logical explanation to this problem, it would not be satisfactory regarding the reliability of the data produced. Outside of a possible but consequent discrepancy in the “whitelist”, this hypothesis would suggest that all the barcodes not found are due to contamination or outliers, something we have not been able to determine as previously demonstrated. Furthermore, such a large amount of contamination would be worrisome, as the large number of barcodes containing less than 500 UMIs is already quite problematic with respect to the performance of this method, and the possibility that barcodes containing tens or hundreds of thousands of UMIs could be caused by contamination would drastically decrease the confidence in the other whitelisted barcodes. As it stands, we do not have the ability to determine the validity of this hypothesis, as the barcode whitelists are 10x Genomics material and not available in the supplementary materials of the article or the public website of the company.

Although the exclusion of these barcodes was normally done before assignment to a cell type by clustering, due to the presence of unassigned barcodes among the authors data. We sought to test both whether these unused barcodes have an outlier expression and whether they can be recovered despite their exclusion from the dataset by inserting them into the used barcode dataset (*i.e.* an addition of 70,639 barcodes) and repeating the UMAP clustering step. The inclusion of cells with aberrant expression in a dataset of "normal" barcodes would normally result in either a complete disruption of the clusters initially defined by the authors or a clear separation of the unused barcodes from those used by the authors

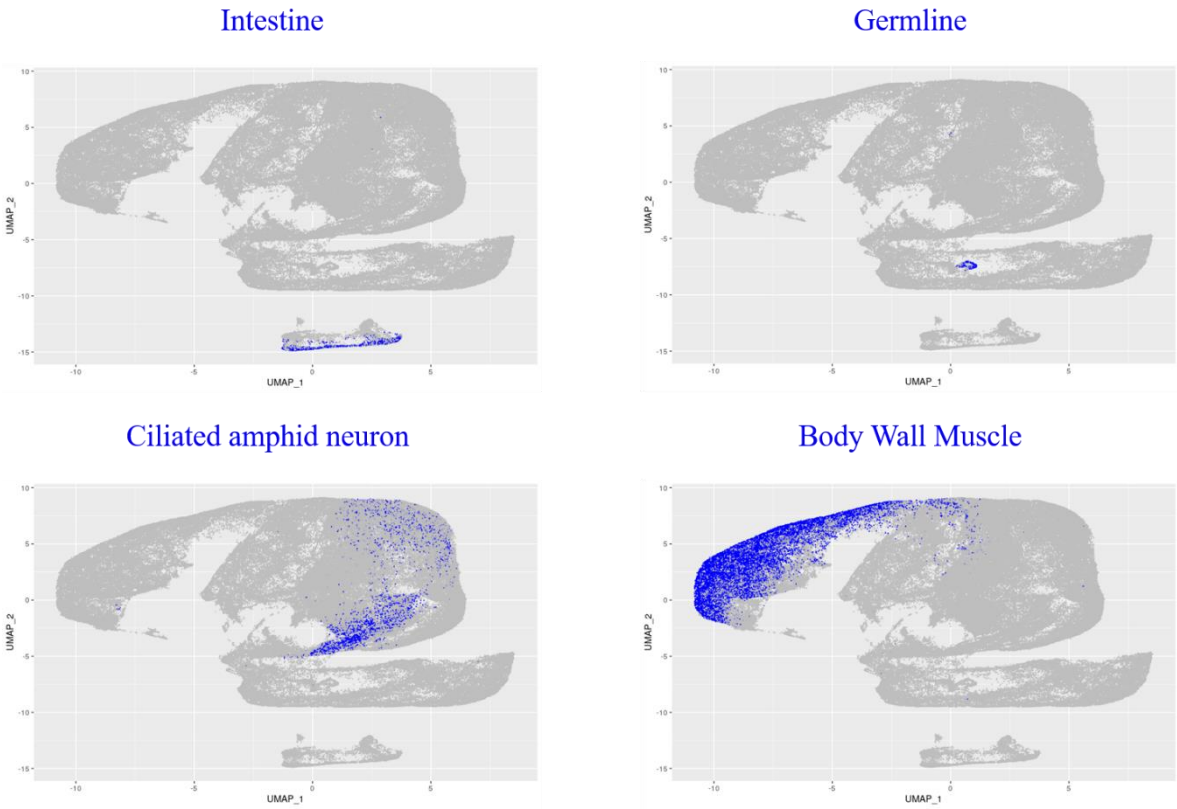


Figure 82. UMAP visualization of clustering of the dataset containing unused barcodes + barcodes used by the authors with coloring for 4 given cell types based on authors assignation.

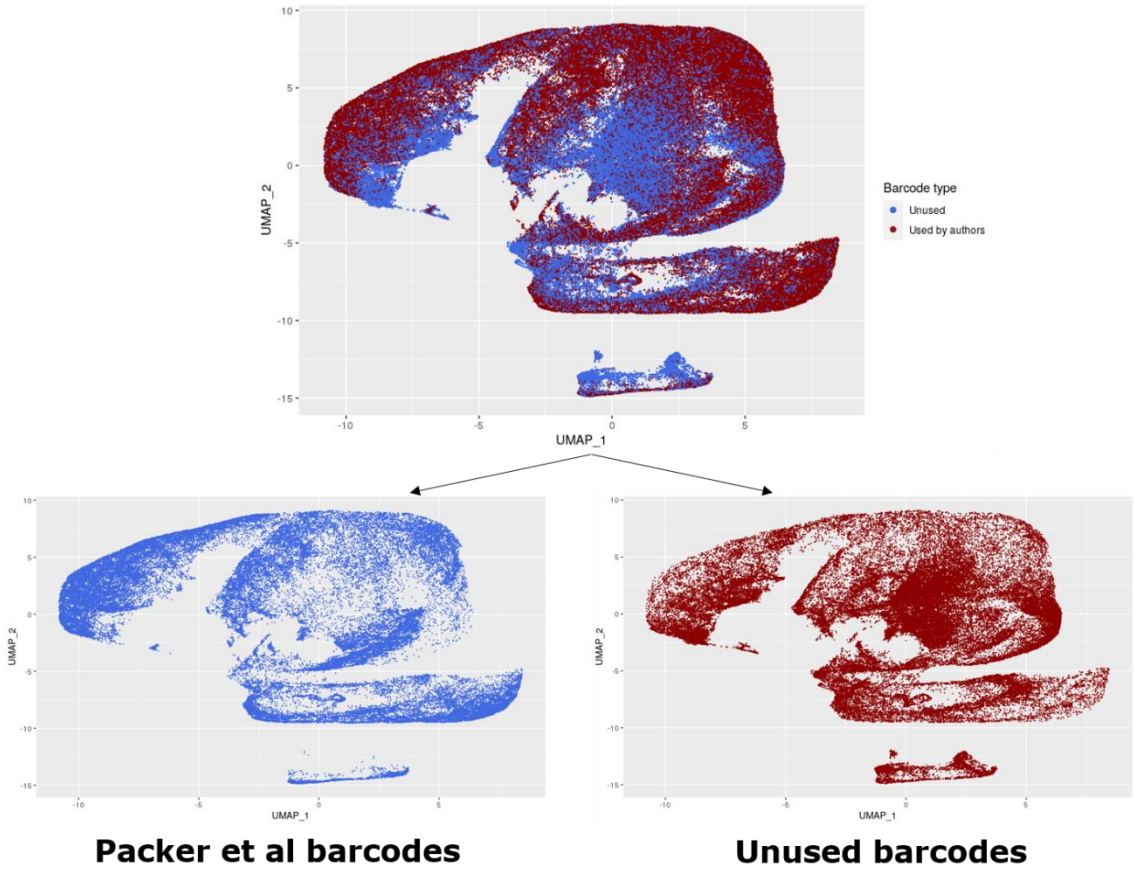


Figure 83. UMAP visualization of clustering of the dataset containing unused barcodes + barcodes used by the authors, coloured depending on the category of use of the barcode.

On the UMAP clustering result of all barcodes > 500 UMIs, we can see that despite the addition of a significant number of cells, the clusters identified by the authors and associated with a cell type are globally preserved (Figures 82-83). Notable differences are nevertheless visible, for example for the cluster assigned to "Germline" which was previously separated from the other barcodes and which is now reintegrated into the rest of the dataset although the barcodes of this cluster remain clustered together, as well as the general structure of all clusters combined but this structure is secondary to the good internal conservation of the clusters. No schism appears that would separate the unused barcodes from the authors barcodes, which reinforces the hypothesis that these are indeed data from authentic cells.

Looking at the 2 categories of barcodes separately, we can see that they share the same overall structure and that many unused barcodes share significant proximity to assigned barcode clusters suggesting that they are barcodes from the cell types in question rather than contamination. There are regions of uneven density in unused barcodes and barcodes used by the authors which could correspond either to clusters representing a cell type not detected in the initial study or to conglomerates of barcodes whose expression is too confusing to assign it to clusters of cells with distinct expression. In the case of the latter, this does not necessarily mean that these barcodes can be considered as having an aberrant expression, because of the close proximity they maintain with the rest of the data and that new iterations of clustering on subsets of the data could allow them to be assigned.

Discussion

Our objective when looking at the data produced in this landmark paper for cell expression by Single-Cell RNA-Seq on *Caenorhabditis elegans* was to evaluate data produced under "optimal" conditions where the initial material can be problematic, in this article case the study of cellular expression in embryos, in order to better anticipate problems that we might encounter in our specific case. The fact that this paper was conducted using the 10X Genomics microfluidic method and the associated analysis pipeline, Cell Ranger, also allowed us to study data generated on one of the most popular single-cell RNA-Seq methods today. We thus wanted to determine if data are lost during production or analysis as well as the typical expression profile obtained from the sequenced cells to evaluate how this technique might fit in our case study.

We started by analyzing the UMAP/Louvain clustering product which is the focus of the bioinformatics analysis of the data to assign barcodes to clusters and thus to cell types. We found relatively variable cluster content with many barcodes assigned to clusters in which they share little or no proximity to other barcodes or subgroups sharing only weak proximity grouped within a single cluster. Apart from these questionable assignments, the UMAP/Louvain method groups barcodes sharing high proximity with each other and its other qualities still make it an essential tool for this cell assignment step in Single-Cell RNA-Seq studies[270]. It could however be considered to "refine" the UMAP/Louvain clustering step by linear regression, as the use of this method to cluster all the barcodes is not feasible from a practical point of view because of the computation time which would be required. It could be used to dissociate within the clusters the possible subgroups formed and to remove the barcodes presenting a weak proximity of transcriptomic expression. This solution could help to reach a more defined cellular expression of the cell types and in particular our cell type of interest by decreasing quantitively but increasing qualitatively the expression found in the clusters.

We also observed that the distribution of the data produced was mainly centered on small barcodes, mainly between 1000 and 2000 UMIs, for a rather minimal amount of High Content Cells which nevertheless allow a better definition of the expression of their cell of origin. In an analysis with a small amount of initial material, one would therefore expect to produce even fewer HCCs and thus potentially miss some of the expression of cell types. Our analysis assessed that the composition of HCCs and smaller barcodes are broadly similar and more importantly that the expression of an HCC can be recovered by clustering of these small

barcodes. This implies that it is not necessary to obtain HCCs to study a cell type, provided of course that enough small barcodes are produced for a given cell type, but that those HCC could offer a preferable solution by obtaining directly the expression of a given cell rather than by groupings that may cause bias.

Finally, the re-processing of raw data made it possible to find a significant quantity of barcodes that were not integrated in the article and thus representing potentially lost data. By comparing these unused barcodes with those from the article, we were not able to find significant differences that could justify their exclusion. However, no reference to these data is made in the article or in the supplementary materials to explain the reason for their exclusion. We have not come up with a satisfactory hypothesis to explain this situation, the only one that seems likely to us being an automatic sorting of the data on the basis of a lack of whitelist validation by the Cell Ranger analysis pipeline used in the article. In any case, in the absence of arguments to the contrary at the present time, these excluded data seem to us usable as is and would allow us to increase the total amount of data produced by almost half.

In summary, this analysis of single-cell data has allowed us to better understand the product of this type of sequencing and the challenges to be anticipated for application to our study context. It should be noted, however, that this analysis was only done on one article, using a specific method and on a specific model. We have briefly studied other single-cell RNA-Seq articles on *C. elegans* but none of them allowed an analysis as deep as Packer *et al*, notably because of the very extensive supplementary data.

The most pressing issue identified in this article is the significant loss of data by this Single-Cell method both by the cell capture method, for which an order of magnitude of efficiency is available in the paper, and by the bioinformatics analysis of the sequenced libraries, either because unused or unassignable. A part of these data seems however to be recoverable in order to mitigate this loss but a sufficient quantity of initial material appears to be essential nonetheless. Therefore, it appears complicated at this time to perform Single-Cell RNA-Seq analysis from the amount of material and extracted RNA produced in our micro-irradiation experiments, especially to study the expression of two specific cells in an organism composed of several hundred cells. A possible solution would be to produce a large number of replicates in order to pool them into a bulk sample containing sufficient material for this type of analysis

or to use other *C.elegans* mutant lines whose irradiated part could be easily dissected in order to obtain an initial material composed largely of the cells of interest.

Once this obstacle of the quantity of material is overcome, the rest of the typical bioinformatics analysis, although it can be corrected on certain points, seems to us globally adapted to lead to a cellular expression of the desired cell types in order to then identify the differences in expression between the conditions.

Conclusion

My thesis work focused on the integration of Oxford Nanopore Technologies (ONT) 3rd generation sequencing method and the associated bio-informatic tools in an interdisciplinary team studying the interactions of ionizing radiation and metal oxide nanoparticles with living organisms at the interface of physics, chemistry and biology. The originality of the research projects carried out in this team is based on the conjoint use and development of the Monte Carlo simulation codes Geant4/Geant4-DNA allowing to model the charged particle-living interactions at the DNA scale, and two nuclear microprobes allowing an irradiation and a quantification of the chemical elements at the cell or sub-cellular compartment scale. These projects are articulated in a dynamic aiming to progress from analyses on restricted models (DNA in suspension, *in cellulo* cells) to the *in vivo* multicellular model *Caenorhabditis elegans* to achieve a characterization of the radio- and nano-induced responses at the cellular level in a living organism. The main techniques for the analysis of biological response being established within the team were imaging by confocal microscopy, developmental monitoring by flow cytometry and expression analysis of genes of interest by qPCR. The integration of a sequencing method was intended to provide a high-throughput analysis method allowing the analysis of the entire cellular pathways of an organism by transcriptomic analysis in order to correlate their possible dysregulation with the results obtained via the other methods. The use of ONT's long-read sequencing method also presented the possibility of studying cellular damage from new angles: (i) by measuring DNA fragmentation thanks to the long-read capacity of the sequencer allowing to theoretically sequence DNA molecules of any size, (ii) by studying modified bases, especially RNA bases, which is made possible by the Direct-RNA sequencing of this sequencer.

However, the applicability of this method in this particular study context remained to be demonstrated, particularly on the ability of the sequencer to provide results usable with Geant4-DNA as well as on its use on samples with a low quantity of initial material. My thesis was therefore divided into 3 main parts dealing with the integration of sequencing and bio-informatic in the research projects of the iRiBio team. Parts I and II focus respectively on a concrete application of ONT's 3rd generation sequencer to an analysis of radiation-induced DNA fragmentation on lyophilized DNA and to a transcriptomic analysis of the radiation- and nano-induced cellular response on *C.elegans*. Part III focused on an exploration of Single-Cell RNA-Seq data in order to evaluate the type of results produced and the potential applicability to the iRiBio team's study context with a view to progressing towards a single-cell scale cellular response.

Part I. Direct measurements of DNA strand breaks by long-read sequencing.

The use of long-read sequencing allowed us to observe and quantify the radiation-induced fragmentation on two DNAs: the genomes of the plasmid pBR322 (4,361 bp) and the phage Lambda (48,502 bp). In spite of the important difference in size between these two molecules, the fragmentation probabilities, considering an exponential model, present almost identical values. The next step is to apply the same analysis on the genome of the T4 phage (168 903 bp), a molecule in the "ultra-long" read category, in order to progress towards genomes closer in size to those of living organisms. This move to longer molecules is still in its infancy, however, due to new issues to be resolved, particularly in terms of the technical limitations of the current sequencer and protocols.

The results obtained on pBR322-DNA were used as elements of comparison and validation of the simulations carried out on Geant4-DNA which thus made it possible to determine values of irradiation parameters (breakthrough energy threshold, maximum radius of dose deposition) starting from a base of values available in the literature. Simulations of the same type on the Lambda phage are also in progress.

With this foundation established, it is now possible to move towards sequencing DNA suspended in water to study fragmentation due to radiolysis of water, which is also simulatable in Geant4-DNA. This would also allow the study of modified bases in DNA that may be caused by radiolysis of water, as an official ONT tool is already available to analyze modified bases (only methylation at the time of writing). At this stage, the DNA bases simulated in Geant4-DNA are all identical, but progress is being made to simulate DNAs of defined sequence. Our sequencing results seem to show that different bases do not exhibit different fragmentation probabilities but this implementation is essential for the analysis of modified DNA bases.

In addition to these advances in sequencing, another method for studying radiation-induced DNA fragmentation is being developed in the team within the framework of the INSIDE project led by F. Gobet. This approach is based on the use of real-time microscopy of T4 phage DNA fragmentation in water on AIFIRA and on the measurement of the size of the fragments produced according to their diffusion constant in water. This method has the advantage of being able to quantify fragmentation at the scale of single DNA molecules and the results obtained will be additional elements for comparison with the Geant4-DNA simulation codes.

Part II. Analysis of radio- and nano-induced cellular expression by transcriptomic analysis

The study of micro-irradiated *C.elegans* worms by sequencing and transcriptomic analysis could be carried out thus validating the developed protocol aiming at defining the modalities of sample preparation, irradiation and analysis of these complex samples. This first experiment allowed us to observe the first elements of the cellular response to irradiation, in particular the activation of genes coding for heat shock proteins and the response to cuticle damage. Some problems of variability were observed between the samples, in particular on 2 samples showing an abnormal cellular expression and because of a slight desiccation phenomenon which can be observed *via* differentially expressed genes. The low magnitude of the observed cellular response confirms our intention to progress towards the use of Single-Cell RNA-Seq methods which would theoretically allow the observation of the radiation-induced cellular response directly on the cells concerned.

