DNA Repair 74 (2019) 26–37

Contents lists available at ScienceDirect

DNA Repair

journal homepage: www.elsevier.com/locate/dnarepair

An open-source algorithm for rapid unbiased determination of DNA fiber length

Pierre Ghesquière[a], Abdelhamid Elsherbiny[a,b], Emile Fortier[a,b], Mary McQuaid[a,b], Javier Mazzaferri[a], François Bélanger[a], Farida Cheriet[a,c], Elliot Drobetsky[a,b,d], Hugo Wurtele[a,b,d], Santiago Costantino[a,e,*]

[a] Centre de Recherche de l'Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada
[b] Programmes de Biologie Moléculaire, Université de Montréal, Montréal, Québec, Canada
[c] Département de Génie Informatique et Génie Logiciel, Polytechnique Montréal, Québec, Canada
[d] Département de Médecine, Université de Montréal, Montréal, Québec, Canada
[e] Département d'Ophtalmologie, Université de Montréal, Montréal, Québec, Canada

ARTICLE INFO

Keywords:
DNA fiber fluorography
Automatic Fiber Quantification
Open-source fiber analysis

ABSTRACT

DNA fiber fluorography is widely employed to study the kinetics of DNA replication, but the usefulness of this approach has been limited by the lack of freely-available automated analysis tools. Quantification of DNA fibers usually relies on manual examination of immunofluorescence microscopy images, which is laborious and prone to inter- and intra-operator variability. To address this, we developed an unbiased, fully automated algorithm that quantifies length and color of DNA fibers from fluorescence microscopy images. Our fiber quantification method, termed FiberQ, is an open-source image processing tool based on edge detection and a novel segment splicing approach. Here, we describe the algorithm in detail, validate our results experimentally, and benchmark the analysis against manual assessments. Our implementation is offered free of charge to the scientific community under the General Public License.

1. Introduction

DNA replication is tightly regulated by a myriad of molecular mechanisms that ensure accurate transmission of genetic information to daughter cells. The fidelity of this process can be compromised by DNA replicative stress, i.e., the abnormal slowing down or stalling of DNA replication forks (RF) [1]. Indeed, stalled RF must be resolved in a timely manner to prevent their "collapse" into highly-genotoxic DNA double-strand breaks (DSB), which in turn engender chromosomal rearrangements and genomic instability [2]. Replicative stress may arise from various impediments to DNA synthesis, such as DNA secondary structures (eg. G-quadruplexes, palindromes) [3], RNA:DNA hybrids (R-loops) [4], collisions between replication and transcription machineries [5], dNTP pool imbalances [6], or DNA adducts induced by any among a plethora of genotoxins (including endogenous agents, environmental mutagens, and chemotherapeutic drugs). Mutation or defective regulation of essential DNA replication factors, as well as activation of certain oncogenes including Ras Myc, and Bcl-2, have also been shown to cause abnormal DNA replication dynamics [7,8]. The resultant replicative stress-induced genomic instability constitutes a critical determinant in both cancer development and treatment. Replicative stress is also implicated in the molecular pathogenesis of aging and neurodegenerative disease, as well as developmental syndromes such as primordial dwarfism [9,10].

DNA fiber fluorography is commonly used to evaluate RF progression at the level of individual DNA molecules [11]. This method is based on the incorporation of halogenated nucleotide analogs, such as chloro- (CldU), iodo- (IdU), or bromo-deoxyuridine (BrdU) into nascent DNA at RF in living cells. In a typical experiment, sequential incorporation of two nucleotide analogs, e.g., IdU and CldU, is performed. Cells are exposed to DNA replication stress-inducing treatments during or after the second labeling period [12]. Following cell lysis and spreading of DNA on microscopy slides, DNA molecules are labeled using anti-IdU and anti-CldU antibodies coupled to different fluorophores. Two-color images generated by fluorescence microscopy then reveal contiguous labeled regions in elongated DNA fibers. Measurement of the respective lengths of these labeled stretches of DNA permits quantification of RF progression. Variations of this general experimental strategy have been used extensively to quantify DNA replication dynamics in the context of replicative stress induced by a plethora of

* Corresponding author at: Centre de Recherche de l'Hôpital Maisonneuve-Rosemont, Montréal, Québec, Canada.
E-mail address: santiago.costantino@umontreal.ca (S. Costantino).

https://doi.org/10.1016/j.dnarep.2019.01.003
Received 21 September 2018; Received in revised form 21 December 2018; Accepted 7 January 2019
Available online 11 January 2019
1568-7864/

experimental conditions, including exposure to chemotherapy drugs and expression of oncogenes [13,14].