The main challenge remaining in this protocol is the low amount of material produced per experiment and therefore low amount of RNAs extracted. This limitation prevents us at this stage to use Direct-RNA sequencing methods that would allow to study the modified bases. Several options are possible to overcome this lack: mass production of samples to pool them (but the availability of the irradiation line is very limited), dissection of the targeted regions of the worm, improvement of the protocol (more efficient RNA extraction, increase of the irradiation capacities by development of an automatic worm targeting software, etc).

A transcriptomic analysis was also performed on *C.elegans* worms exposed to TiO_2 nanoparticles. In these samples, we were able to observe a significant cellular response despite a total absence of internalization of the NPs which remain in the intestinal lumen, thus revealing a new mode of action of these particles on the biological material. However, this effect was only observed on P25s and not on TNs, despite the fact that the latter presented the most important toxicity during the previous *in vitro* experiments. This difference in response is probably due to the shape and size of these NPs, the TNs being more difficult to ingest and therefore entering the worms' intestine less easily. These results also confirm our intention to progress towards Single-Cell RNA-Seq which would allow us to study the cellular response of the intestinal cells which are the most directly exposed to the NPs, which would perhaps also allow us to observe a toxic effect of the NPs on these cells. This experiment is the only one in which we could test the use of a modified base detection software. Although it did not yield

significant results, it allowed us to become familiar with this tool in order to integrate it in future Direct-RNA experiments.

In these two experiments, although the capabilities of the microbeam lines allow either to irradiate a precise region or to quantify the amount of nanoparticles at the cell level, the homogeneity of the dose and thus of the biological response in the whole sample cannot be guaranteed. The deposition of these physical agents on the biological material is done in a random manner which means that, although everything is done to minimize the impact, a level of heterogeneity persists. One prospect to overcome this shortcoming would be to analyze single worms, something that is already possible by microscopy but remains to be done for transcriptomic analysis. For this purpose, a protocol is being developed by D. Dupuy to optimize the RNA extraction protocols in order to produce enough material for single worm sequencing on the ONT sequencer. This protocol is still under development but could be a solution to avoid the risk of dose heterogeneity

Part III. Evaluation of single-cell RNA-Seq applicability in a low yield and high complexity experiment.

As described above, our goal is to move toward using the Single-Cell RNA-Seq which would allow us to study the radio- and nano-induced cellular response at the cellular level rather than the organism level.

However, as this technology is still new, the analysis methods developed and the relevance of the results produced are still under debate. We therefore analyzed data from a reference article produced from *C.elegans* embryos in order to judge the potential application of this method to our case of complex and low yield samples. Through this analysis, we were able to identify the following: (i) the UMAP cell clustering method is globally efficient and groups highly correlated cells but suffers from some anomalies (cells grouped by mistake, distinct expression groups grouped into a single cluster) which seem however correctable by linear regression; (ii) some cells show a much more complete expression than the majority of the other cells and thus represent a better snapshot of their cell of origin, these cells are not mandatory for the analysis, as the grouping of smaller cells allows to approximate their expression, but they represent a significant advantage if it is possible to obtain them on our targeted cell types; (iii) a significant amount of data present in the raw data seems to be unused despite the absence of an obvious exclusion factor for a significant portion of these data, although these data are excluded by the

analysis pipeline, they still appear to be usable. In summary, the current method of analysis of these data appears to us to be relatively robust, although it can be corrected on a few points.

Regarding the analysis of specific cells, particularly for micro-irradiation studies, it appears difficult to consider the use of this method at the present time due to the number of cells required to perform a grouping by cell type and the small number of cells studied in our configuration. Indeed, although the analysis of the more abundant peripheral cells can be interesting to study the "proximity" effect, the expression of the 2 targeted cells (Z2-Z3) risks to be "drowned" in the totality of the sequenced cells. It would be necessary to either produce a sufficiently large number of cells for the expression of Z2-Z3 to be distinguishable or to limit the sequenced cells to the target region (dissection).

In the perspective of the application of this technology, it should be noted that it will probably be limited in its usage for the study of samples irradiated at high doses. The observation by confocal microscopy revealed that the targeted cells did not grow at all in a large proportion of cases due to the cellular damage caused. This cellular damage, if it is too severe and destroys the cell or at least the RNAs, will therefore prevent the use of Single-Cell RNA-Seq on these particular cells. However, peripheral cells, which are normally less exposed, will still available to study and should make it possible to study phenomena such as the bystander effect or the result of cellular cascades on tissue development caused by the destruction of precursor cells.

Through these different research projects, the technology of ONT sequencing and the bioinformatics methods necessary for its use have been integrated into the team's skills. This integration has allowed to produce first results in the study and modeling of the biological consequences of IR and metal oxide NPs. Several phenomena having slowed down the production of samples in the various research projects (Covid pandemic, technical breakdowns, etc.), we were not able to make as much use as we would have liked of all the capacities of the sequencer, particularly the analysis of modified bases. We were able to implement the use of software for the analysis of this type of data but we did not have the opportunity to study in more detail on a larger number of samples the interesting perspectives that could be drawn from it: correlation between modified bases and differential expression, relationship between dose and quantity of modified bases, analysis of specific target genes (PCN-1). The development of these tools by ONT is to be followed as announcements have been made about the future

implementation of new modified bases in their official software (8-oxoguanine especially which is a common marker of radiation-induced DNA damage). Another solution would be the in-house development of a modified base detection tool on DNA or RNA to study the desired bases using synthetic sequences for calibration.
The other important step in the progression of sample analysis by sequencing is the implementation of Single-Cell RNA-Seq. The application of this method will pose a significant challenge on sample preparation in order to meet the necessary conditions for its use. Codes for processing these data have been developed but they only allow a preliminary analysis that will have to be completed and adapted to future data.

In conclusion, the implementation of 3rd generation sequencing and bioinformatics techniques in the team's methods during these 3 years is part of a dynamic progression towards the use of some of the most promising methods in a specific framework of study of radio- and nano-induced biological damage[271].

References

[1] Gregor Mendel, « Versuche über pflanzen-hybriden », *Verhandlungen der naturfoschung Vereins*, nº 4 (1866): 3-47.

[2] H. M. Vernon, « Elemente Der Exakten Erblichkeitslehre », *Nature* 81, nº 2084 (octobre 1909): 424-424, https://doi.org/10.1038/081424a0.

[3] O. T. Avery, C. M. Macleod, et M. McCarty, « STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III », *The Journal of Experimental Medicine* 79, nº 2 (1 février 1944): 137-58, https://doi.org/10.1084/jem.79.2.137.

[4] J. D. Watson et F. H. Crick, « Molecular Structure of Nucleic Acids; a Structure for Deoxyribose Nucleic Acid », *Nature* 171, nº 4356 (25 avril 1953): 737-38, https://doi.org/10.1038/171737a0.

[5] Doris T. Zallen, « Despite Franklin's Work, Wilkins Earned His Nobel », *Nature* 425, nº 6953 (4 septembre 2003): 15, https://doi.org/10.1038/425015b.

[6] S. Brenner, F. Jacob, et M. Meselson, « An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis », *Nature* 190 (13 mai 1961): 576-81, https://doi.org/10.1038/190576a0.

[7] F. Jacob et J. Monod, « Genetic Regulatory Mechanisms in the Synthesis of Proteins », *Journal of Molecular Biology* 3 (juin 1961): 318-56, https://doi.org/10.1016/s0022-2836(61)80072-7.

[8] F. Crick, « Central Dogma of Molecular Biology », *Nature* 227, nº 5258 (8 août 1970): 561-63, https://doi.org/10.1038/227561a0.

[9] « Central Dogma - An Inheritance Mechanism », Byju's, consulté le 3 janvier 2022, https://byjus.com/biology/central-dogma-inheritance-mechanism/.

[10] F. Sanger et H. Tuppy, « The Amino-Acid Sequence in the Phenylalanyl Chain of Insulin. I. The Identification of Lower Peptides from Partial Hydrolysates », *The Biochemical Journal* 49, nº 4 (septembre 1951): 463-81, https://doi.org/10.1042/bj0490463.

[11] F. Sanger et E. O. P. Thompson, « The Amino-Acid Sequence in the Glycyl Chain of Insulin. I. The Identification of Lower Peptides from Partial Hydrolysates », *The Biochemical Journal* 53, nº 3 (février 1953): 353-66, https://doi.org/10.1042/bj0530353.

[12] Clyde A. Hutchison, « DNA Sequencing: Bench to Bedside and Beyond », *Nucleic Acids Research* 35, nº 18 (2007): 6227-37, https://doi.org/10.1093/nar/gkm688.

[13] W. Min Jou et al., « Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein », *Nature* 237, nº 5350 (12 mai 1972): 82-88, https://doi.org/10.1038/237082a0.

[14] A. M. Maxam et W. Gilbert, « A New Method for Sequencing DNA », *Proceedings of the National Academy of Sciences of the United States of America* 74, nº 2 (février 1977): 560-64, https://doi.org/10.1073/pnas.74.2.560.

[15] A. M. Maxam et W. Gilbert, « Sequencing End-Labeled DNA with Base-Specific Chemical Cleavages », *Methods in Enzymology* 65, nº 1 (1980): 499-560, https://doi.org/10.1016/s0076-6879(80)65059-9.

[16] F. Sanger et A. R. Coulson, « A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase », *Journal of Molecular Biology* 94, nº 3 (25 mai 1975): 441-48, https://doi.org/10.1016/0022-2836(75)90213-2.

[17] F. Sanger, S. Nicklen, et A. R. Coulson, « DNA Sequencing with Chain-Terminating Inhibitors », *Proceedings of the National Academy of Sciences of the United States of America* 74, nº 12 (décembre 1977): 5463-67, https://doi.org/10.1073/pnas.74.12.5463.

[18] Graziano Pesole et Cecilia Saccone, « Handbook of comparative genomics: principles and methodology » (New York: Wiley-Liss, 2003), p133.

[19] Z. G. Chidgeavadze et al., « 2',3'-Dideoxy-3' Aminonucleoside 5'-Triphosphates Are the Terminators of DNA Synthesis Catalyzed by DNA Polymerases », *Nucleic Acids Research* 12, nº 3 (10 février 1984): 1671-86, https://doi.org/10.1093/nar/12.3.1671.

[20] Michel Gauthier, « Simulation of polymer translocation through small channels: A molecular dynamics study and a new Monte Carlo approach » (2007).

[21] Jay Shendure et Hanlee Ji, « Next-Generation DNA Sequencing », *Nature Biotechnology* 26, nº 10 (octobre 2008): 1135-45, https://doi.org/10.1038/nbt1486.

[22] Linnea M. Baudhuin et al., « Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing », *The Journal of Molecular Diagnostics* 17, nº 4 (1 juillet 2015): 456-61, https://doi.org/10.1016/j.jmoldx.2015.03.004.

[23] Eugene Y. Chan, « Next-Generation Sequencing Methods: Impact of Sequencing Accuracy on SNP Discovery », in *Single Nucleotide Polymorphisms: Methods and Protocols*, éd. par Anton A. Komar, Methods in Molecular Biology™ (Totowa, NJ: Humana Press, 2009), 95-111, https://doi.org/10.1007/978-1-60327-411-1_5.

[24] R. Staden, « A Strategy of DNA Sequencing Employing Computer Programs », *Nucleic Acids Research* 6, nº 7 (11 juin 1979): 2601-10, https://doi.org/10.1093/nar/6.7.2601.

[25] Eric S. Lander et al., « Initial Sequencing and Analysis of the Human Genome », *Nature* 409, nº 6822 (février 2001): 860-921, https://doi.org/10.1038/35057062.

[26] James L. Weber et Eugene W. Myers, « Human Whole-Genome Shotgun Sequencing », *Genome Research* 7, nº 5 (5 janvier 1997): 401-9, https://doi.org/10.1101/gr.7.5.401.

[27] Evan E. Eichler, Royden A. Clark, et Xinwei She, « An Assessment of the Sequence Gaps: Unfinished Business in a Finished Human Genome », *Nature Reviews Genetics* 5, nº 5 (mai 2004): 345-54, https://doi.org/10.1038/nrg1322.

[28] D. E. Koshland, « Sequences and Consequences of the Human Genome », *Science (New York, N.Y.)* 246, nº 4927 (13 octobre 1989): 189, https://doi.org/10.1126/science.2799380.

[29] Barton E. Slatko, Andrew F. Gardner, et Frederick M. Ausubel, « Overview of Next Generation Sequencing Technologies », *Current protocols in molecular biology* 122, nº 1 (avril 2018): e59, https://doi.org/10.1002/cpmb.59.

[30] A. D. Mirzabekov, « DNA Sequencing by Hybridization--a Megasequencing Method and a Diagnostic Tool? », *Trends in Biotechnology* 12, nº 1 (janvier 1994): 27-32, https://doi.org/10.1016/0167-7799(94)90008-6.

[31] Carl W. Fuller et al., « The Challenges of Sequencing by Synthesis », *Nature Biotechnology* 27, nº 11 (novembre 2009): 1013-23, https://doi.org/10.1038/nbt.1585.

[32] Elaine R. Mardis, « Next-Generation DNA Sequencing Methods », *Annual Review of Genomics and Human Genetics* 9 (2008): 387-402, https://doi.org/10.1146/annurev.genom.9.081307.164359.

[33] Snezana Drmanac et al., « Accurate Sequencing by Hybridization for DNA Diagnostics and Individual Genomics », *Nature Biotechnology* 16, nº 1 (janvier 1998): 54-58, https://doi.org/10.1038/nbt0198-54.

[34] S. Drmanac et al., « Elevated Baseline Triglyceride Levels Modulate Effects of HMGCoA Reductase Inhibitors on Plasma Lipoproteins », *Journal of Cardiovascular Pharmacology and Therapeutics* 6, nº 1 (1 mars 2001): 47-56, https://doi.org/10.1177/107424840100600106.

[35] P. Nyren, B. Pettersson, et M. Uhlen, « Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay », *Analytical Biochemistry* 208, nº 1 (1 janvier 1993): 171-75, https://doi.org/10.1006/abio.1993.1024.

[36] M. Ronaghi et al., « Real-Time DNA Sequencing Using Detection of Pyrophosphate Release », *Analytical Biochemistry* 242, nº 1 (1 novembre 1996): 84-89, https://doi.org/10.1006/abio.1996.0432.

[37] P. Nyrén, « Enzymatic Method for Continuous Monitoring of DNA Polymerase Activity », *Analytical Biochemistry* 167, nº 2 (décembre 1987): 235-38, https://doi.org/10.1016/0003-2697(87)90158-8.

[38] M. Ronaghi, M. Uhlén, et P. Nyrén, « A Sequencing Method Based on Real-Time Pyrophosphate », *Science (New York, N.Y.)* 281, nº 5375 (17 juillet 1998): 363, 365, https://doi.org/10.1126/science.281.5375.363.

[39] Marcel Margulies et al., « Genome Sequencing in Microfabricated High-Density Picolitre Reactors », *Nature* 437, nº 7057 (15 septembre 2005): 376-80, https://doi.org/10.1038/nature03959.

[40] D. S. Tawfik et A. D. Griffiths, « Man-Made Cell-like Compartments for Molecular Evolution », *Nature Biotechnology* 16, nº 7 (juillet 1998): 652-56, https://doi.org/10.1038/nbt0798-652.

[41] Richard Williams et al., « Amplification of Complex Gene Libraries by Emulsion PCR », *Nature Methods* 3, nº 7 (juillet 2006): 545-50, https://doi.org/10.1038/nmeth896.

[42] Wei Shao et al., « Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA », *Retrovirology* 10 (13 février 2013): 454, https://doi.org/10.1186/1742-4690-10-18.

[43] David A. Wheeler et al., « The Complete Genome of an Individual by Massively Parallel DNA Sequencing », *Nature* 452, nº 7189 (17 avril 2008): 872-76, https://doi.org/10.1038/nature06884.

[44] Magda Rybicka, Piotr Stalke, et Krzysztof Bielawski, « Current molecular methods for the detection of hepatitis B virus quasispecies », *Reviews in Medical Virology* 26 (1 août 2016), https://doi.org/10.1002/rmv.1897.

[45] B. Canard et R. S. Sarfati, « DNA Polymerase Fluorescent Substrates with Reversible 3'-Tags », *Gene* 148, nº 1 (11 octobre 1994): 1-6, https://doi.org/10.1016/0378-1119(94)90226-7.

[46] David R. Bentley et al., « Accurate Whole Human Genome Sequencing Using Reversible Terminator Chemistry », *Nature* 456, nº 7218 (6 novembre 2008): 53-59, https://doi.org/10.1038/nature07517.

[47] Gerardo Turcatti et al., « A New Class of Cleavable Fluorescent Nucleotides: Synthesis and Optimization as Reversible Terminators for DNA Sequencing by Synthesis », *Nucleic Acids Research* 36, nº 4 (mars 2008): e25, https://doi.org/10.1093/nar/gkn021.

[48] Jia Guo et al., « Four-color DNA sequencing with 3′-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides », *Proceedings of the National Academy of Sciences* 105, nº 27 (8 juillet 2008): 9145-50, https://doi.org/10.1073/pnas.0804023105.

[49] Mahdi Heydari et al., « Illumina error correction near highly repetitive DNA regions improves de novo genome assembly », *BMC Bioinformatics* 20, nº 1 (3 juin 2019): 298, https://doi.org/10.1186/s12859-019-2906-2.

[50] Juliane C. Dohm et al., « Substantial biases in ultra-short read data sets from high-throughput DNA sequencing », *Nucleic Acids Research* 36, nº 16 (1 septembre 2008): e105, https://doi.org/10.1093/nar/gkn425.

[51] Lincoln D Stein, « The case for cloud computing in genome informatics », *Genome Biology* 11, nº 5 (2010): 207, https://doi.org/10.1186/gb-2010-11-5-207.