Evaluation of DNA fiber length is generally performed manually using simple image manipulation tools. This procedure is laborious and subject to inter-user variability stemming in part from unintended bias in the choice of fibers to be measured. These problems highlight the need for a reliable computational method for unbiased analysis of DNA fiber immunofluorescence images. To the best of our knowledge, the only available tool is not free [15], and the source-code not open, rendering it unavailable for wide distribution and public validation. Here we present FiberQ, a novel, fully-automated algorithm to segment (ie. delineate) and quantify labeled DNA fiber length from fluorescence microscopy images. FiberQ is based on edge detection filters and splicing techniques, and provides rapid, reliable, and unbiased analysis of DNA fibers. We describe our algorithm in full detail, and use images obtained under different experimental conditions to compare its performance with manual segmentation. Our open-source software is offered free of charge to the scientific community.

2. Results

2.1. Inter-user variability upon manual quantification of DNA fibers on immunofluorescence images

DNA fiber immunofluorescence images display variations in fiber density, straightness, branching, and staining intensity. Inter-user variability due to biased identification and inaccurate measurement of isolated fibers is expected. To evaluate this variability, three experienced users were asked to manually segment the same set of 6 images obtained from two different experiments (3 images per experiment). Fig. 1A shows an example of such images extracted from the second sample. Cells were pulsed sequentially with IdU and CldU such that bicolor contiguous regions of DNA fibers represent progressing RF which first incorporated IdU, and then CldU, into nascent DNA. Both experiments only differ in the incubation time of the second pulse (20 min for Experiment 1 and 30 min for Experiment 2). Using a widely-available open-source image manipulation software, GIMP (GNU Image Manipulation Program), users colored CldU and IdU fiber sections in red and green, respectively. For each bicolor fiber, the ratio r ($r = \frac{CldU\ length}{IdU\ length}$), a commonly-used metric to quantify the dynamics of RFs [11,13,14], was measured.

Despite the global increase of ratio distributions between both experiments showing that the manual quantification is consistent with the biological process, statistically significant differences were observed between User 3 and the other two users (see Fig. 1 and Table 1).

Disagreement among users illustrates the challenges of manual selection of fibers and can be explained by several factors:

1) The number of segmented fibers varied between users (Table 2). For example, User 2 measured 61 more fibers than User 1 and 37 more fibers than User 3 in Experiment 1, which represents respectively 21% and 13% of all the 285 segmented fibers (Table 2). Moreover, only 16% and 23% of the total number of segmented fibers were quantified by every user for experiment 1 and 2 respectively. As an illustration, Fig. 1B shows how many users chose each segmented fiber in an image extracted from Experiment 2.
2) Overlap between IdU and CldU signals complicates the precise localization of label changes, leading to variations in r ratio quantification between users for a given fiber (see Fig. 2A–C).
3) Staining gaps that split fibers into smaller segments may cause disparities between users. Indeed, splicing such segments into a single fiber is a subjective choice (see Fig. 2D).
4) Some users may tolerate high radii of curvature in single fibers whereas others prefer straight fibers (see Fig. 2E,F).
5) Entangled fibers, debris and non-specific antibody staining can interfere with accurate measurements (see Fig. 2G).
6) Loss of focus during long quantification sessions yield mistakes ranging from delineation outside of the fiber (Fig. 2B-User 1) to omission of small low contrasted segments (Fig. 2H-User 1 and 3).

Importantly, most biases are not only a source of variability for human segmentation, they also pose a challenge for the development of image processing segmentation algorithms. Thus, FiberQ includes mathematically well-defined criteria for identification of DNA fibers within a noisy image, robust fiber splicing rules for bridging the gaps along elongated straight fibers, and strategies for removal of unexploitable tangled fibers.

The pipeline of FiberQ is summarized in Fig. 3 and detailed in the Materials and Methods section. Briefly, after a preprocessing step in which fibers are enhanced with respect to the background, DNA fibers are detected using an *ad hoc* edge detection method inspired from Canny's [16] and Marr Hildreth's [17]. The splicing of nearby segments that may belong to a single fiber is based on their curvature and distance. Unexploitable clusters of fibers are removed by establishing the maximum local fiber density. Finally, color transitions are determined by analyzing the difference in fluorescence between IdU and CldU channels.