[52] Jason A. Reuter, Damek V. Spacek, et Michael P. Snyder, « High-Throughput Sequencing Technologies », *Molecular Cell* 58, nº 4 (21 mai 2015): 586-97, https://doi.org/10.1016/j.molcel.2015.05.004.

[53] « Illumina Dye Sequencing », in *Wikipedia*, 24 novembre 2022, https://en.wikipedia.org/w/index.php?title=Illumina_dye_sequencing&oldid=1123482176.

[54] « DNA Sequencing Costs: Data », Genome.gov, consulté le 4 janvier 2023, https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

[55] Md Altaf-Ul-Amin et al., « Systems Biology in the Context of Big Data and Networks », *BioMed Research International* 2014 (27 mai 2014): e428570, https://doi.org/10.1155/2014/428570.

[56] Vivien Marx, « The Big Challenges of Big Data », *Nature* 498, nº 7453 (juin 2013): 255-60, https://doi.org/10.1038/498255a.

[57] Matthew B. Scholz, Chien-Chi Lo, et Patrick SG Chain, « Next Generation Sequencing and Bioinformatic Bottlenecks: The Current State of Metagenomic Data Analysis », *Current Opinion in Biotechnology*, Analytical biotechnology, 23, nº 1 (1 février 2012): 9-15, https://doi.org/10.1016/j.copbio.2011.11.013.

[58] Jeff Gauthier et al., « A brief history of bioinformatics », *Briefings in Bioinformatics* 20, nº 6 (27 novembre 2019): 1981-96, https://doi.org/10.1093/bib/bby063.

[59] Lonnie Welch et al., « Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies », *PLOS Computational Biology* 10, nº 3 (6 mars 2014): e1003496, https://doi.org/10.1371/journal.pcbi.1003496.

[60] Shoba Ranganathan, « Bioinformatics Education—Perspectives and Challenges », *PLOS Computational Biology* 1, nº 6 (25 novembre 2005): e52, https://doi.org/10.1371/journal.pcbi.0010052.

[61] Olena Morozova, Martin Hirst, et Marco A. Marra, « Applications of New Sequencing Technologies for Transcriptome Analysis », *Annual Review of Genomics and Human Genetics* 10, nº 1 (1 septembre 2009): 135-51, https://doi.org/10.1146/annurev-genom-082908-145957.

[62] Mark T. W. Ebbert et al., « Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight », *Genome Biology* 20, nº 1 (20 mai 2019): 97, https://doi.org/10.1186/s13059-019-1707-2.

[63] Eric E. Schadt, Steve Turner, et Andrew Kasarskis, « A Window into Third-Generation Sequencing », *Human Molecular Genetics* 19, nº R2 (15 octobre 2010): R227-240, https://doi.org/10.1093/hmg/ddq416.

[64] Ido Braslavsky et al., « Sequence Information Can Be Obtained from Single DNA Molecules », *Proceedings of the National Academy of Sciences of the United States of America* 100, nº 7 (1 avril 2003): 3960-64, https://doi.org/10.1073/pnas.0230489100.

[65] M. J. Levene et al., « Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations », *Science (New York, N.Y.)* 299, nº 5607 (31 janvier 2003): 682-86, https://doi.org/10.1126/science.1079700.

[66] John Eid et al., « Real-Time DNA Sequencing from Single Polymerase Molecules », *Science (New York, N.Y.)* 323, nº 5910 (2 janvier 2009): 133-38, https://doi.org/10.1126/science.1162986.

[67] Benjamin A. Flusberg et al., « Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing », *Nature Methods* 7, nº 6 (juin 2010): 461-65, https://doi.org/10.1038/nmeth.1459.

[68] Gang Fang et al., « Genome-Wide Mapping of Methylated Adenine Residues in Pathogenic Escherichia Coli Using Single-Molecule Real-Time Sequencing », *Nature Biotechnology* 30, nº 12 (décembre 2012): 1232-39, https://doi.org/10.1038/nbt.2432.

[69] « Séquençage à "longues lectures SMRT" », France Génomique, consulté le 4 janvier 2023, https://www.france-genomique.org/expertises-technologiques/genome-entier/sequencage-a-longues-lectures-smrt/.

[70] Erwin L. van Dijk et al., « Ten Years of Next-Generation Sequencing Technology », *Trends in Genetics: TIG* 30, nº 9 (septembre 2014): 418-26, https://doi.org/10.1016/j.tig.2014.07.001.

[71] Kin Fai Au et al., « Improving PacBio Long Read Accuracy by Short Read Alignment », *PLOS ONE* 7, nº 10 (4 octobre 2012): e46679, https://doi.org/10.1371/journal.pone.0046679.

[72] Dandan Lang et al., « Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore », *GigaScience* 9, nº 12 (30 novembre 2020): giaa123, https://doi.org/10.1093/gigascience/giaa123.

[73] Peter A. Larsen, Amy M. Heilman, et Anne D. Yoder, « The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms », *BMC Genomics* 15, nº 1 (26 août 2014): 720, https://doi.org/10.1186/1471-2164-15-720.

[74] « How HiFi Sequencing Works », PacBio, consulté le 29 juillet 2022, https://www.pacb.com/technology/hifi-sequencing/how-it-works/.

[75] Miten Jain et al., « The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community », *Genome Biology* 17, nº 1 (25 novembre 2016): 239, https://doi.org/10.1186/s13059-016-1103-0.

[76] Farzin Haque et al., « Solid-State and Biological Nanopore for Real-Time Sensing of Single Chemical and Sequencing of DNA », *Nano Today* 8, nº 1 (février 2013): 56-74, https://doi.org/10.1016/j.nantod.2012.12.008.

[77] David Deamer, Mark Akeson, et Daniel Branton, « Three Decades of Nanopore Sequencing », *Nature Biotechnology* 34, nº 5 (6 mai 2016): 518-24, https://doi.org/10.1038/nbt.3423.

[78] Ryan R. Wick, Louise M. Judd, et Kathryn E. Holt, « Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing », *Genome Biology* 20, nº 1 (24 juin 2019): 129, https://doi.org/10.1186/s13059-019-1727-y.

[79] Mikhail A. Pyatnitskiy et al., « Oxford Nanopore MinION Direct RNA-Seq for Systems Biology », *Biology* 10, nº 11 (novembre 2021): 1131, https://doi.org/10.3390/biology10111131.

[80] « What Is Oxford Nanopore Technology (ONT) Sequencing? », @yourgenome · Science website, consulté le 4 janvier 2023, https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing/.

[81] « Oxford Nanopore Integrates “Remora”: A Tool to Enable Real-Time, High-Accuracy Epigenetic Insights with Nanopore Sequencing Software MinKNOW », Oxford Nanopore Technologies, consulté le 29 juillet 2022, http://nanoporetech.com/about-us/news/oxford-nanopore-integrates-remora-tool-enable-real-time-high-accuracy-epigenetic.

[82] Huanle Liu, Oguzhan Begik, et Eva Maria Novoa, « EpiNano: Detection of M6A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing », in *RNA Modifications: Methods and Protocols*, éd. par Mary McMahon, Methods in Molecular Biology (New York, NY: Springer US, 2021), 31-52, https://doi.org/10.1007/978-1-0716-1374-0_3.

[83] Mattia Furlan et al., « Computational methods for RNA modification detection from nanopore direct RNA sequencing data », *RNA Biology* 18, nº sup1 (15 octobre 2021): 31-40, https://doi.org/10.1080/15476286.2021.1978215.

[84] « London Calling 2022: Update from Oxford Nanopore Technologies », Oxford Nanopore Technologies, 20 mai 2022, http://nanoporetech.com/resource-centre/london-calling-2022-update-oxford-nanopore-technologies.

[85] T. Laver et al., « Assessing the Performance of the Oxford Nanopore Technologies MinION », *Biomolecular Detection and Quantification* 3 (1 mars 2015): 1-8, https://doi.org/10.1016/j.bdq.2015.02.001.

[86] Clara Delahaye et Jacques Nicolas, « Sequencing DNA with Nanopores: Troubles and Biases », *PloS One* 16, nº 10 (2021): e0257521, https://doi.org/10.1371/journal.pone.0257521.

[87] Lee J Kerkhof, « Is Oxford Nanopore sequencing ready for analyzing complex microbiomes? », *FEMS Microbiology Ecology* 97, nº 3 (1 mars 2021): fiab001, https://doi.org/10.1093/femsec/fiab001.

[88] Franka J. Rang, Wigard P. Kloosterman, et Jeroen de Ridder, « From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy », *Genome Biology* 19, nº 1 (13 juillet 2018): 90, https://doi.org/10.1186/s13059-018-1462-9.

[89] Alexander S. Mikheyev et Mandy M. Y. Tin, « A First Look at the Oxford Nanopore MinION Sequencer », *Molecular Ecology Resources* 14, nº 6 (2014): 1097-1102, https://doi.org/10.1111/1755-0998.12324.

[90] Philip M. Ashton et al., « MinION Nanopore Sequencing Identifies the Position and Structure of a Bacterial Antibiotic Resistance Island », *Nature Biotechnology* 33, nº 3 (mars 2015): 296-300, https://doi.org/10.1038/nbt.3103.

[91] « At NCM, Announcements Include Single-Read Accuracy of 99.1% on New Chemistry and Sequencing a Record 10 Tb in a Single PromethION Run », Oxford Nanopore Technologies, consulté le 29 juillet 2022, http://nanoporetech.com/about-us/news/ncm-announcements-include-single-read-accuracy-991-new-chemistry-and-sequencing.

[92] Kathryn Dumschott et al., « Oxford Nanopore sequencing: new opportunities for plant genomics? », *Journal of Experimental Botany* 71, nº 18 (19 septembre 2020): 5313-22, https://doi.org/10.1093/jxb/eraa263.

[93] Eun Jung Koh et Seung Yong Hwang, « Multi-Omics Approaches for Understanding Environmental Exposure and Human Health », *Molecular & Cellular Toxicology* 15, nº 1 (1 janvier 2019): 1-7, https://doi.org/10.1007/s13273-019-0001-4.

[94] Gerd P. Pfeifer, « Environmental Exposures and Mutational Patterns of Cancer Genomes », *Genome Medicine* 2, nº 8 (16 août 2010): 54, https://doi.org/10.1186/gm175.

[95] Serena Nik-Zainal et al., « The genome as a record of environmental exposure », *Mutagenesis* 30, nº 6 (1 novembre 2015): 763-70, https://doi.org/10.1093/mutage/gev073.

[96] Carmen J. Marsit, « Influence of environmental exposure on human epigenetic regulation », éd. par Hans H. Hoppeler, *Journal of Experimental Biology* 218, nº 1 (1 janvier 2015): 71-79, https://doi.org/10.1242/jeb.106971.

[97] Lianguo Chen et al., « High-Throughput Transcriptome Sequencing Reveals the Combined Effects of Key e-Waste Contaminants, Decabromodiphenyl Ether (BDE-209) and Lead, in Zebrafish Larvae », *Environmental Pollution* 214 (1 juillet 2016): 324-33, https://doi.org/10.1016/j.envpol.2016.04.040.

[98] Himangi Srivastava et al., « Protein Prediction Models Support Widespread Post-Transcriptional Regulation of Protein Abundance by Interacting Partners », *PLOS Computational Biology* 18, nº 11 (10 novembre 2022): e1010702, https://doi.org/10.1371/journal.pcbi.1010702.

[99] Denis Mariano-Goulart, « Physical properties of ionizing radiations » (France, 2017).

[100] New York United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) NY (United States), *Sources and effects of ionizing radiation UNSCEAR 1996 report to the General Assembly, with scientific annex* (Austria: UN, 1996).

[101] Charles A. Waldren, « Classical Radiation Biology Dogma, Bystander Effects and Paradigm Shifts », *Human & Experimental Toxicology* 23, nº 2 (février 2004): 95-100, https://doi.org/10.1191/0960327104ht425oa.

[102] Julie A. Reisz et al., « Effects of Ionizing Radiation on Biological Molecules—Mechanisms of Damage and Emerging Methods of Detection », *Antioxidants & Redox Signaling* 21, nº 2 (10 juillet 2014): 260-92, https://doi.org/10.1089/ars.2013.5489.

[103] Jean Cadet, Thierry Douki, et Jean-Luc Ravanat, « Oxidatively Generated Base Damage to Cellular DNA », *Free Radical Biology & Medicine* 49, nº 1 (1 juillet 2010): 9-21, https://doi.org/10.1016/j.freeradbiomed.2010.03.025.

[104] E. S. Kempner, « Damage to Proteins Due to the Direct Action of Ionizing Radiation », *Quarterly Reviews of Biophysics* 26, nº 1 (février 1993): 27-48, https://doi.org/10.1017/S0033583500003954.

[105] Sophie Le Caër, « Water Radiolysis: Influence of Oxide Surfaces on H2 Production under Ionizing Radiation », *Water* 3, nº 1 (mars 2011): 235-53, https://doi.org/10.3390/w3010235.

[106] Dhriti Kapoor et Mahendra P. Singh, « Nanoparticles: Sources and Toxicity », in *Plant Responses to Nanomaterials: Recent Interventions, and Physiological and Biochemical Responses*, éd. par Vijay Pratap Singh et al., Nanotechnology in the Life Sciences (Cham: Springer International Publishing, 2021), 217-32, https://doi.org/10.1007/978-3-030-36740-4_9.

[107] Carlos A. Silvera Batista, Ronald G. Larson, et Nicholas A. Kotov, « Nonadditivity of Nanoparticle Interactions », *Science (New York, N.Y.)* 350, nº 6257 (9 octobre 2015): 1242477, https://doi.org/10.1126/science.1242477.

[108] Laura Rubio et al., « Biological Effects, Including Oxidative Stress and Genotoxic Damage, of Polystyrene Nanoparticles in Different Human Hematopoietic Cell Lines », *Journal of Hazardous Materials* 398 (5 novembre 2020): 122900, https://doi.org/10.1016/j.jhazmat.2020.122900.

[109] Rui Chen et al., « Endoplasmic Reticulum Stress Induced by Zinc Oxide Nanoparticles Is an Earlier Biomarker for Nanotoxicological Evaluation », *ACS Nano* 8, nº 3 (25 mars 2014): 2562-74, https://doi.org/10.1021/nn406184r.

[110] Gevdeep Bhabra et al., « Nanoparticles Can Cause DNA Damage across a Cellular Barrier », *Nature Nanotechnology* 4, nº 12 (décembre 2009): 876-83, https://doi.org/10.1038/nnano.2009.313.

[111] Meenal Gupta et al., « Genes Affecting Ionizing Radiation Survival Identified through Combined Exome Sequencing and Functional Screening », *Human Mutation* 42, nº 9 (2021): 1124-38, https://doi.org/10.1002/humu.24241.

[112] William F. Morgan, « Non-targeted and Delayed Effects of Exposure to Ionizing Radiation: I. Radiation-Induced Genomic Instability and Bystander Effects In Vitro », *Radiation Research* 159, nº 5 (1 mai 2003): 567-80, https://doi.org/10.1667/0033-7587(2003)159[0567:NADEOE]2.0.CO;2.

[113] M. K. Sarmast et al., « Silver Nanoparticles Affect ACS Expression in Tecomella Undulata in Vitro Culture », *Plant Cell, Tissue and Organ Culture (PCTOC)* 121, nº 1 (1 avril 2015): 227-36, https://doi.org/10.1007/s11240-014-0697-8.

[114] Lufeng Li et al., « Deep Sequencing Analysis of the Kineococcus Radiotolerans Transcriptome in Response to Ionizing Radiation », *Microbiological Research* 170 (1 janvier 2015): 248-54, https://doi.org/10.1016/j.micres.2014.10.003.

[115] Yeo Jin Kim et al., « Assessment of in vivo genotoxicity of citrated-coated silver nanoparticles via transcriptomic analysis of rabbit liver tissue », *International Journal of Nanomedicine* 14 (8 janvier 2019): 393-405, https://doi.org/10.2147/IJN.S174515.

[116] Adeline Buisset-Goussen et al., « Effects of Chronic Gamma Irradiation: A Multigenerational Study Using Caenorhabditis Elegans », *Journal of Environmental Radioactivity* 137 (novembre 2014): 190-97, https://doi.org/10.1016/j.jenvrad.2014.07.014.

[117] S. Gerardi, « A Comparative Review of Charged Particle Microbeam Facilities », *Radiation Protection Dosimetry* 122, nº 1-4 (2006): 285-91, https://doi.org/10.1093/rpd/ncl444.

[118] Laurence Tartier et al., « Cytoplasmic Irradiation Induces Mitochondrial-Dependent 53BP1 Protein Relocalization in Irradiated and Bystander Cells », *Cancer Research* 67, nº 12 (15 juin 2007): 5872-79, https://doi.org/10.1158/0008-5472.CAN-07-0188.

[119] Cintil Jose Chirayil et al., « Chapter 1 - Instrumental Techniques for the Characterization of Nanoparticles », in *Thermal and Rheological Measurement Techniques for Nanomaterials Characterization*, éd. par Sabu Thomas et al., Micro and Nano Technologies (Elsevier, 2017), 1-36, https://doi.org/10.1016/B978-0-323-46139-9.00001-3.

[120] Barberet, Ph. et al., « Development of a focused charged particle microbeam for the irradiation of individual cells », *Review of Scientific Instruments* Vol 76, nº 1 (2005), https://aip.scitation.org/doi/abs/10.1063/1.1823551.

[121] Philippe Barberet et al., « Cell Micro-Irradiation with MeV Protons Counted by an Ultra-Thin Diamond Membrane », *Applied Physics Letters* 111, nº 24 (11 décembre 2017): 243701, https://doi.org/10.1063/1.5009713.

[122] P. Barberet et H. Seznec, « Advances in Microbeam Technologies and Applications to Radiation Biology », *Radiation Protection Dosimetry* 166, nº 1-4 (septembre 2015): 182-87, https://doi.org/10.1093/rpd/ncv192.