2.2. Comparing FiberQ vs manual quantification

Manual quantifications were compared with our algorithm by computing the correlations between both methods, as well as inter-operator differences for humans and FiberQ. We used a database of 98 images and measured the median CldU/IdU length ratios for each image. We observed a very good correlation between the algorithm and users (Pearson's coefficient of 0.79), demonstrating that FiberQ is consistent with trained users' observations (Fig. 4A).

To evaluate inter-operator differences for individual fibers, 12 images were segmented manually by three experienced users and FiberQ. For each fiber segmented by two operators, we compared their respective lengths and ratios r by using three metrics: $C^{green}_{opi,opj}$, $C^{red}_{opi,opj}$ and $\Delta ratio_{opi,opj}$.

- $C^{green}_{opi,opj}$ *and* $C^{red}_{opi,opj}$ measure the difference of length between operator i and operator j normalized by the mean of the 2 lengths:

$$C^{green}_{opi,opj} = \frac{(l^{green}_{opi} - l^{green}_{opj})}{\frac{l^{green}_{opi} + l^{green}_{opj}}{2}}$$

  where l^{green}_{opi} (resp. l^{green}_{opj}) are the length of the green part of the fiber measured by operator i or j.
- $\Delta ratio_{opi,opj}$ is the difference of CldU/IdU ratios for a fiber segmented by op_i and opj.

For the above metrics, only bicolor fibers were considered, and their distributions were computed for each possible pair of operators (Fig. 4B–D). We observed similar inter-operator variability when comparing either FiberQ vs users, or User i vs User j. We also counted the fibers segmented by FiberQ that were also quantified by the human users (Table 3) and found that the majority of fibers (73%) are shared by at least one user. The remaining 27% include a mix of what seem to be "good" fibers forgotten by users, mistakes made by FiberQ and also ambiguous situations (see S.Fig. 1).

2.3. Performance of FiberQ in biologically-relevant experimental conditions

We next sought to validate FiberQ experimentally by comparing the RF progression in samples treated with hydroxyurea (HU) vs untreated controls. Hydroxyurea inhibits the activity of the ribonucleotide reductase enzymatic complex [18], thereby depleting deoxyribonucleotide pools and strongly slowing down RF progression. We first