[123] Stéphane Bourret et al., « Fluorescence Time-Lapse Imaging of Single Cells Targeted with a Focused Scanning Charged-Particle Microbeam », *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 325 (15 avril 2014): 27-34, https://doi.org/10.1016/j.nimb.2014.02.004.

[124] Giovanna Muggiolu et al., « Single α-Particle Irradiation Permits Real-Time Visualization of RNF8 Accumulation at DNA Damaged Sites », *Scientific Reports* 7 (31 janvier 2017): 41764, https://doi.org/10.1038/srep41764.

[125] Dietrich W. M. Walsh et al., « Live Cell Imaging of Mitochondria Following Targeted Irradiation in Situ Reveals Rapid and Highly Localized Loss of Membrane Potential », *Scientific Reports* 7, nº 1 (25 avril 2017): 46684, https://doi.org/10.1038/srep46684.

[126] Eva Torfeh et al., « Monte-Carlo Dosimetry and Real-Time Imaging of Targeted Irradiation Consequences in 2-Cell Stage Caenorhabditis Elegans Embryo », *Scientific Reports* 9, nº 1 (22 juillet 2019): 10568, https://doi.org/10.1038/s41598-019-47122-7.

[127] Giovanna Muggiolu et al., « In Situ Detection and Single Cell Quantification of Metal Oxide Nanoparticles Using Nuclear Microprobe Analysis », *Journal of Visualized Experiments : JoVE*, nº 132 (3 février 2018): 55041, https://doi.org/10.3791/55041.

[128] Dirk P. Kroese et Reuven Y. Rubinstein, « Monte Carlo Methods », *WIREs Computational Statistics* 4, nº 1 (2012): 48-58, https://doi.org/10.1002/wics.194.

[129] Dirk P. Kroese et al., « Why the Monte Carlo Method Is so Important Today », *WIREs Computational Statistics* 6, nº 6 (2014): 386-92, https://doi.org/10.1002/wics.1314.

[130] S. Agostinelli et al., « Geant4—a Simulation Toolkit », *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506, nº 3 (1 juillet 2003): 4, https://doi.org/10.1016/S0168-9002(03)01368-8.

[131] J. Allison et al., « Geant4 developments and applications », *IEEE Transactions on Nuclear Science* 53, nº 1 (février 2006): 4, https://doi.org/10.1109/TNS.2006.869826.

[132] J. Allison et al., « Recent Developments in Geant4 », *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 835 (1 novembre 2016): 186-225, https://doi.org/10.1016/j.nima.2016.06.125.

[133] S. Incerti et al., « The geant4-dna project », *International Journal of Modeling, Simulation, and Scientific Computing* 01, nº 02 (juin 2010): 157-78, https://doi.org/10.1142/S1793962310000122.

[134] S. Incerti et al., « Comparison of GEANT4 Very Low Energy Cross Section Models with Experimental Data in Water », *Medical Physics* 37, nº 9 (2010): 4692-4708, https://doi.org/10.1118/1.3476457.

[135] M. A. Bernal 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 », *Physica Medica: European Journal of Medical Physics* 31, nº 8 (1 décembre 2015): 861-74, https://doi.org/10.1016/j.ejmp.2015.10.087.

[136] S. Incerti et al., « Geant4-DNA Example Applications for Track Structure Simulations in Liquid Water: A Report from the Geant4-DNA Project », *Medical Physics*, 14 juin 2018, https://doi.org/10.1002/mp.13048.

[137] Hoang Ngoc Tran et al., « Geant4-DNA Modeling of Water Radiolysis beyond the Microsecond: An On-Lattice Stochastic Approach », *International Journal of Molecular Sciences* 22, nº 11 (juin 2021): 4, https://doi.org/10.3390/ijms22116023.

[138] Wook-Geun Shin et al., « Geant4-DNA Simulation of the Pre-Chemical Stage of Water Radiolysis and Its Impact on Initial Radiochemical Yields », *Physica Medica* 88 (1 août 2021): 86-90, https://doi.org/10.1016/j.ejmp.2021.05.029.

[139] N. Watson, « A New Revision of the Sequence of Plasmid PBR322 », *Gene* 70, nº 2 (30 octobre 1988): 399-403, https://doi.org/10.1016/0378-1119(88)90212-0.

[140] John Gregor Sutcliffe, « Complete nucleotide sequence of the Escherichia coli plasmid pBR322. », 1979.

[141] Esther M. Lederberg et Joshua Lederberg, « Genetic Studies of Lysogenicity in Escherichia Coli », *Genetics* 38, nº 1 (janvier 1953): 51-64.

[142] Sherwood R. Casjens et Roger W. Hendrix, « Bacteriophage lambda: early pioneer and still relevant », *Virology* 0 (mai 2015): 310-30, https://doi.org/10.1016/j.virol.2015.02.010.

[143] Steven Fuller et al., « Immunogenicity of a lambda phage-based anti-cancer vaccine targeting HAAH », *Journal for Immunotherapy of Cancer* 1, nº Suppl 1 (7 novembre 2013): P210, https://doi.org/10.1186/2051-1426-1-S1-P210.

[144] Trent M. Prall et al., « Consistent ultra-long DNA sequencing with automated slow pipetting », *BMC Genomics* 22, nº 1 (12 mars 2021): 182, https://doi.org/10.1186/s12864-021-07500-w.

[145] P. G. Leiman et al., « Structure and Morphogenesis of Bacteriophage T4 », *Cellular and Molecular Life Sciences: CMLS* 60, nº 11 (novembre 2003): 2356-70, https://doi.org/10.1007/s00018-003-3072-1.

[146] Moh Lan Yap et Michael G Rossmann, « Structure and function of bacteriophage T4 », *Future Microbiolgy* 9, nº 12 (17 décembre 2014), https://doi.org/10.2217/fmb.14.91.

[147] Andreas Kuhn et Julie A. Thomas, « The Beauty of Bacteriophage T4 Research: Lindsay W. Black and the T4 Head Assembly », *Viruses* 14, nº 4 (28 mars 2022): 700, https://doi.org/10.3390/v14040700.

[148] Takuma Hayashi et al., « Biological characterization of soft tissue sarcomas », *Annals of Translational Medicine* 3, nº 22 (décembre 2015): 368, https://doi.org/10.3978/j.issn.2305-5839.2015.12.33.

[149] P. Rous, « A TRANSMISSIBLE AVIAN NEOPLASM. (SARCOMA OF THE COMMON FOWL.) », *The Journal of Experimental Medicine* 12, nº 5 (1 septembre 1910): 696-705, https://doi.org/10.1084/jem.12.5.696.

[150] Sean M. Post, « Mouse models of sarcomas: critical tools in our understanding of the pathobiology », *Clinical Sarcoma Research* 2, nº 1 (4 octobre 2012): 20, https://doi.org/10.1186/2045-3329-2-20.

[151] Fredrik Mertens, Ioannis Panagopoulos, et Nils Mandahl, « Genomic Characteristics of Soft Tissue Sarcomas », *Virchows Archiv: An International Journal of Pathology* 456, nº 2 (février 2010): 129-39, https://doi.org/10.1007/s00428-009-0736-8.

[152] P Lagarde et al., « Stable Instability of Sarcoma Cell Lines Genome Despite Intra-Tumoral Heterogeneity: A Genomic and Transcriptomic Study of Sarcoma Cell Lines », *Journal of Genetics and Genomic Research*, 2015.

[153] E. Maupas, *Modes et Formes de Reproduction Des Nématodes*, 1900.

[154] G. Osche, 'Systematik Und Phylogenie Der Gattung Rhabditis (Nematoda)', *Zoologische Jahrbucher, Abteilung Fur Systematik, Okologie Und Geographie Der Tiere* 81, no. 3 (2 April 1952): 190–280.

[155] C. elegans Sequencing Consortium, 'Genome Sequence of the Nematode C. Elegans: A Platform for Investigating Biology', *Science (New York, N.Y.)* 282, no. 5396 (11 December 1998): 2012–18, https://doi.org/10.1126/science.282.5396.2012.

[156] 'The Nobel Prize in Physiology or Medicine 2002', NobelPrize.org, accessed 15 October 2022, https://www.nobelprize.org/prizes/medicine/2002/summary/.

[157] 'The Nobel Prize in Physiology or Medicine 2006 - NobelPrize.Org', accessed 15 October 2022, https://www.nobelprize.org/prizes/medicine/2006/summary/.

[158] R. C. Cassada and R. L. Russell, 'The Dauerlarva, a Post-Embryonic Developmental Variant of the Nematode Caenorhabditis Elegans', *Developmental Biology* 46, no. 2 (October 1975): 326–42, https://doi.org/10.1016/0012-1606(75)90109-8.

[159] Siwen Zhang et al., 'Caenorhabditis Elegans as a Useful Model for Studying Aging Mutations', *Frontiers in Endocrinology* 11 (2020), https://www.frontiersin.org/articles/10.3389/fendo.2020.554994.

[160] Seung-Jae Lee and Cynthia Kenyon, 'Regulation of the Longevity Response to Temperature by Thermosensory Neurons in Caenorhabditis Elegans', *Current Biology* 19, no. 9 (12 May 2009): 715–22, https://doi.org/10.1016/j.cub.2009.03.041.

[161] J. E. Sulston et al., 'The Embryonic Cell Lineage of the Nematode Caenorhabditis Elegans', *Developmental Biology* 100, no. 1 (November 1983): 64–119, https://doi.org/10.1016/0012-1606(83)90201-4.

[162] J. E. Sulston and H. R. Horvitz, 'Post-Embryonic Cell Lineages of the Nematode, Caenorhabditis Elegans', *Developmental Biology* 56, no. 1 (March 1977): 110–56, https://doi.org/10.1016/0012-1606(77)90158-0.

[163] Marina Simon et al., « In Situ Quantification of Diverse Titanium Dioxide Nanoparticles Unveils Selective Endoplasmic Reticulum Stress-Dependent Toxicity », *Nanotoxicology* 11, nº 1 (février 2017): 134-45, https://doi.org/10.1080/17435390.2017.1278803.

[164] Marina Simon et al., « Titanium Dioxide Nanoparticles Induced Intracellular Calcium Homeostasis Modification in Primary Human Keratinocytes. Towards an in Vitro Explanation of Titanium Dioxide Nanoparticles Toxicity », *Nanotoxicology* 5, nº 2 (juin 2011): 125-39, https://doi.org/10.3109/17435390.2010.502979.

[165] Lawrence A. Loeb et Raymond J. Monnat, « DNA Polymerases and Human Disease », *Nature Reviews. Genetics* 9, nº 8 (août 2008): 594-604, https://doi.org/10.1038/nrg2345.

[166] Sabrina L. Andersen et Jeff Sekelsky, « Meiotic versus Mitotic Recombination: Two Different Routes for Double-Strand Break Repair », *BioEssays : news and reviews in molecular, cellular and developmental biology* 32, nº 12 (décembre 2010): 1058-66, https://doi.org/10.1002/bies.201000087.

[167] Cadet, Douki, et Ravanat, « Oxidatively Generated Base Damage to Cellular DNA ».

[168] Julien Vignard, Gladys Mirey, et Bernard Salles, « Ionizing-Radiation Induced DNA Double-Strand Breaks: A Direct and Indirect Lighting Up », *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 108, nº 3 (septembre 2013): 362-69, https://doi.org/10.1016/j.radonc.2013.06.013.

[169] P. D. Lawley, « Effects of Some Chemical Mutagens and Carcinogens on Nucleic Acids », *Progress in Nucleic Acid Research and Molecular Biology* 5 (1966): 89-131, https://doi.org/10.1016/s0079-6603(08)60232-9.

[170] Alberto Ciccia et Stephen J. Elledge, « The DNA Damage Response: Making it safe to play with knives », *Molecular cell* 40, nº 2 (22 octobre 2010): 179-204, https://doi.org/10.1016/j.molcel.2010.09.019.

[171] Li Lan et al., « In Situ Analysis of Repair Processes for Oxidative DNA Damage in Mammalian Cells », *Proceedings of the National Academy of Sciences of the United States of America* 101, nº 38 (21 septembre 2004): 13738-43, https://doi.org/10.1073/pnas.0406048101.

[172] Ciccia et Elledge, « The DNA Damage Response ».

[173] Giovanna MUGGIOLU, « Deciphering the biological effects of ionizing radiations using charged particle microbeam: from molecular mechanisms to perspectives in emerging cancer therapies » (Université de Bordeaux, 2017).

[174] S. Nowsheen et E.S. Yang, « THE INTERSECTION BETWEEN DNA DAMAGE RESPONSE AND CELL DEATH PATHWAYS », *Experimental oncology* 34, nº 3 (octobre 2012): 243-54.

[175] Rinne De Bont et Nik van Larebeke, « Endogenous DNA Damage in Humans: A Review of Quantitative Data », *Mutagenesis* 19, nº 3 (mai 2004): 169-85, https://doi.org/10.1093/mutage/geh025.

[176] Omar Desouky, Nan Ding, et Guangming Zhou, « Targeted and Non-Targeted Effects of Ionizing Radiation », *Journal of Radiation Research and Applied Sciences* 8, nº 2 (1 avril 2015): 247-54, https://doi.org/10.1016/j.jrras.2015.03.003.

[177] Hatim Fakir et al., « Clusters of DNA Double-Strand Breaks Induced by Different Doses of Nitrogen Ions for Various LETs: Experimental Measurements and Theoretical Analyses », *Radiation Research* 166, nº 6 (1 décembre 2006): 917-27, https://doi.org/10.1667/RR0639.1.

[178] Megumi Hada et Alexandros G. Georgakilas, « Formation of Clustered DNA Damage after High-LET Irradiation: A Review », *Journal of Radiation Research* advpub (2008): 0804090035-0804090035, https://doi.org/10.1269/jrr.07123.

[179] Harald Paganetti, « Proton Relative Biological Effectiveness – Uncertainties and Opportunities », *International Journal of Particle Therapy* 5, nº 1 (21 septembre 2018): 2-14, https://doi.org/10.14338/IJPT-18-00011.1.

[180] Yoshitaka Matsumoto et al., « Enhanced Radiobiological Effects at the Distal End of a Clinical Proton Beam: In Vitro Study », *Journal of Radiation Research* 55, nº 4 (juillet 2014): 816-22, https://doi.org/10.1093/jrr/rrt230.

[181] « Ionizing Radiation - Control and Prevention | Occupational Safety and Health Administration », consulté le 2 septembre 2022, https://www.osha.gov/ionizing-radiation/control-prevention.

[182] Edward J. Calabrese, « The Linear No-Threshold (LNT) Dose Response Model: A Comprehensive Assessment of Its Historical and Scientific Foundations », *Chemico-Biological Interactions* 301 (1 mars 2019): 6-25, https://doi.org/10.1016/j.cbi.2018.11.020.

[183] Maurice Tubiana et al., « The Linear No-Threshold Relationship Is Inconsistent with Radiation Biologic and Experimental Data », *Radiology* 251, nº 1 (avril 2009): 13-22, https://doi.org/10.1148/radiol.2511080671.

[184] « Les effets des faibles doses de rayonnements ionisants », consulté le 27 janvier 2023, https://www.irsn.fr/FR/Larecherche/Organisation/Collaborations/Melodi/Pages/Melodi-effets-faibles-doses-rayonnements-ionisants.aspx#.Y9OcFq2ZOMo.

[185] Maurice Tubiana et André AURENGO, « La relation dose-effet et l'estimation des effets cancérogènes des faibles doses de rayonnements ionisants » (Académie des sciences, 2005), https://www.academie-sciences.fr/archivage_site/activite/rapport/rapport070405.pdf.

[186] William R. Holley et Aloke Chatterjee, « Clusters of DNA Damage Induced by Ionizing Radiation: Formation of Short DNA Fragments. I. Theoretical Modeling », *Radiation Research* 145, nº 2 (1 février 1996): 188-99, https://doi.org/10.2307/3579174.

[187] Björn Rydberg, « Clusters of DNA Damage Induced by Ionizing Radiation: Formation of Short DNA Fragments. II. Experimental Detection », *Radiation Research* 145, nº 2 (1 février 1996): 200-209, https://doi.org/10.2307/3579175.

[188] Maria P. Souli et al., « Clustered DNA Damage Patterns after Proton Therapy Beam Irradiation Using Plasmid DNA », *International Journal of Molecular Sciences* 23, nº 24 (janvier 2022): 15606, https://doi.org/10.3390/ijms232415606.

[189] Eugene Surdutovich et Andrey V. Solov'yov, « Multiscale Approach to the Physics of Radiation Damage with Ions », *The European Physical Journal D* 68, nº 11 (25 novembre 2014): 353, https://doi.org/10.1140/epjd/e2014-50004-0.

[190] Kentaro Baba et al., « A simulation-based study on water radiolysis species for 1H+, 4He2+, and 12C6+ ion beams with multiple ionization using Geant4-DNA », *Journal of Applied Physics* 129, nº 24 (28 juin 2021): 244702, https://doi.org/10.1063/5.0054665.

[191] Dylan Peukert et al., « Validation and Investigation of Reactive Species Yields of Geant4-DNA Chemistry Models », *Medical Physics* 46, nº 2 (2019): 983-98, https://doi.org/10.1002/mp.13332.

[192] Konstantinos P. Chatzipapas et al., « Simulation of DNA damage using Geant4-DNA: an overview of the "moleculardna" example application » (arXiv, 6 octobre 2022), https://doi.org/10.48550/arXiv.2210.01564.

[193] Thomas P. A. Devasagayam et al., « Singlet Oxygen Induced Single-Strand Breaks in Plasmid PBR322 DNA: The Enhancing Effect of Thiols », *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1088, nº 3 (26 mars 1991): 409-12, https://doi.org/10.1016/0167-4781(91)90133-7.

[194] S. K. Sahu et al., « The Effects of Indium-111 Decay on pBR322 DNA », *Radiation Research* 141, nº 2 (1 février 1995): 193-98, https://doi.org/10.2307/3579047.

[195] Simone Schröder et al., « Quantitative Gel Electrophoresis: Sources of Variation », *Journal of Proteome Research* 7, nº 3 (1 mars 2008): 1226-34, https://doi.org/10.1021/pr700589s.

[196] B. W. Birren et al., « Pulsed Field Gel Electrophoresis Techniques for Separating 1- to 50-Kilobase DNA Fragments », *Analytical Biochemistry* 177, nº 2 (mars 1989): 282-86, https://doi.org/10.1016/0003-2697(89)90052-3.

[197] Heng Li, « New strategies to improve minimap2 alignment accuracy », *Bioinformatics* 37, nº 23 (1 décembre 2021): 4572-74, https://doi.org/10.1093/bioinformatics/btab705.

[198] Petr Danecek et al., « Twelve years of SAMtools and BCFtools », *GigaScience* 10, nº 2 (1 février 2021): giab008, https://doi.org/10.1093/gigascience/giab008.

[199] Vinod K.S. Shante et Scott Kirkpatrick, « An introduction to percolation theory », *Advances in Physics* 20, nº 85 (1 mai 1971): 325-57, https://doi.org/10.1080/00018737100101261.

[200] Nathanael Lampe et al., « Mechanistic DNA Damage Simulations in Geant4-DNA Part 1: A Parameter Study in a Simplified Geometry », *Physica Medica: PM: An International Journal Devoted to the Applications of Physics to Medicine and Biology: Official Journal of the Italian Association of Biomedical Physics (AIFB)* 48 (avril 2018): 135-45, https://doi.org/10.1016/j.ejmp.2018.02.011.

[201] S. Razin, « The Minimal Cellular Genome of Mycoplasma », *Indian Journal of Biochemistry & Biophysics* 34, nº 1-2 (1997): 124-30.

[202] Laver et al., « Assessing the Performance of the Oxford Nanopore Technologies MinION ».

[203] Mikheyev et Tin, « A First Look at the Oxford Nanopore MinION Sequencer ».

[204] Roham Razaghi, « Shedding light on the long-range interaction of the human epigenome using ultra-long nanopore sequencin » (Oxford Nanopore Technologies London Calling, London, 21 mai 2021).

[205] « From Kilobases to "Whales": A Short History of Ultra-Long Reads and High-Throughput Genome Sequencing », Oxford Nanopore Technologies, consulté le 10 décembre 2022, http://nanoporetech.com/about-us/news/blog-kilobases-whales-short-history-ultra-long-reads-and-high-throughput-genome.

[206] « Beads-Free ONT Ligation Kit Library Preparation for Ultra-Long Read Sequencing », 4 mai 2020, https://gih.uq.edu.au/research/long-read-sequencing/beads-free-ont-ligation-kit-library-preparation-ultra-long-read-sequencing.

[207] G. K. Sim et al., 'Use of a CDNA Library for Studies on Evolution and Developmental Expression of the Chorion Multigene Families', *Cell* 18, no. 4 (December 1979): 1303–16, https://doi.org/10.1016/0092-8674(79)90241-1.

[208] M. D. Adams et al., 'Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project', *Science (New York, N.Y.)* 252, no. 5013 (21 June 1991): 1651–56, https://doi.org/10.1126/science.2047873.

[209] M. A. Marra, L. Hillier, and R. H. Waterston, 'Expressed Sequence Tags--ESTablishing Bridges between Genomes', *Trends in Genetics: TIG* 14, no. 1 (January 1998): 4–7, https://doi.org/10.1016/S0168-9525(97)01355-3.

[210] Peng Liang and Arthur B Pardee, 'Recent Advances in Differential Display', *Current Opinion in Immunology* 7, no. 2 (1 April 1995): 274–80, https://doi.org/10.1016/0952-7915(95)80015-8.

[211] V. E. Velculescu et al., 'Serial Analysis of Gene Expression', *Science (New York, N.Y.)* 270, no. 5235 (20 October 1995): 484–87, https://doi.org/10.1126/science.270.5235.484.

[212] M. Schena et al., 'Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray', *Science (New York, N.Y.)* 270, no. 5235 (20 October 1995): 467–70, https://doi.org/10.1126/science.270.5235.467.

[213] N. J. Nelson, 'Microarrays Have Arrived: Gene Expression Tool Matures', *Journal of the National Cancer Institute* 93, no. 7 (4 April 2001): 492–94, https://doi.org/10.1093/jnci/93.7.492.

[214] Matthew N. Bainbridge et al., 'Analysis of the Prostate Cancer Cell Line LNCaP Transcriptome Using a Sequencing-by-Synthesis Approach', *BMC Genomics* 7 (29 September 2006): 246, https://doi.org/10.1186/1471-2164-7-246.

[215] Andreas P.M. Weber, 'Discovering New Biology through Sequencing of RNA1', *Plant Physiology* 169, no. 3 (November 2015): 1524–31, https://doi.org/10.1104/pp.15.01081.

[216] Rory Stark, Marta Grzelak, and James Hadfield, 'RNA Sequencing: The Teenage Years', *Nature Reviews Genetics* 20, no. 11 (November 2019): 631–56, https://doi.org/10.1038/s41576-019-0150-2.

[217] R. H. Brakenhoff, J. G. Schoenmakers, and N. H. Lubsen, 'Chimeric CDNA Clones: A Novel PCR Artifact', *Nucleic Acids Research* 19, no. 8 (25 April 1991): 1949, https://doi.org/10.1093/nar/19.8.1949.

[218] X. Qiu et al., 'Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S RRNA Gene-Based Cloning', *Applied and Environmental Microbiology* 67, no. 2 (February 2001): 880–87, https://doi.org/10.1128/AEM.67.2.880-887.2001.

[219] M. F. Polz and C. M. Cavanaugh, 'Bias in Template-to-Product Ratios in Multitemplate PCR', *Applied and Environmental Microbiology* 64, no. 10 (October 1998): 3724–30, https://doi.org/10.1128/AEM.64.10.3724-3730.1998.

[220] M. T. Suzuki and S. J. Giovannoni, 'Bias Caused by Template Annealing in the Amplification of Mixtures of 16S RRNA Genes by PCR', *Applied and Environmental Microbiology* 62, no. 2 (February 1996): 625–30, https://doi.org/10.1128/aem.62.2.625-630.1996.

[221] Mikhail A. Pyatnitskiy et al., 'Oxford Nanopore MinION Direct RNA-Seq for Systems Biology', *Biology* 10, no. 11 (November 2021): 1131, https://doi.org/10.3390/biology10111131.

[222] Emily M. Harcourt, Anna M. Kietrys, and Eric T. Kool, 'Chemical and Structural Effects of Base Modifications in Messenger RNA', *Nature* 541, no. 7637 (18 January 2017): 339–46, https://doi.org/10.1038/nature21351.

[223] Huanle Liu, Oguzhan Begik, and Eva Maria Novoa, 'EpiNano: Detection of M6A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing', in *RNA Modifications: Methods and Protocols*, ed. Mary McMahon, Methods in Molecular Biology (New York, NY: Springer US, 2021), 31–52, https://doi.org/10.1007/978-1-0716-1374-0_3.

[224] Ploy N. Pratanwanich et al., 'Identification of Differential RNA Modifications from Nanopore Direct RNA Sequencing with XPore', *Nature Biotechnology* 39, no. 11 (November 2021): 1394–1402, https://doi.org/10.1038/s41587-021-00949-w.

[225] Matthew T. Parker, Geoffrey J. Barton, and Gordon G. Simpson, 'Yanocomp: Robust Prediction of M6A Modifications in Individual Nanopore Direct RNA Reads' (bioRxiv, 16 June 2021), https://doi.org/10.1101/2021.06.15.448494.

[226] Mattia Furlan et al., 'Computational Methods for RNA Modification Detection from Nanopore Direct RNA Sequencing Data', *RNA Biology* 18, no. sup1 (15 October 2021): 31–40, https://doi.org/10.1080/15476286.2021.1978215.

[227] Muggiolu et al., « Single α-Particle Irradiation Permits Real-Time Visualization of RNF8 Accumulation at DNA Damaged Sites ».

[228] G. Seydoux, C. Savage, et I. Greenwald, « Isolation and Characterization of Mutations Causing Abnormal Eversion of the Vulva in Caenorhabditis Elegans », *Developmental Biology* 157, nº 2 (juin 1993): 423-36, https://doi.org/10.1006/dbio.1993.1146.

[229] Yunshun Chen et al., « edgeR: Empirical Analysis of Digital Gene Expression Data in R » (Bioconductor version: Release (3.16), 2022), https://doi.org/10.18129/B9.bioc.edgeR.

[230] Gordon Smyth et al., « limma: Linear Models for Microarray Data » (Bioconductor version: Release (3.16), 2022), https://doi.org/10.18129/B9.bioc.limma.

[231] Winston Haynes, « Bonferroni Correction », in *Encyclopedia of Systems Biology*, éd. par Werner Dubitzky et al. (New York, NY: Springer, 2013), 154-154, https://doi.org/10.1007/978-1-4419-9863-7_1213.

[232] Uku Raudvere et al., « g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) », *Nucleic Acids Research* 47, nº W1 (2 juillet 2019): W191-98, https://doi.org/10.1093/nar/gkz369.

[233] Zhu-hui Yuan et al., « Fatty Acids Metabolism: The Bridge Between Ferroptosis and Ionizing Radiation », *Frontiers in Cell and Developmental Biology* 9 (2021), https://www.frontiersin.org/articles/10.3389/fcell.2021.675617.

[234] Otilia Antal et al., « Combination of Unsaturated Fatty Acids and Ionizing Radiation on Human Glioma Cells: Cellular, Biochemical and Gene Expression Analysis », *Lipids in Health and Disease* 13, nº 1 (2 septembre 2014): 142, https://doi.org/10.1186/1476-511X-13-142.

[235] Elizabeth Dufourcq-Sekatcheff et al., « Deciphering Differential Life Stage Radioinduced Reproductive Decline in Caenorhabditis Elegans through Lipid Analysis », *International Journal of Molecular Sciences* 22, nº 19 (24 septembre 2021): 10277, https://doi.org/10.3390/ijms221910277.

[236] Charles P. Hinzman et al., « Exposure to Ionizing Radiation Causes Endoplasmic Reticulum Stress in the Mouse Hippocampus », *Radiation Research* 190, nº 5 (7 août 2018): 483-93, https://doi.org/10.1667/RR15061.1.

[237] Lei Li, Michael Story, et Randy J Legerski, « Cellular Responses to Ionizing Radiation Damage », *International Journal of Radiation Oncology*Biology*Physics* 49, nº 4 (15 mars 2001): 1157-62, https://doi.org/10.1016/S0360-3016(00)01524-8.

[238] Giana Angelo et Marc R. Van Gilst, « Starvation Protects Germline Stem Cells and Extends Reproductive Longevity in C. elegans », *Science* 326, nº 5955 (13 novembre 2009): 954-58, https://doi.org/10.1126/science.1178343.

[239] Maria Antonietta Zoroddu et al., « Toxicity of Nanoparticles », *Current Medicinal Chemistry* 21, nº 33 (2014): 3837-53, https://doi.org/10.2174/09298673216661406011623l4.

[240] Simon et al., « In Situ Quantification of Diverse Titanium Dioxide Nanoparticles Unveils Selective Endoplasmic Reticulum Stress-Dependent Toxicity ».

[241] Adrien Leger et al., « RNA Modifications Detection by Comparative Nanopore Direct RNA Sequencing », *Nature Communications* 12, nº 1 (10 décembre 2021): 7198, https://doi.org/10.1038/s41467-021-27393-3.

[242] Simpson Jt et al., « Detecting DNA Cytosine Methylation Using Nanopore Sequencing », *Nature Methods* 14, nº 4 (avril 2017), https://doi.org/10.1038/nmeth.4184.

[243] Sara Novak et al., « Internalization of consumed TiO2 nanoparticles by a model invertebrate organism », *Journal of Nanomaterials* 2012 (1 janvier 2012): 3:3, https://doi.org/10.1155/2012/658752.

[244] « What Is Bulk RNA Sequencing? », *Single Cell Discoveries* (blog), consulté le 9 octobre 2022, https://www.scdiscoveries.com/support/what-is-bulk-rna-sequencing/.

[245] J Eberwine et al., « Analysis of gene expression in single live neurons. », *Proceedings of the National Academy of Sciences* 89, nº 7 (avril 1992): 3010-14, https://doi.org/10.1073/pnas.89.7.3010.

[246] Bertrand Lambolez et al., « AMPA Receptor Subunits Expressed by Single Purkinje Cells », *Neuron* 9, nº 2 (1 août 1992): 247-58, https://doi.org/10.1016/0896-6273(92)90164-9.

[247] Sarah K. Whitley, William T. Horne, et Jay K. Kolls, « Research Techniques Made Simple: Methodology and Clinical Applications of RNA Sequencing », *The Journal of Investigative Dermatology* 136, nº 8 (août 2016): e77-82, https://doi.org/10.1016/j.jid.2016.06.003.

[248] Fuchou Tang et al., « MRNA-Seq Whole-Transcriptome Analysis of a Single Cell », *Nature Methods* 6, nº 5 (mai 2009): 377-82, https://doi.org/10.1038/nmeth.1315.

[249] Valentine Svensson, Roser Vento-Tormo, et Sarah A. Teichmann, « Exponential Scaling of Single-Cell RNA-Seq in the Past Decade », *Nature Protocols* 13, nº 4 (avril 2018): 599-604, https://doi.org/10.1038/nprot.2017.149.

[250] Saiful Islam et al., « Characterization of the Single-Cell Transcriptional Landscape by Highly Multiplex RNA-Seq », *Genome Research* 21, nº 7 (juillet 2011): 1160-67, https://doi.org/10.1101/gr.110882.110.

[251] Tamar Hashimshony et al., « CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification », *Cell Reports* 2, nº 3 (27 septembre 2012): 666-73, https://doi.org/10.1016/j.celrep.2012.08.003.

[252] Diego Adhemar Jaitin et al., « Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types », *Science (New York, N.Y.)* 343, nº 6172 (14 février 2014): 776-79, https://doi.org/10.1126/science.1247651.

[253] Daniel Ramsköld et al., « Full-Length MRNA-Seq from Single-Cell Levels of RNA and Individual Circulating Tumor Cells », *Nature Biotechnology* 30, nº 8 (août 2012): 777-82, https://doi.org/10.1038/nbt.2282.

[254] Simone Picelli et al., « Smart-Seq2 for Sensitive Full-Length Transcriptome Profiling in Single Cells », *Nature Methods* 10, nº 11 (novembre 2013): 1096-98, https://doi.org/10.1038/nmeth.2639.

[255] Sayantan Bose et al., « Scalable microfluidics for single-cell RNA printing and sequencing », *Genome Biology* 16, nº 1 (6 juin 2015): 120, https://doi.org/10.1186/s13059-015-0684-3.

[256] Junyue Cao et al., « Comprehensive single cell transcriptional profiling of a multicellular organism », *Science (New York, N.Y.)* 357, nº 6352 (18 août 2017): 661-67, https://doi.org/10.1126/science.aam8940.

[257] Allon M Klein et al., « Droplet barcoding for single cell transcriptomics applied to embryonic stem cells », *Cell* 161, nº 5 (21 mai 2015): 1187-1201, https://doi.org/10.1016/j.cell.2015.04.044.

[258] Sarah Aldridge et Sarah A. Teichmann, « Single Cell Transcriptomics Comes of Age », *Nature Communications* 11, nº 1 (27 août 2020): 4307, https://doi.org/10.1038/s41467-020-18158-5.

[259] Byungjin Hwang, Ji Hyun Lee, et Duhee Bang, « Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines », *Experimental & Molecular Medicine* 50, nº 8 (août 2018): 1-14, https://doi.org/10.1038/s12276-018-0071-8.

[260] Olivier B. Poirion et al., « Single-Cell Transcriptomics Bioinformatics and Computational Challenges », *Frontiers in Genetics* 7 (2016), https://www.frontiersin.org/articles/10.3389/fgene.2016.00163.

[261] Luwen Ning et al., « Current Challenges in the Bioinformatics of Single Cell Genomics », *Frontiers in Oncology* 4 (2014), https://www.frontiersin.org/articles/10.3389/fonc.2014.00007.

[262] Rapolas Zilionis et al., « Single-Cell Barcoding and Sequencing Using Droplet Microfluidics », *Nature Protocols* 12, nº 1 (janvier 2017): 44-73, https://doi.org/10.1038/nprot.2016.154.

[263] Zhichao Miao et al., « Putative Cell Type Discovery from Single-Cell Gene Expression Data », *Nature Methods* 17, nº 6 (juin 2020): 621-28, https://doi.org/10.1038/s41592-020-0825-9.

[264] Ramiro Lorenzo et al., « Combining Single-Cell RNA-Sequencing with a Molecular Atlas Unveils New Markers for Caenorhabditis Elegans Neuron Classes », *Nucleic Acids Research* 48, nº 13 (27 juillet 2020): 7119-34, https://doi.org/10.1093/nar/gkaa486.

[265] Jordan W. Squair et al., « Confronting False Discoveries in Single-Cell Differential Expression », *Nature Communications* 12, nº 1 (28 septembre 2021): 5692, https://doi.org/10.1038/s41467-021-25960-2.

[266] Tara Chari, Joeyta Banerjee, et Lior Pachter, « The Specious Art of Single-Cell Genomics » (bioRxiv, 27 septembre 2021), https://doi.org/10.1101/2021.08.25.457696.

[267] Michelle Ying Ya Lee, Klaus H. Kaestner, et Mingyao Li, « Benchmarking Algorithms for Joint Integration of Unpaired and Paired Single-Cell RNA-Seq and ATAC-Seq Data » (bioRxiv, 3 février 2023), https://doi.org/10.1101/2023.02.01.526609.

[268] Jonathan S. Packer et al., « A Lineage-Resolved Molecular Atlas of C. Elegans Embryogenesis at Single-Cell Resolution », *Science* 365, nº 6459 (20 septembre 2019), https://doi.org/10.1126/science.aax1971.

[269] « What is Cell Ranger? -Software -Single Cell Gene Expression -Official 10x Genomics Support », consulté le 9 octobre 2022, https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger.