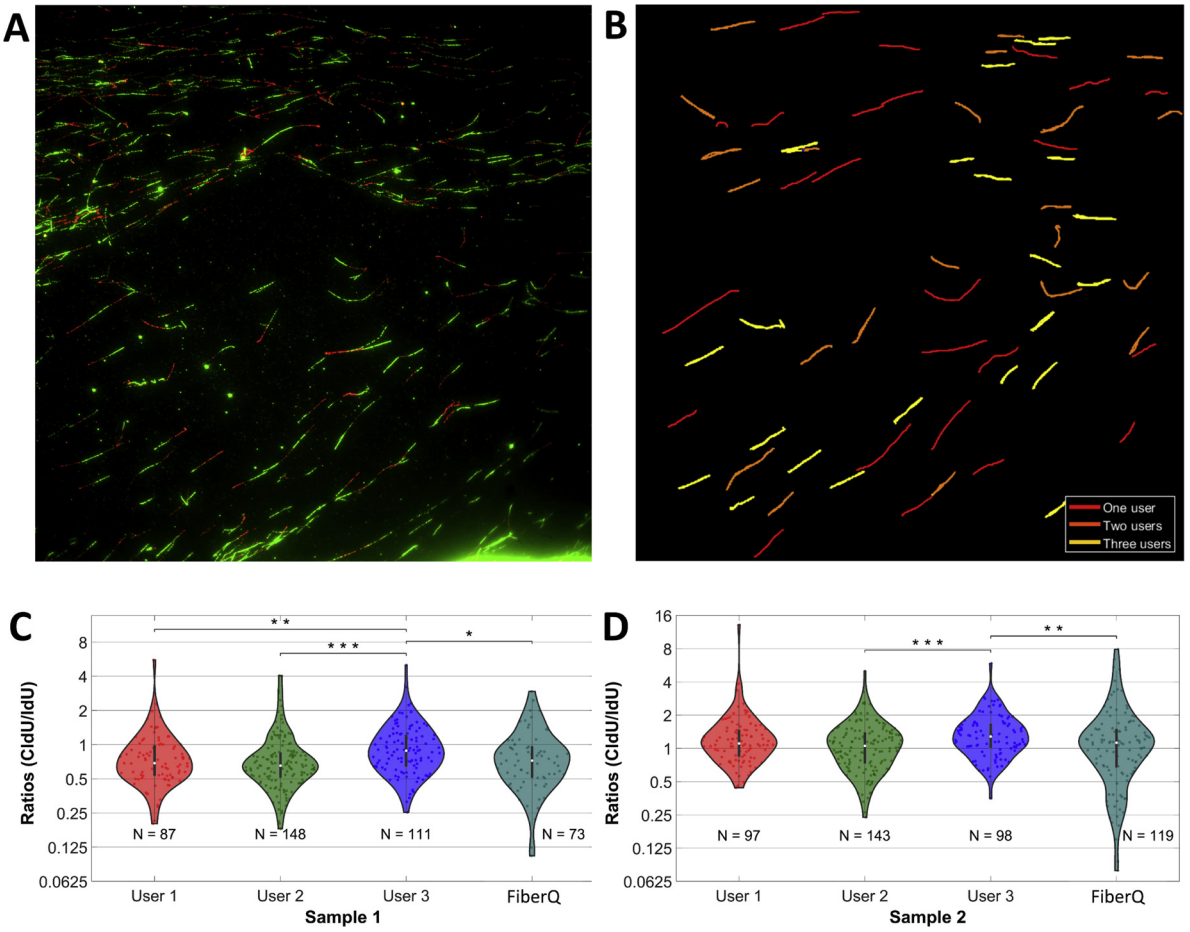

Fig. 1. Inter-user variability. Three experienced users segmented the same set of 6 images obtained from two samples. A- Example of one fluorescent image from the second sample. B- Illustration of the different segmentations performed by three experienced users. Yellow fibers were segmented by the 3 users, orange fibers by only 2 users and red fibers by only 1 user. C- Ratio (CldU/IdU) distribution for each user (sample 1). The distribution of the 3rd user shows a significative difference ($p < 10^{-2}$ between User 1 and User 3, $p < 10^{-4}$ between User 2 and User 3, $p < 0.02$ between FiberQ and User 3, Mann Whitney test). D-Ratio (CldU/IdU) distribution for each user (sample 2). The distribution of User 3 shows a significative difference with User 2 ($p < 10^{-4}$) and with FiberQ ($p < 10^{-2}$).

quantified images acquired from a typical experiment in which HU was included in the culture media of HeLa cells during the second pulse with halogenated nucleotides (Fig. 5A-B). IdU/CldU pulses were 30 min/60 min for the first experiment, and 30 min/90 min for the second experiment. As expected, both FiberQ and manual quantifications reveal that CldU-labeled tracks are shorter in HU-treated samples than in control untreated samples, leading to reduced CldU/IdU length ratio (Fig. 5A-B).

Our group and others have previously shown that nascent DNA is unstable in certain cell lines due to aberrant nuclease activity at stalled RF [13,14]. We also recently showed that overexpression of all three subunits of the Replication Protein A complex suppresses such nascent DNA instability in the ovarian cancer cell line OV1946 [14]. To evaluate the stability of nascent DNA at stalled RFs, OV1946 cells were exposed to HU for 3 h *after* the second pulse. Reduction in the length of the second label (CldU) upon incubation with HU reflects nuclease-mediated degradation of nascent DNA, which is rescued by RPA overexpression. Both manual quantification and our algorithm confirmed that GFP-RPA expression leads to higher CldU/IdU ratios vs GFP alone, as expected (Fig. 5C). We note that measurements performed with FiberQ generally display higher variance, partly due to a much larger number of measured fibers.

We also validated FiberQ by varying the duration of the nucleotide analog pulse and evaluating the effect on labeled track length. Incubation time for the second nucleotide analogue (CldU) was incrementally increased from 10 to 90 min, whereas the IdU pulse period remained constant at 20 min. Manual and FiberQ quantification of the images indicate, as expected, an increase in CldU/IdU ratio for CldU labeling periods of up to 60 min (Fig. 6A-B). Intriguingly, we observed that the length of the CldU tracks reaches a plateau between 60 and 90 min of labeling. Examination of the images reveals that this is likely due to the presence of extremely long fibers, which are almost invariably entangled in clusters or cut out of the image field. Only very short isolated DNA molecules can be detected under these conditions, which introduces biases in both manual and automatic segmentations.