[270] Etienne Becht et al., « Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP », *Nature Biotechnology* 37, nº 1 (janvier 2019): 38-44, https://doi.org/10.1038/nbt.4314.

[271] « Method of the Year 2022: Long-Read Sequencing », *Nature Methods* 20, nº 1 (janvier 2023): 1-1, https://doi.org/10.1038/s41592-022-01759-x.

Appendix: Effect of nano-sensitization on the cellular response of irradiated sarcoma lines

Introduction

The use of ionizing radiation in the treatment of cancer is a procedure that dates back to the beginning of the 20th century and has since been commonly used under the name of radiation therapy. The principle behind this technique is based on the use of the ability of ionizing radiation to produce sufficient cell damage to trigger cell death processes and thus destroy cancer cells. Radiation therapy is one of the key treatments for cancer, used for nearly 50% of patients during the course of their disease and often applied in combination with surgery or chemotherapy in a neo-adjuvant or adjuvant manner (in order to increase its positive effects)[1]. Improving its effectiveness remains an essential challenge and new strategies must emerge. Its principle is based on the use of high-energy ionizing radiation (photons, electrons) to minimize tumor proliferation. The success of radiation therapy depends on the "biological effective dose"[2] delivered to the tumor, and the higher the dose delivered, the better the control: this is the concept of "dose escalation"[3]. However, this increase in dose within the tumor tissue generally implies a similar increase in the surrounding healthy tissue, which is responsible for side effects that limit the continuation of treatment[4,5].

Current development strategies in radiation therapy therefore aim to achieve better protection of healthy tissue while increasing the dose delivered to the heart of the tumor, *i.e.* to improve the differential effect between tumor and healthy tissue. These ways explored include the study of new irradiation techniques based on ballistics and beam delivery, allowing more targeted and more precise irradiation, the use of charged particles (protons, hadrons), and the use of radio-sensitizing, radio-enhancing or radioprotective agents with respect to ionizing radiation. Among these agents, metal and metal oxide nanoparticles with a high atomic number (Z) have been proposed, because the interactions between these metal and metal oxide nanoparticles with high electron density and radiation imply the induction of physicochemical and biological mechanisms underlying these radio-sensitizing and/or radio-enhancement effects[6]. Although studies have already validated this radio-sensitization process with several types of nanoparticles and on different types of tumors[7,8], the cellular mechanisms involved in this phenomenon are still obscure, especially in the rarest types of cancers which sometimes results in difficulties in moving from experimental study to clinical application[9].

In this context, a project to study the radio-sensitization of two patient-derived sarcoma cell lines, IB106 and IB115, by TiO_2 nanoparticles in nanosheets shape (TNs) was carried out with

the IB115 line presenting characteristics of radio-resistance contrary to IB106 which will allow to better study the presence or not of a radio-sensitization.

Monte Carlo simulations were performed to ensure homogeneous dose deposition on a sarcoma cell monolayer and the effects of nanoparticles alone, multi-dose proton irradiation and the two previous factors combined were studied at the scale of cell proliferation, chemical quantification analysis and transcriptomic expression.

(A)

(B)

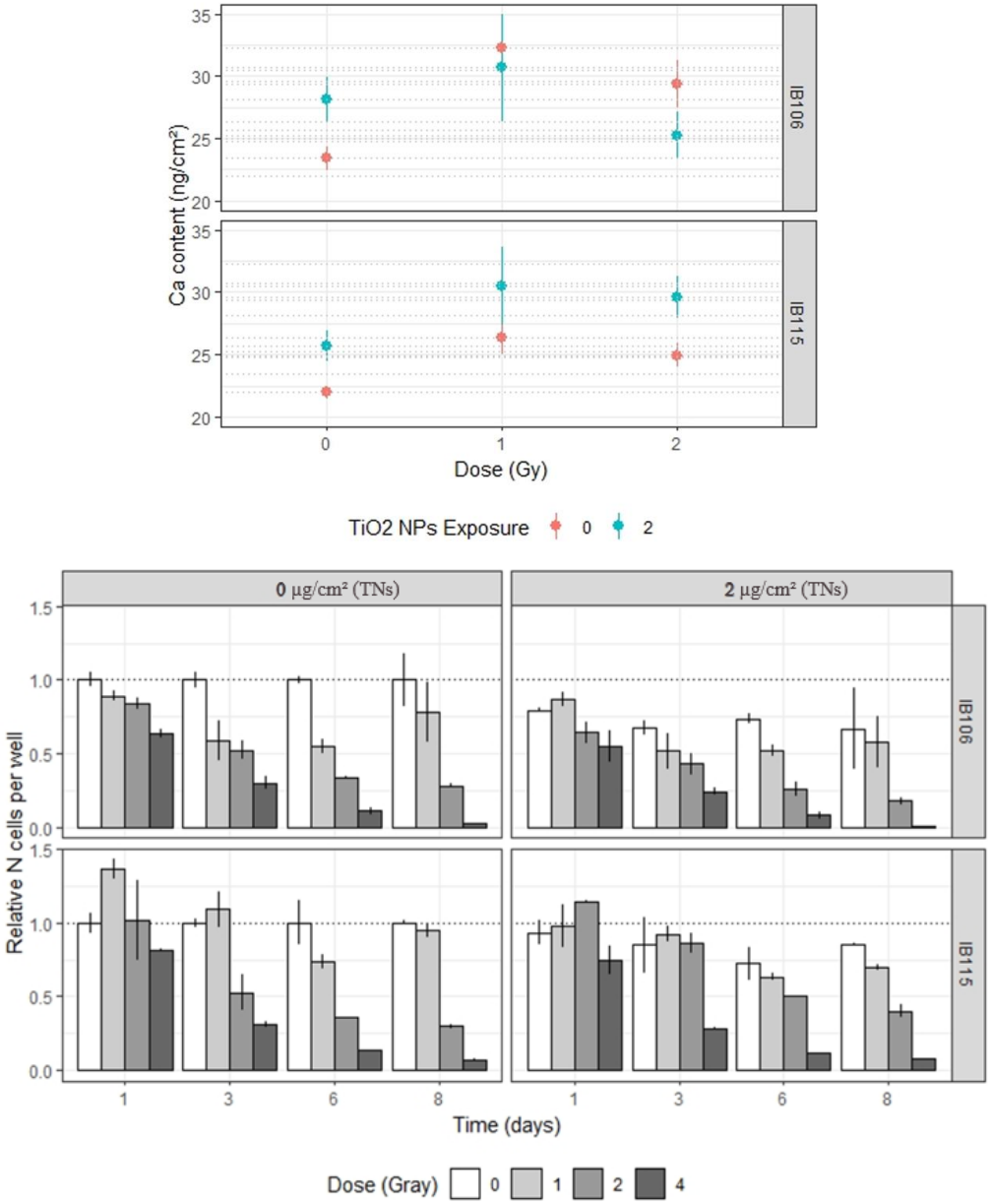


Figure 84. Effect of exposure to TNs (2µg.cm^{-2}) and ionizing radiation (1, 2 and 4 Gy, 3 MeV protons by macro-irradiation) on (A) cellular calcium concentration and (B) cell proliferation on IB106 and IB115 cells.

Previous cell proliferation and chemical quantification analyses produced have already established the impact of exposure to nanoparticles or ionizing radiation separately on the intracellular calcium content of cells and on cell survival in the 8 days following exposure.

However, the study of the exposure to both factors is complex when it comes to identifying if the resulting cellular damage is caused by only one factor or by both, and in the last case which cellular mechanisms are involved. qPCR on selected genes linked to known stress pathways was used but the technique fails to capture the entirety of the cellular response.

The contribution of the sequencing method in this project is therefore to study the cellular pathways impacted under the experimental conditions in order to identify whether radio-sensitization is caused by the addition of TiO_2 nanoparticles in the form of beads or nanosheets on irradiated sarcoma cells.

Materials and methods

1/ Sarcoma culture

Sarcoma cell lines were characterized, from a genetic point of view, in previous works from F. Chibon laboratory. Cell lines (IB115 and IB106) were maintained in RPMI 1640 GlutaMAX supplemented with Fetal Bovine Serum (10% v/v, FBS) and streptomycin/penicillin (100 µg/ml). Cells were kept in a humidified atmosphere at 37 °C and 5% (v/v) CO_2. Sarcoma cell lines were grown in defined medium at 37°C in a 5% (v/v) CO_2 humidified atmosphere and passages are realized at 80 % confluency.

2/ TNs exposure

The suspensions of TNs were prepared in ultrapure water at a concentration of 1 $mg.ml^{-1}$. TNs were dispersed by intense sonication pulses of 1 min at RT (750 W, 20 kHz, with 30 % amplitude) using a Vibra-Cell™ and a dedicated 3MM conical microprobe (ThermoFischer Scientific). Suspensions were hereby known as "stock suspensions". Stock suspensions were diluted at the appropriate concentration in defined culture medium in order to obtain an exposure suspension at 2 $\mu g.cm^{-2}$ (final concentration). Briefly, 20,000 cells were seeded in a single drop in the middle of a 6-well plates for 24 hours in appropriate culture medium, and then exposed to TNs for 16 to 24 h.

3/ Preparation for irradiation

Sarcoma cells were cultured directly onto ion beam microprobe sample holders as adapted from previous studies (Le Trequesser et al. 2014, Muggiolu et al. JoVe 2018). Briefly, cells were directly grown on 2 µm-thick polycarbonate foil for 24 hours in appropriate culture medium, and then exposed (or not) to TNs for 16 to 24 h before irradiation. Control cells were prepared similarly with no addition of TNs and no irradition exposure. 24h after irradiation sequence, cells were rinsed once in culture medium, and very briefly rinsed twice in ultrapure water to remove excess of extracellular salts from culture medium.

4/ Irradiation

Cell irradiation was carried out at the Department of Radiotherapy, Institut Bergonié (Bordeaux, France) using a Clinical Linear Accelerator (CLINAC 21EX, Varian Medical Systems) used in routine for the treatment patients. Sarcoma cell lines were irradiated using 6 MeV photon beams with 1, 2 and 4 Gy delivered with a dose-rate of 2 $Gy.min^{-1}$. These doses are selected as irradiation doses because 2 Gy is the fractionated dose used during patient treatments; 1 and 4 Gy are lower and higher doses with respect to this standard value. The photon beam was collimated in 15 x 15 cm^2 square field. Surface distance of 100 cm was applied, and the irradiations were carried out with a single beam oriented at 0° (single vertical beam). As a control, a mock sample was used, which was treated in the same way with the exception of TNs exposure and/or irradiation. The cell monolayer was covered by 10 mm of medium to achieve electronic equilibrium and reach 93% of the maximum dose. One hour before irradiation 9.6 ml of growth medium were added to achieve this depth.

5/ Sequencing

Whole-transcriptome cDNA libraries were first constructed from extracted mRNA using a PCR-cDNA barcoding kit (SQK-PCB109; Oxford Nanopore Technologies) following the standard associated protocol. Three libraries, each containing barcoded cDNAs from all the studied experimental conditions, were produced from different biological samples. The libraries then were sequenced on a Mk1C MinION using R9.4.1 flow cells with a min_qscore of 7 and live basecalling until the flow cell runs out of active pores.

6/ Bioinformatic analysis

The fastq files were merged and mapped to the GrCh38 human reference transcriptome using minimap2 with the option –ax map-ont and the alignment files were processed using samtools. Alignment results were converted into an expression matrix with an associated metadata table using a custom Python script. The differential expression analysis was then performed in R using the edgeR and limma libraries. The expression matrix was inserted in a DGEList object (edgeR package). Genes with a <1 CPM (counts per million) were removed and gene counts were then normalized to log2-CPM (functions calcNormFactors and voom). A linear model was fitted for each gene (function lmFit) and contrasts between experimental conditions were

extracted (functions makeContrasts and constrasts.fit). The log odds of differential expression for each gene was then determined using an empirical Bayes test (function eBayes) and pvalues adjusted using the Bonferroni method (function p.adjust) and differentially expressed genes obtained (function decideTests). Enrichment analysis was then performed using the gprofiler g:GOSt functional profiling method with default settings. The codes used are available at:

https://github.com/pelotbdr/iribio_scripts/tree/main/bulk_transcriptome_analysis

Experimental results

The two sarcoma lines studied, IB106 and IB115, were exposed to various experimental conditions in order to study the transcriptomic response to irradiation, to TiO_2 nanoparticles in the form of nanosheets (TNs) and to both factors simultaneously. The conditions were as follows: TNs, 1Gy, 1Gy+TNs, 2Gy, 2Gy+TNs. Exposure to nanoparticles under the relevant conditions was always at 2µg/cm^{-2}, transcriptomes were extracted 24h post-exposure/irradiation and each condition was performed in biological triplicate for each line. The extracted transcriptomes were then prepared for sequencing using an Oxford Nanopore barcoding kit to sequence multiple conditions on a single chip. The three triplicates were performed on different chips.

1. Differential expression analysis

We started by visualizing all experimental conditions on PCA (Principal Component Analysis), plots for each cell line. This method does not provide precise information on transcriptomic expression profiles but it allows to get an overview of the global distribution of conditions and to see if any initial trends emerge. We also extracted the positions of the points on these plots and calculated the Euclidean distance in order to represent them in the form of a clustered heatmap (clustermap) to better observe the distances between conditions.

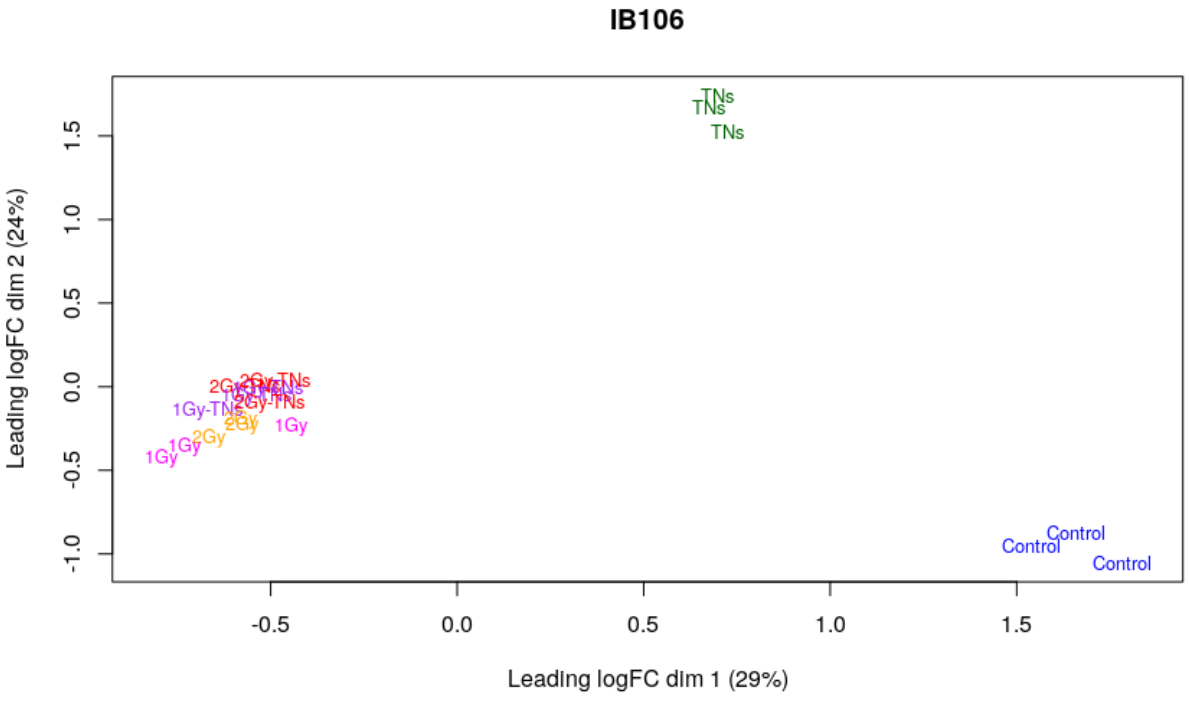


Figure 85. PCA of libraries of all experimental conditions in IB106 cell line using all expressed genes.

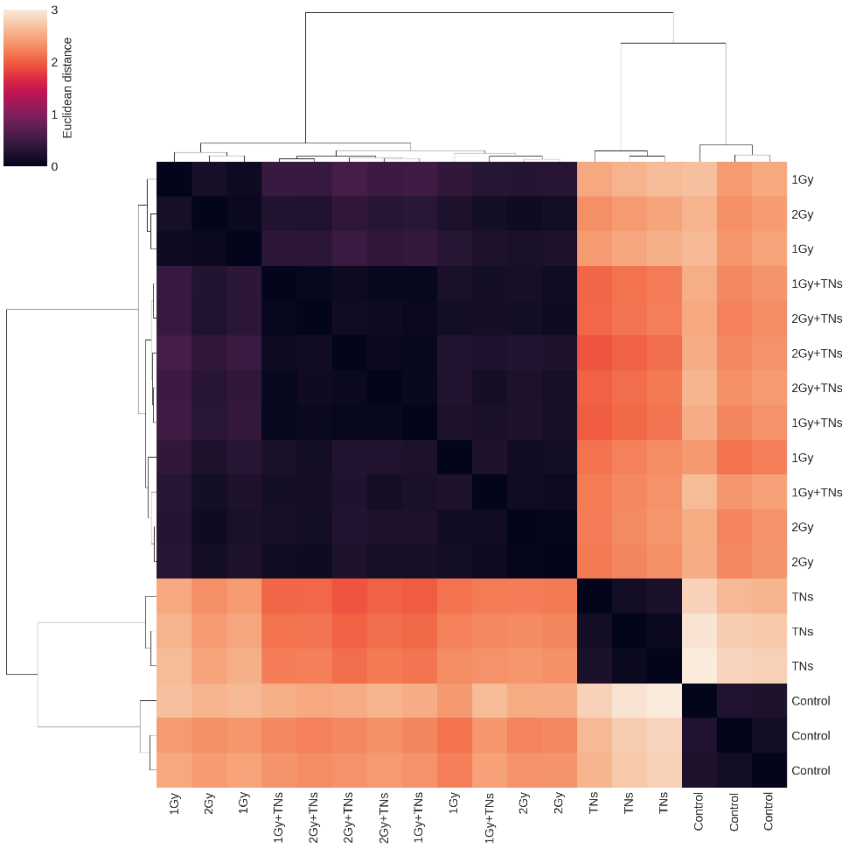


Figure 86. Clustermap of Euclidean distances extracted from the PCA plot.