Table 1
Inter-User variability: p-values for each Mann-Whitney test on the ratio distributions (CldU/IdU).

	User 1 vs User 2	User 1 vs User 3	User 2 vs User 3	FiberQ vs User 1	FiberQ vs User 2	FiberQ vs User 3
Experiment 1	0.38	**2.7e-3**	**1.6e-5**	0.90	0.35	**0.01**
Experiment 2	0.057	0.084	**4.0e-4**	0.21	0.85	**9.5e-3**

Table 2
Inter-user variability: manual quantification of CldU- and IdU-labeled DNA fibers by three experienced users. The results of two independent experiments are presented. The last three columns show the P-value when comparing distribution of ratios r ($r = \frac{CldU\ length}{IdU\ length}$) between users.

	Total nb of manually segmented fibers	Fibers segmented by User 1	Fibers segmented by User 2	Fibers segmented by User 3	Fibers chosen by only one user	Fibers chosen by only two users	Fibers chosen by the three users
Experiment 1	285	87 (31%)	148 (52%)	111 (39%)	166 (58%)	72 (25%)	47 (16%)
Experiment 2	273	97 (36%)	143 (52%)	98 (36%)	100 (37%)	111 (41%)	62 (23%)
Total	558	184 (33%)	291 (52%)	209 (37%)	266 (48%)	183 (33%)	109 (20%)

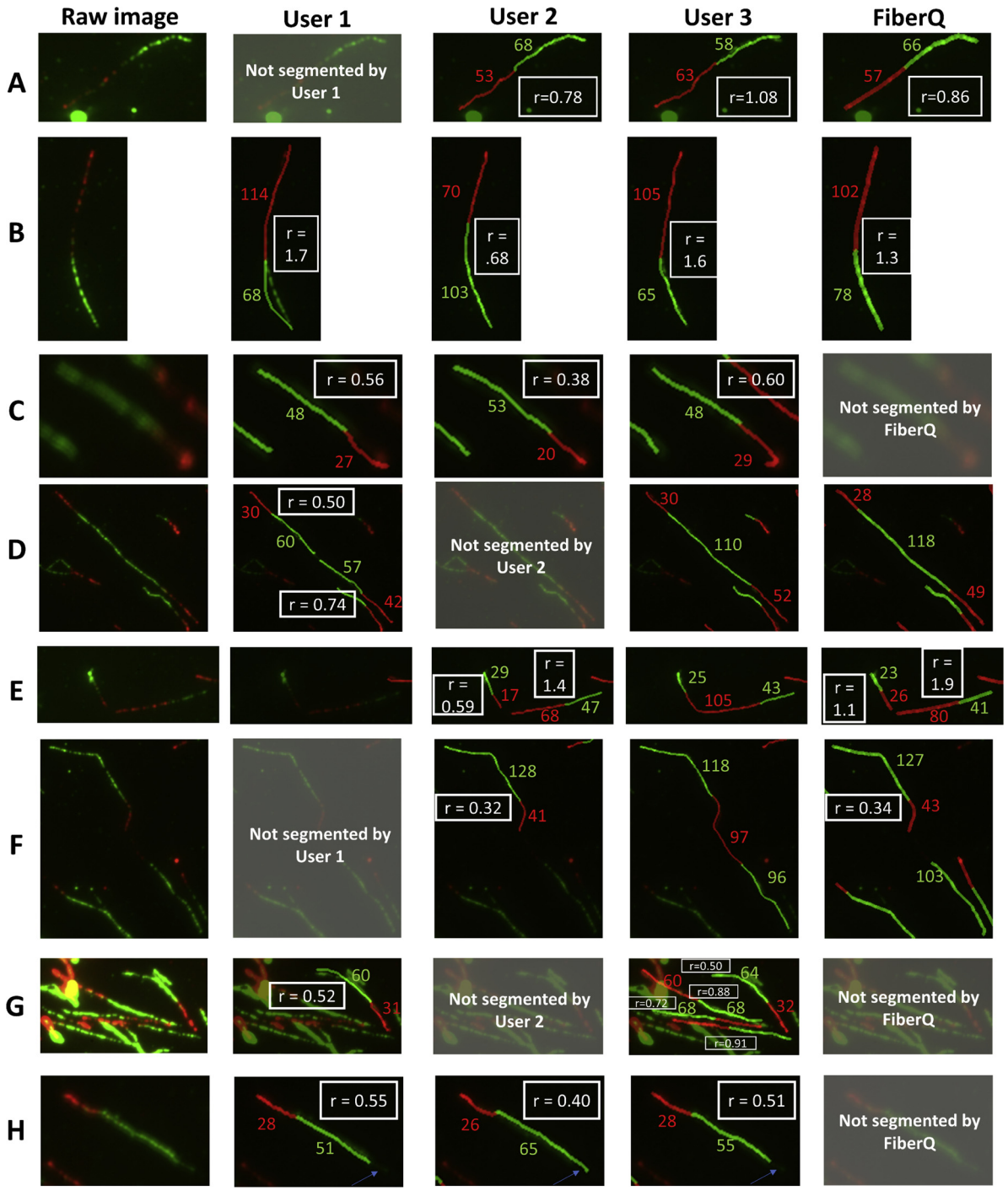

Fig. 2. Examples of segmented fibers by manual users and FiberQ. Lengths of the IdU and CldU are respectively written in green and red. For bicolor fibers, the ratios (CldU/IdU) are indicated in white.

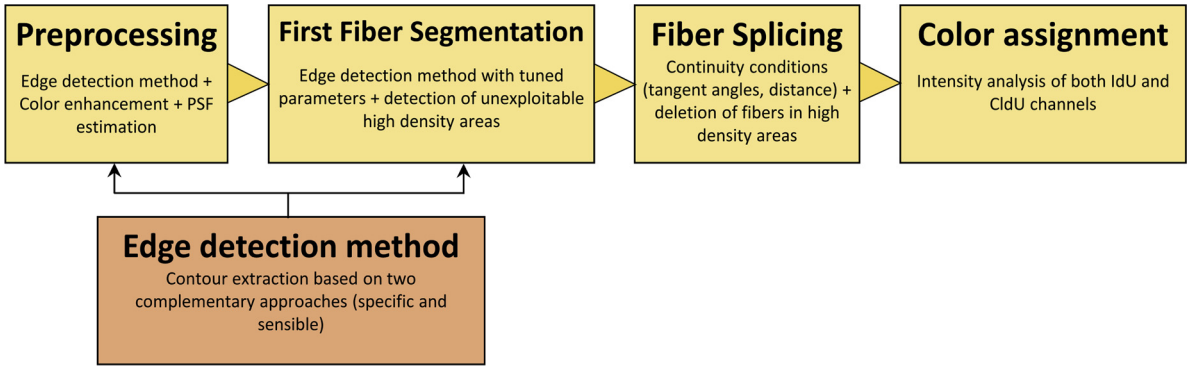

Fig. 3. General Framework of FiberQ. After a preprocessing step in which fibers -roughly segmented with the Edge detection method- are enhanced with respect to the background, the Point Spread Function (PSF) of the imaging system is calculated. The PSF aims at tuning spatial parameters (convolution filter size, maximum splicing distance, etc.) Then, a better segmentation is obtained on the enhanced image with tuned parameters. Unexploitable clusters of fibers are detected by measuring local fiber density. A Fiber Splicing algorithm connects nearby segments that belong to the same fiber and deletes fibers passing through hugh fiber density zones. Finally, a color label (e.g. red or green) is assigned to differentiate CldU vs IdU signals. This color assignment is based on the analysis of the intensity of both IdU and CldU channels.