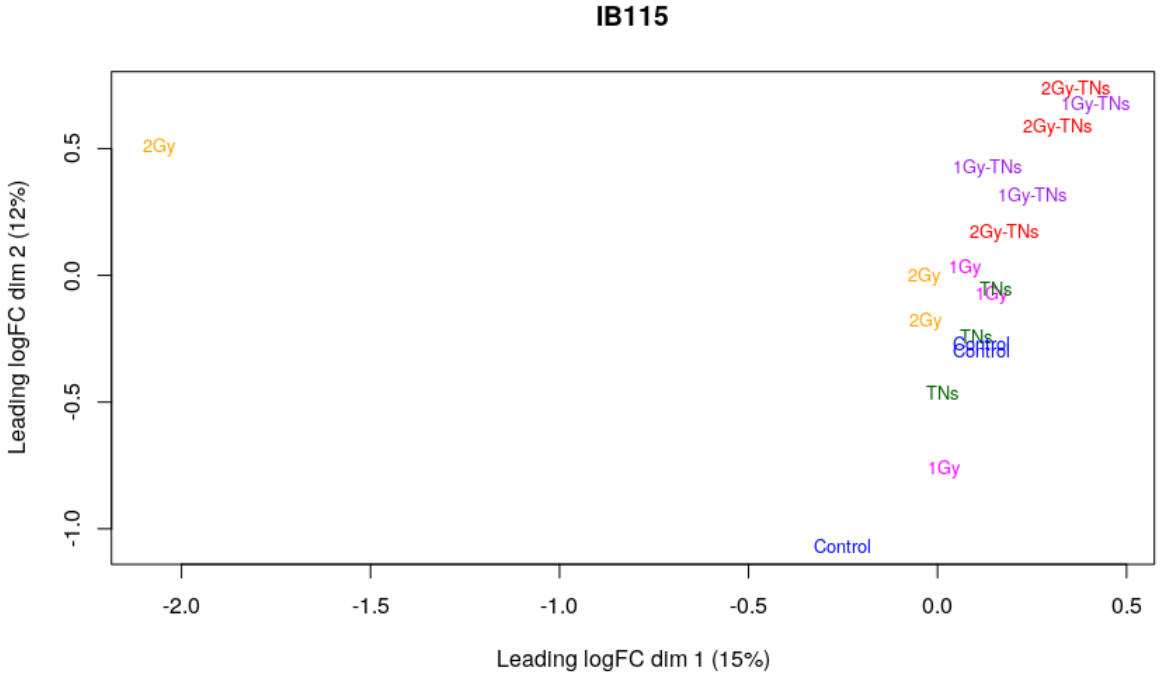


Figure 87. PCA of libraries of all experimental conditions in IB106 cell line using all expressed genes.

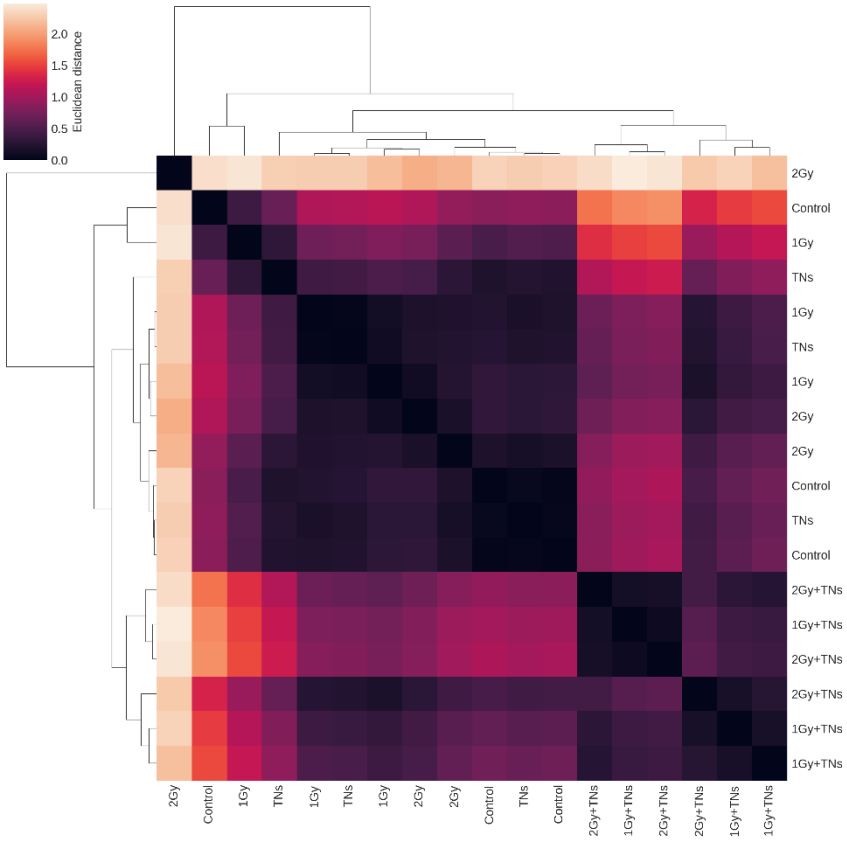


Figure 88. Clustermap of Euclidean distances extracted from the PCA plot.

The IB106 cell line shows a distinct grouping of conditions in 3 groups: controls, TNs exposure only and irradiated (Figures 85-86). Among the irradiated conditions, we can also observe a clustering, although quite weak, between the irradiation+TNs conditions on one side and irradiation only on the other.

On the other hand, no similar phenomenon is observed in line IB115 and no distinct group is discernible (Figures 87-88). One of the 2Gy replicates appears to be separate from the other conditions, which could be due to a problem in RNA extraction or library preparation, and the irradiation+TNs conditions seem to cluster roughly together but apart from that no clear trend is apparent. The low value of the eigenvalues for these PC1 and PC2 (15 and 12% respectively) also indicates the difficulty for PCA to find effective variance factors to distinguish the conditions under study.

From these plots, we can thus observe a non-negligible difference between the two cell lines in terms of dissimilarity between conditions. The IB106 cell line exhibits two types of response

characteristic of the type of stress undergone, response to irradiation or response to nanoparticles, and the latter seems to be overshadowed by the former when irradiation takes place. For the IB115 line, on the other hand, no specific response emerged overall, indicating a possible resistance to these types of stress.

To further explore the difference in response between the different conditions, we then performed a differential expression analysis on all conditions compared to the control to compute differentially expressed (DE) genes.

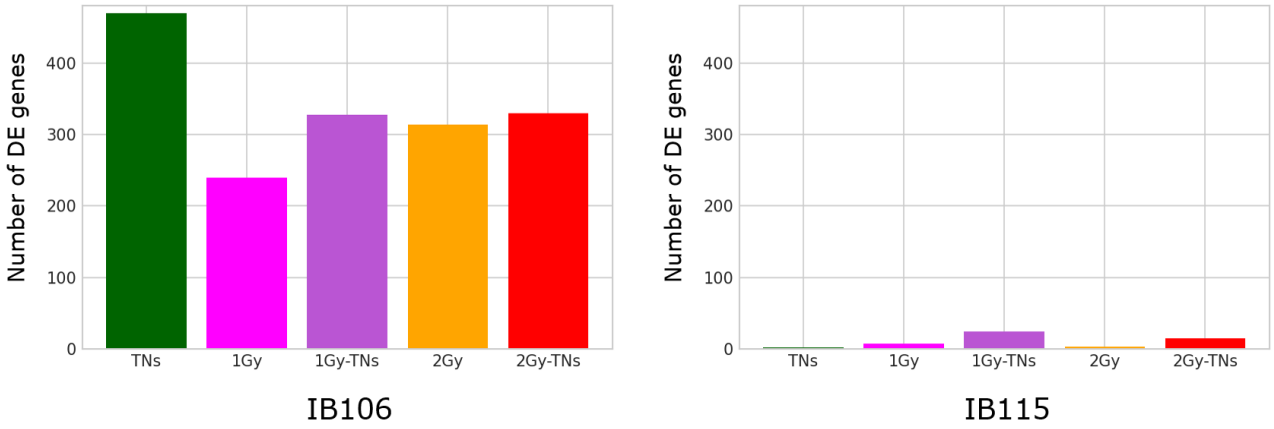


Figure 89. Number of DE genes per condition per cell line.

The difference in the number of DE genes obtained seems to be in the same direction as the results observed previously on the PCA, with line IB106 in which a transcriptomic response clearly appears and line IB115 in which this phenomenon does not appear (Figure 89). We can note that in line IB106, the TNs condition is the one in which the most DE genes are found. Although we cannot quantify the level of response to the number of DE genes, it is interesting to note that the irradiation+TNs response results in fewer DE genes than TNs alone, which again points to a cellular response to irradiation that overshadows that to nanoparticles. For the IB115 line, although the number of DE genes is low, it can be observed that this number is slightly higher for the irradiation+TNs conditions confirming that these conditions are slightly different compared to the controls.

2. Comparison of DE genes between conditions

We also looked at the intersections of DE genes between conditions, to see if the impacted genes are broadly the same in all conditions or if specificities are observable, in UpSet plots.

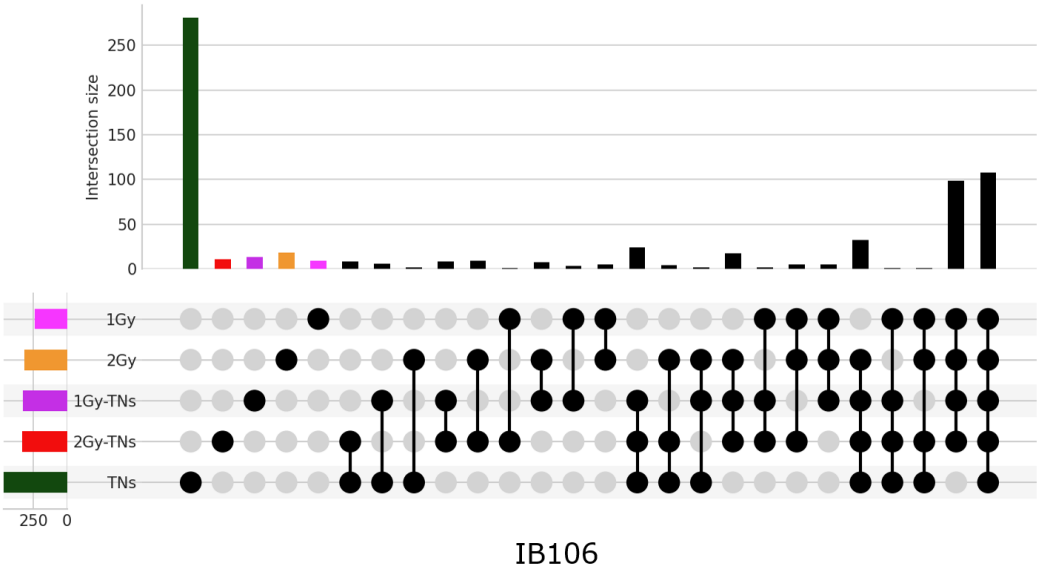


Figure 90. UpSet plot of DE genes intersection between experimental conditions in IB106 cell line.

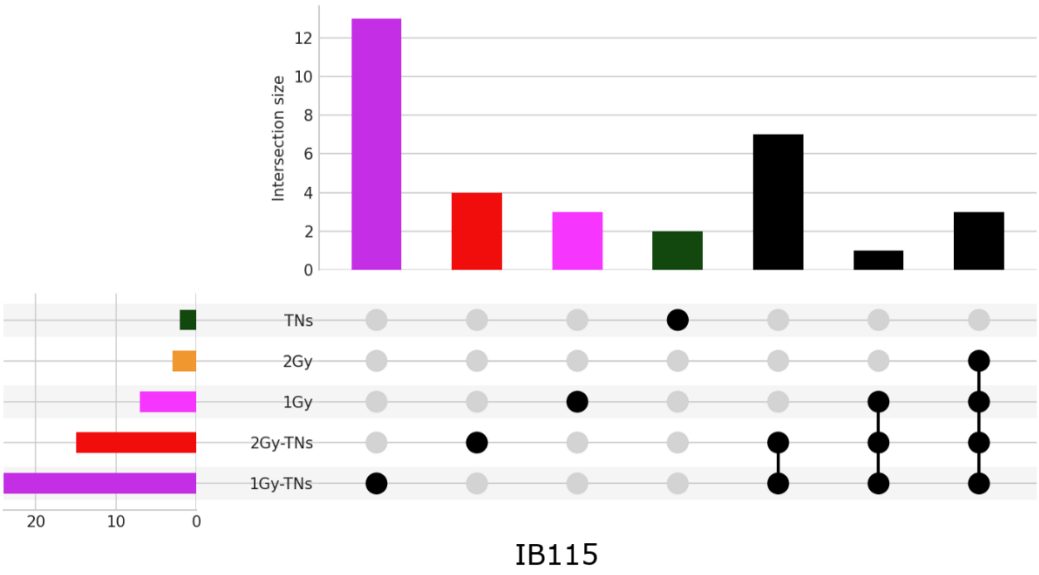


Figure 91. UpSet plot of DE genes intersection between experimental condition in IB115 cell line.

For the IB106 line, we observe that the response in the TNs condition appears to be significantly different from the irradiation conditions with over 250 DE genes specific to this condition (Figure 90). About 100 genes are common to all conditions and thus seem to be part of more generalized cellular response mechanisms while another 100 genes are common only to the irradiation conditions. These results are consistent with previous observations that point to a distinct response to TNs and irradiation. For the IB115 line, the low number of DE genes does not really allow any conclusion to be drawn, but we can nevertheless note the presence of some genes common to all the irradiation conditions as well as genes common to the irradiation+TNs conditions (Figure 91).

3. GO enrichment analysis

From these DE genes per condition, we then sought to know which cellular pathways are impacted by performing an enrichment analysis with gprofiler which determines them based on the number of impacted genes within the cellular pathway. We did not separate the DE genes here based on whether they are under- or over-expressed because we are primarily interested in the types of pathways impacted in the cellular response to our exposure conditions. By this method, we therefore obtained the significantly impacted Gene Ontologies (GO) per condition.

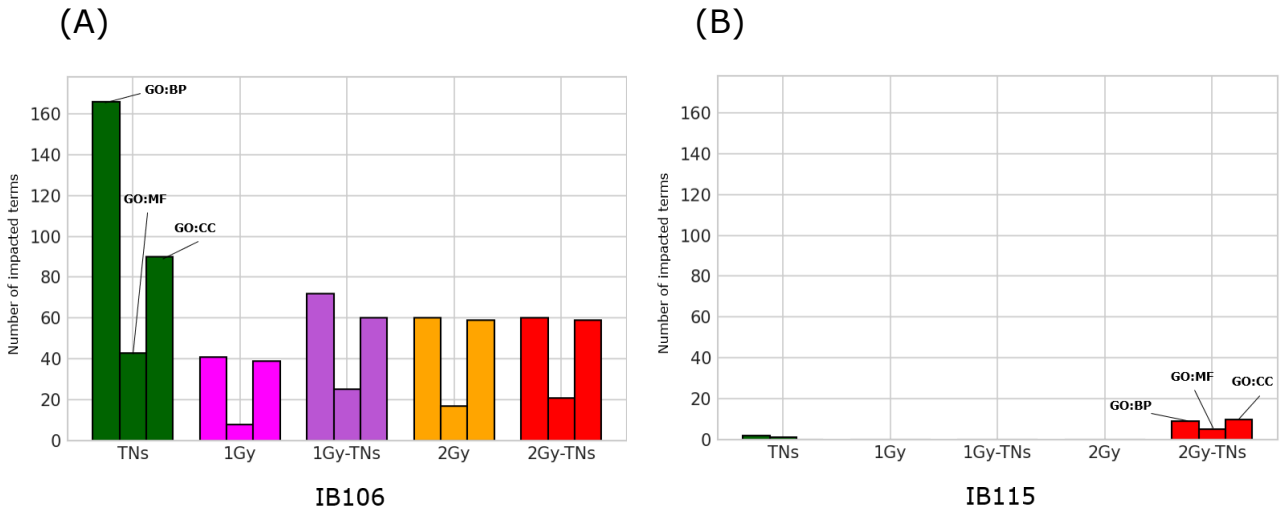


Figure 92. Number of significantly impacted GO terms per condition in (A) IB106 and (B) cell lines.

For line IB106, we see that the number of impacted ontologies correlates well with the number of DE genes (Figure 92A), contrary to line IB115 in which the low number of DE genes does not allow us to classify GOs as significantly impacted in the 1Gy+TNs condition, for example, despite the fact that it was the one with the most DE genes within this line (Figure 92B).

We were mainly interested in GO Biological Processes (GO:BP) that directly refer to the relevant cellular pathways and we identified that a large part of those significantly impacted GO:BP in the IB106 line were related to cellular pathways that we grouped under 4 main families: protein metabolism, cell cycle, cellular respiration and cell stress/death.

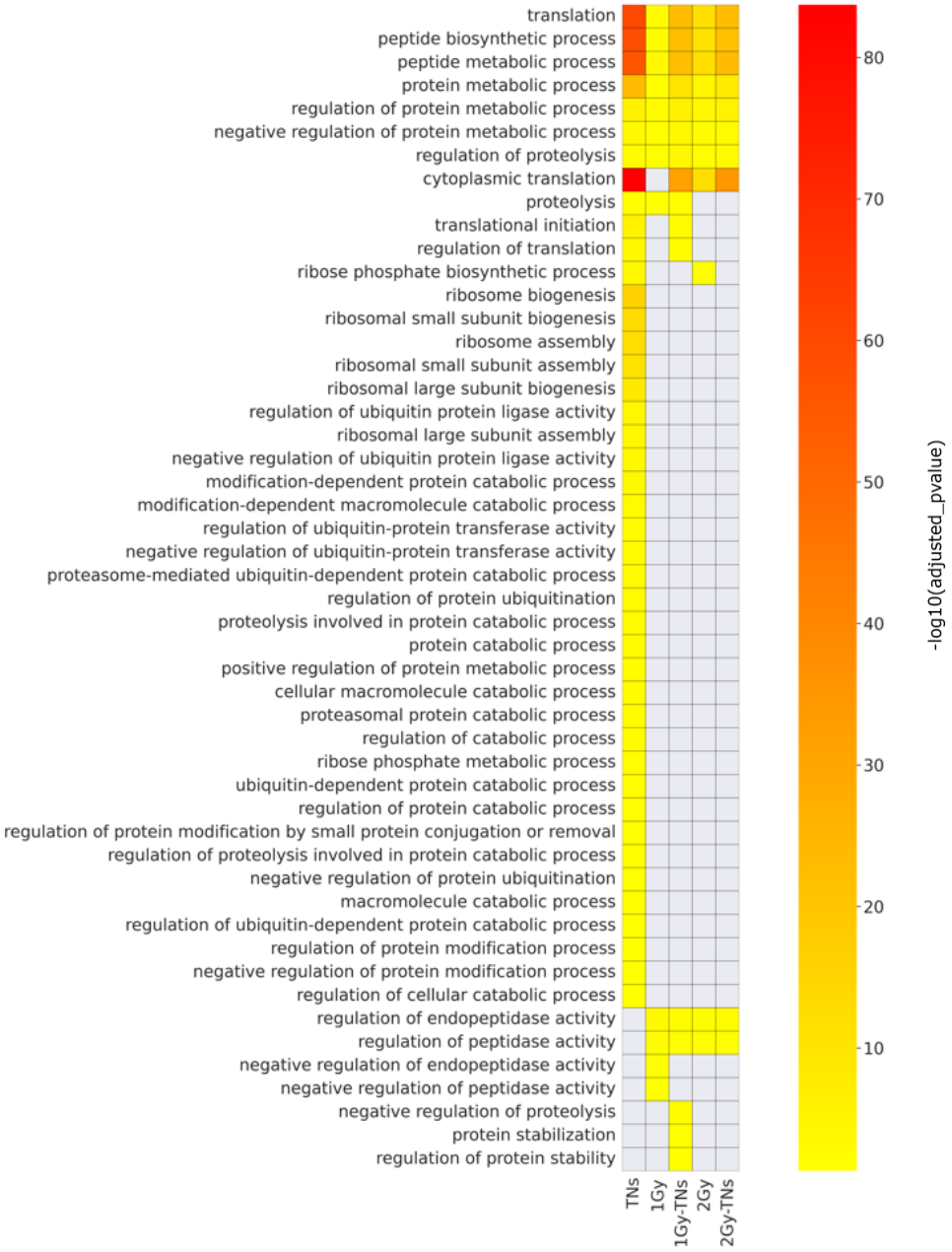


Figure 93. GO:BP significantly impacted (adjusted pvalue <0.05) related to protein metabolism in the different experimental conditions of the IB106 line.

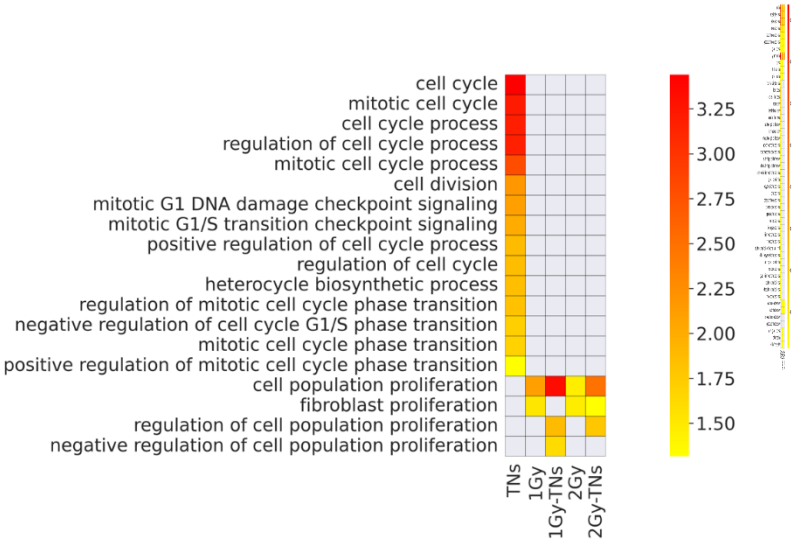


Figure 94. GO:BP significantly impacted (adjusted pvalue <0.05) related to the cell cycle in the different experimental conditions in the IB106 line.

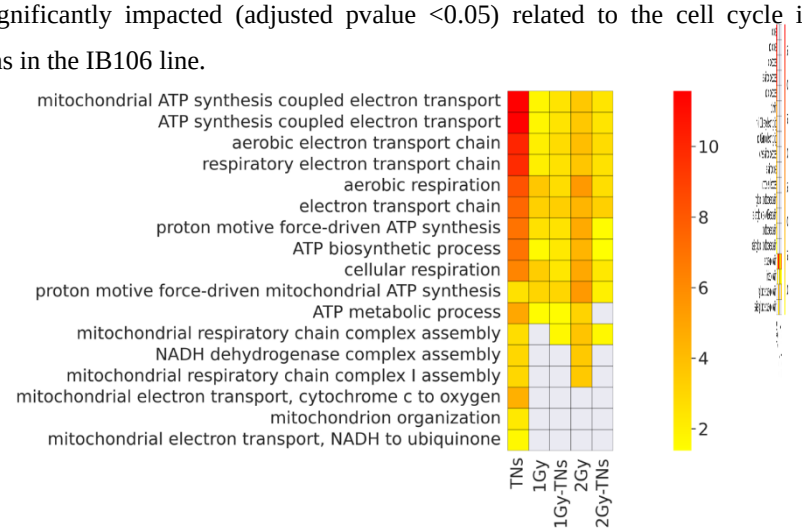


Figure 95. GO:BP significantly impacted (adjusted pvalue <0.05) related to cell respiration in the different experimental conditions in the IB106 line.

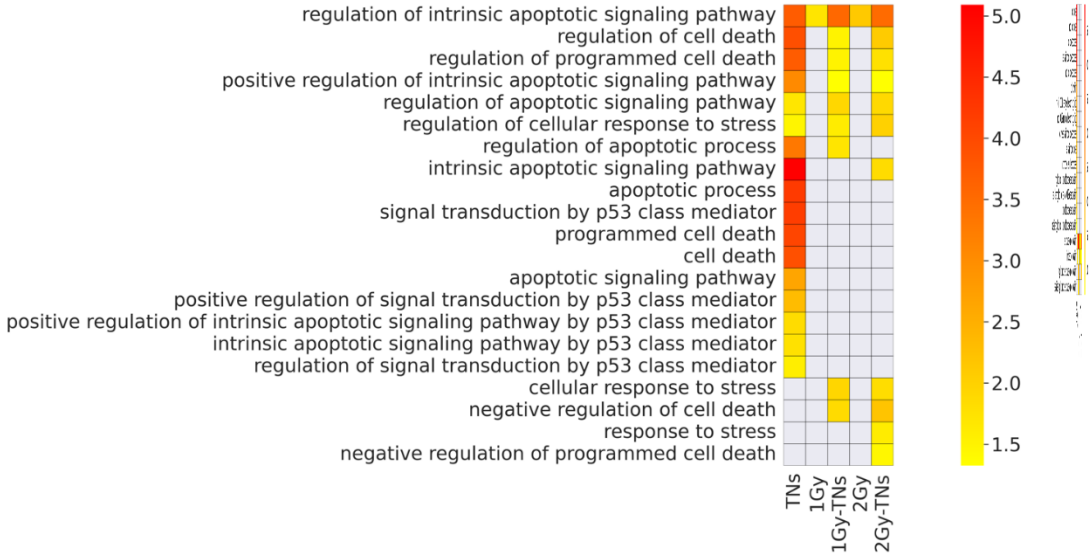


Figure 96. GO:BP significantly impacted (adjusted pvalue <0.05) related to stress and cell death in the different experimental conditions in the IB106 line.

For protein metabolism (Figure 93), the most impacted ontologies are common to all conditions but outside of those, the majority of impacted ontologies are specific to the TNs condition with only a handful of ontologies specific to irradiation conditions. This could suggest a greater magnitude of cellular response, involving a larger part of the cellular pathways, when cells are exposed only to TNs rather than in irradiation conditions.

For the cellular respiration pathways (Figure 94), we observe here that the majority of the impacted ontologies, which are also the most impacted, are common to all conditions, which seems to indicate that they are common cellular pathways in the stress response in these 2 conditions.

For pathways related to the cell cycle (Figure 95), a large number of pathways related to the regulation of the cell cycle and the different checkpoints are impacted only in the TNs condition, however the pathways related to cell proliferation are only impacted in the irradiation conditions.

Cellular pathways related to stress and cell death (Figure 96) seem to be mostly impacted only under TNs exposure conditions with only one pathway common to all conditions and related to the regulation of intrinsic apoptosis. Several terms are common to all these TNs conditions but there are always cellular pathways specific to the TNs-only condition. This seems to indicate that nanoparticles are conducive to triggering cell death pathways but that some of these pathways would no longer be impacted when this exposure is combined with irradiation.

In summary for the IB106 line, an apparently stronger cellular response is observed in the condition of exposure to TNs only but this response is eclipsed when irradiation is performed in addition to the exposure for a "weaker" response in terms of cellular pathways impacted. This cellular response to irradiation is however not constant and varies significantly between the different experimental conditions despite a common base of impacted genes and cellular pathways.

For the IB115 line, although few ontologies were impacted, we can see that the only condition in which several cellular pathways really stand out (the 2 ontologies of the TNs condition dealing with the response to leptomycin and thus appearing rather as an artifact) is the 2Gy+TNs condition which could potentially go in the direction of a nano-sensitization of cells.

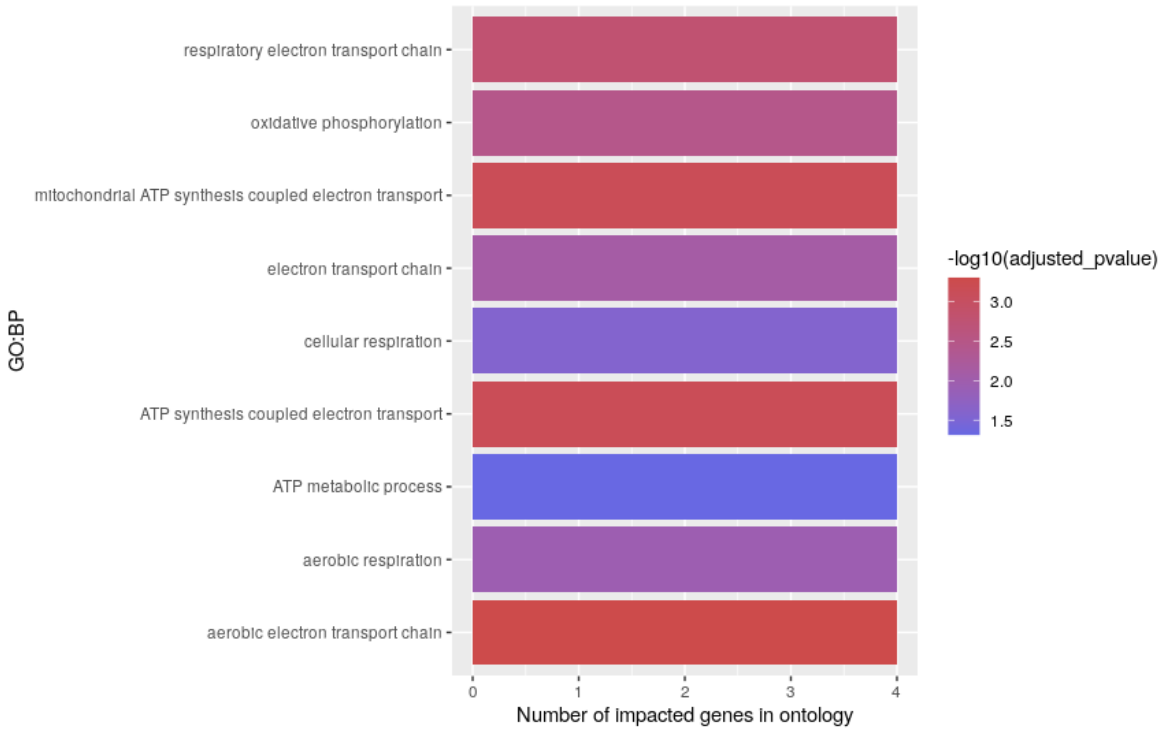


Figure 97. GO:BP terms impacted in the 2Gy+TNs condition of the IB115 line.

We find ontologies related to cellular respiration (Figure 97), in the same way as in the IB106 line. These ontologies are only supported by a few genes indicating that overall, this cellular response is rather weak but the results are coherent nonetheless. As neither the 2Gy nor the TNs condition express these ontologies individually, it is possible that this is indeed a case of radio-sensitization even if the low level of response calls for caution.

Discussion

Transcriptome analysis by sequencing on IB106 and IB115 sarcoma lines was used to study the cellular response to combined exposure to TiO_2 nanoparticles and controlled dose irradiation. We first observed that the IB106 line is much more sensitive to the exposure conditions used than the IB115 line whose radio-resistance characteristics had already been established but whose lack of cellular response to nanoparticle exposure and thus potential nano-resistance to TiO_2 was not. We thus find in the IB106 line two distinct types of cellular response which are either the response to TNs or the response to irradiation. In the conditions combining irradiation and TNs, the response observed is of the same nature and of the same level as in the irradiation-only conditions, which does not seem to indicate a nano-sensitization phenomenon in the cells of this line. For the IB115 line, if some DE genes can be detected, they do not translate for most conditions into a significant impact on cellular pathways except for the 2Gy+TNs condition. In the latter condition, an impact on cellular respiration pathways is observed, which is a common pathway in the response to radiation-induced damage[10,11]. It is therefore possible that a nano-sensitization of the cells has taken place, leading to the appearance of a cellular response to irradiation despite the radio-resistance of this cell line.

Cell survival studies carried out on these lines under the same experimental conditions but over several days of exposure seem to indicate that a phenomenon of nano-sensitization does take place on the 2 lines but that it can take effect over a longer time, in particular for the IB115 line (3-6 days). The study of the cellular response by transcriptomic analysis after 24 hours of exposure is perhaps too early to observe a greater response in this radio-resistant line and a similar analysis after 3 or 6 days of exposure would produce more results, although the decrease in cell survival would undoubtedly complicate the RNA extraction. Samples with more cells, to increase mRNA yields, could also allow Direct-RNA sequencing to address modified bases.

References

[1] Rajamanickam Baskar et al., 'Cancer and Radiation Therapy: Current Advances and Future Directions', *International Journal of Medical Sciences* 9, no. 3 (27 February 2012): 193–99, https://doi.org/10.7150/ijms.3635.

[2] B. Jones et al., 'The Role of Biologically Effective Dose (BED) in Clinical Oncology', *Clinical Oncology* 13, no. 2 (1 April 2001): 71–81, https://doi.org/10.1053/clon.2001.9221.

[3] Nicholas G Zaorsky et al., 'Impact of Radiation Therapy Dose Escalation on Prostate Cancer Outcomes and Toxicities', *American Journal of Clinical Oncology* 41, no. 4 (April 2018): 409–15, https://doi.org/10.1097/COC.0000000000000285.

[4] Syed S. Mahmood and Anju Nohria, 'Cardiovascular Complications of Cranial and Neck Radiation', *Current Treatment Options in Cardiovascular Medicine* 18, no. 7 (July 2016): 45, https://doi.org/10.1007/s11936-016-0468-4.

[5] Cristina Carretero et al., 'Gastroduodenal Injury after Radioembolization of Hepatic Tumors', *The American Journal of Gastroenterology* 102, no. 6 (June 2007): 1216–20, https://doi.org/10.1111/j.1572-0241.2007.01172.x.

[6] Deep Kwatra, Anand Venugopal, and Shrikant Anant, 'Nanoparticles in Radiation Therapy: A Summary of Various Approaches to Enhance Radiosensitization in Cancer', *Translational Cancer Research* 2, no. 4 (September 2013), https://doi.org/10.3978/j.issn.2218-676X.2013.08.06.

[7] James F. Hainfeld et al., 'Gold Nanoparticles Enhance the Radiation Therapy of a Murine Squamous Cell Carcinoma', *Physics in Medicine & Biology* 55, no. 11 (May 2010): 3045, https://doi.org/10.1088/0031-9155/55/11/004.

[8] Anastasia K. Hauser et al., 'Targeted Iron Oxide Nanoparticles for the Enhancement of Radiation Therapy', *Biomaterials* 105 (1 October 2016): 127–35, https://doi.org/10.1016/j.biomaterials.2016.07.032.

[9] Paul Retif et al., 'Nanoparticles for Radiation Therapy Enhancement: The Key Parameters', *Theranostics* 5, no. 9 (11 June 2015): 1030–44, https://doi.org/10.7150/thno.11642.

[10] Tohru Yamamori et al., 'Ionizing Radiation Induces Mitochondrial Reactive Oxygen Species Production Accompanied by Upregulation of Mitochondrial Electron Transport Chain Function and Mitochondrial Content under Control of the Cell Cycle Checkpoint', *Free Radical Biology and Medicine* 53, no. 2 (15 July 2012): 260–70, https://doi.org/10.1016/j.freeradbiomed.2012.04.033.

[11] Winnie Wai-Ying Kam and Richard B. Banati, 'Effects of Ionizing Radiation on Mitochondria', *Free Radical Biology and Medicine* 65 (1 December 2013): 607–19, https://doi.org/10.1016/j.freeradbiomed.2013.07.024.