3. Materials and Methods

3.1. DNA Fiber assay

Exponentially growing Hela cells were labeled with 10 µL IdU for a duration T_1. Cell were washed twice with 3 ml PBS, then labeled with 250 µL CldU for a duration T_2. Cells were washed, harvested, and re-suspended in PBS at a final concentration of 500 cells/µL. Two µL were transferred to a slide, overlaid with 7.5 µL lysis buffer (0.5% SDS, 200 mM Tris-HCl (pH 7.4), and 50 mM EDTA), and incubated at room temperature for 3 min. Slides were tilted to allow DNA to spread by gravity, air-dried for 7 min, fixed for 10 min with freshly prepared 3:1 methanol/acetic acid, and air-dried for 7 min. DNA was denatured by incubating the slide in 2.5 M HCl for 80 min, followed by three washes with PBS. Blocking was performed with 200 µL 5% BSA for 20 min. For immunostaining, slides were incubated for 2 h with primary antibodies; ab6326 anti-BrdU (cross-reacts with CldU) antibody (rat) (1:400) and BD Biosciences 347580 anti-BrdU (cross-reacts with ldU) antibody (mouse) (1:25) in 5% BSA in PBS. Slides were washed three times with PBS-T (PBS + 0.05% tween), then once with PBS. Next, slides were incubated for one hour with the secondary antibodies; anti-rat AIexa-594 (1:100) and goat anti-mouse Alexa—488 (1:100) in 5% BSA in PBS. Slides were washed three times with PBS-T (PBS + 0.05% tween), then once with PBS. Slides were allowed to dry in air for few minutes then mounting medium was added and images acquired using two different microscopes: either GE Healthcare Deltavision or ZEISS Axio Imager 2.

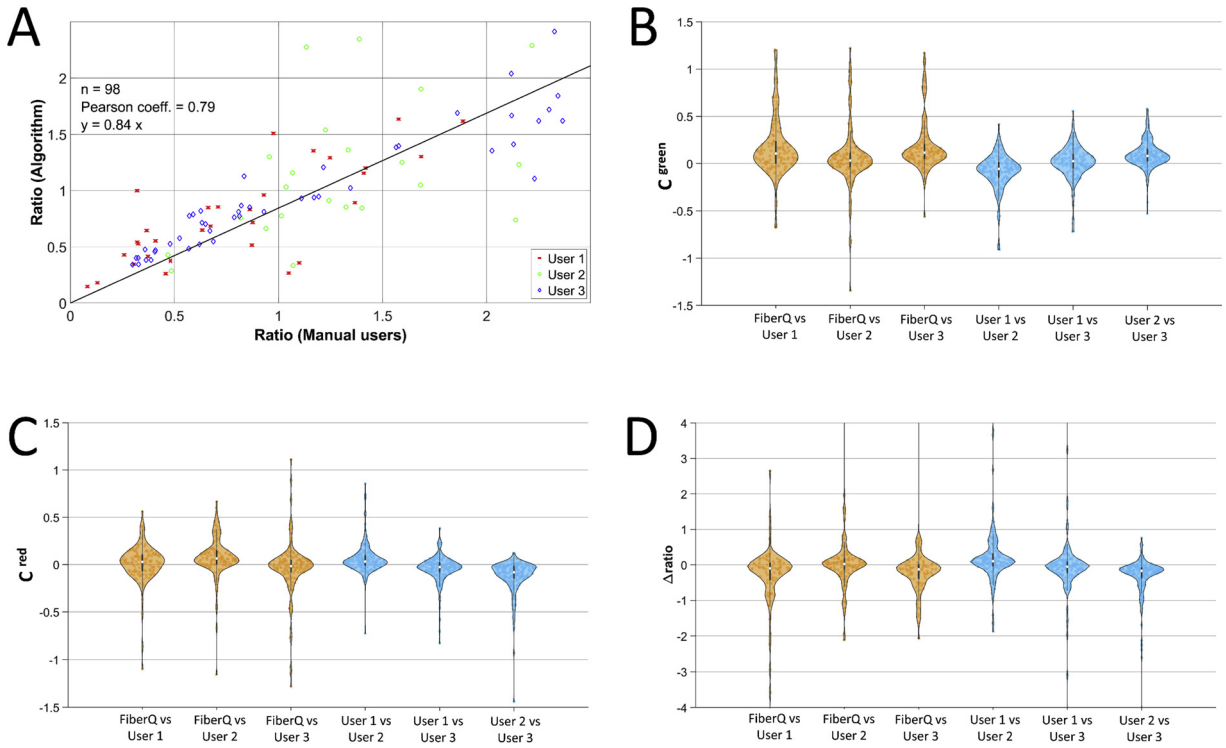

Fig. 4. Comparison of FiberQ vs human users. A- Correlation plot for 98 images quantified by FiberQ and manual users. B- Distribution of $C^{green}_{opi,opj}$ (normalized difference of lengths for green portions) for each pair of operators. C- Distribution of $C^{red}_{opi,opj}$ (normalized difference of lengths of red portions) for each pair of operators. D- Distribution of $\Delta ratio_{opi,opj}$ (difference of ratios) for each pair of operators.

Table 3
Overlap between FiberQ and human operators: Number of automatically segmented fibers (FiberQ fibers) also segmented by human users.

Total nb of FiberQ fibers	FiberQ fibers segmented by User 1	FiberQ fibers segmented by User 2	FiberQ fibers segmented by User 3	FiberQ fibers chosen by only one user	FiberQ fibers chosen by only two users	FiberQ fibers chosen by the three users	FiberQ fibers chosen by at least one user
269	130 (48%)	171 (64%)	132 (50%)	48 (18%)	59 (22%)	89 (33%)	196 (73%)

3.2. FiberQ algorithm: Method

From a raw image made up of two color channels (eg. IdU and CldU), DNA fibers are segmented and the length of each fluorescent marker is quantified to evaluate DNA replication dynamics. Our framework (summarized in Fig. 3) is divided in four steps. Fibers are first enhanced with respect to the background and then extracted with an ad-hoc edge detection method. As this first segmentation is interfered with by gaps within strands, a splicing step reconnects sections belonging to single DNA fibers. Finally, we quantify the length of each fluorophore by analysing channel intensity differences.

In our analysis pipeline, some parameters have been experimentally optimized using a large image database obtained with two imaging systems (GE Healthcare Deltavision and ZEISS Axio Imager 2). All these Experimentally Optimized Parameters (noted EOP_i hereafter) are expressed related to the diameter of the Point Spread Function of the imaging system to derive spatial metrics. If necessary, they can easily be tuned by the user.

Table 5 displays the values that we have established in our implementation of FiberQ.

3.2.1. Preprocessing: color enhancement and point spread function estimation

To smooth the raw image without altering the edge, a 4 × 4 × 1 median filter is applied to both channels. For simplicity, we call I_1, the channel of the first nucleotide analog, and I_2 the channel of the second one. The two channels are combined into a grayscale image, $I_{gray1} = \frac{I_1 + I_2}{2}$.

An *ad-hoc* edge detection algorithm (see details below) is applied on I_{gray1} to obtain a first rough segmentation of DNA fibers. This rough segmentation (BW_{DNA1}) is a binary image in which true pixels represent fiber pixels and false pixels are considered background.

Intensity normalisation is performed separately on both I_1 and I_2. On each of these two channels, we calculate the 5th and the 95th intensity percentiles of *fiber pixels* (ie. pixels belonging to the foreground mask of BW_{DNA1}). We linearly map the intensity values of both channels by saturating the bottom and top intensities to those two percentiles. The new normalised channels I_{1N} and I_{2N} which are matrices of doubles in the interval [0,1], have enhanced fiber fluorescence with respect to the background (Fig. 7B). A new grayscale image, I_{gray2}, is obtained by combining I_{1N} and I_{2N}.

The rough DNA mask, BW_{DNA1}, is also used to estimate the diameter of the Point Spread Function (PSF) of the imaging system in pixel units. Throughout the analysis pipeline, this PSF is used as a characteristic metric to adjust all spatial parameters (morphological operators, structuring elements, convolutional filters kernels, etc.) to the individual images. The diameter of the PSF is estimated by measuring the fiber width: for each fiber, the intensity distribution on a cross section s is fitted by a Gaussian function $P(\mu, \sigma) = A.\ exp(-\frac{(s-\mu)^2}{2\sigma^2})$. The PSF diameter is set to the median of all σ measured.

3.2.2. First fiber segmentation

To obtain a first fiber segmentation adapted to the structure of the input image, the ad-hoc *edge detection method* is applied again but this time on the enhanced grayscale image (I_{gray2}) with tuned spatial parameters based on the PSF (see details below). At the output of the edge detection method, we get a rough fiber segmentation BW_{DNA2}.

Large clusters of overlapping DNA molecules cannot be adequately analysed and therefore need to be removed from BW_{DNA2}. We calculate the foreground local pixel density $d(x, y)$ by convolving BW_{DNA2} with a gaussian kernel of standard deviation $\sigma = EOP_1.\ PSF$. Zones where $d(x, y) < EOP_2$ (ie. zones with a high concentration of fibers) are discarded for fiber analysis (Fig. 7C).

A second filter removes objects of width larger than EOP_3 times the PSF. More precisely, we discard objects for which the minor axis of the

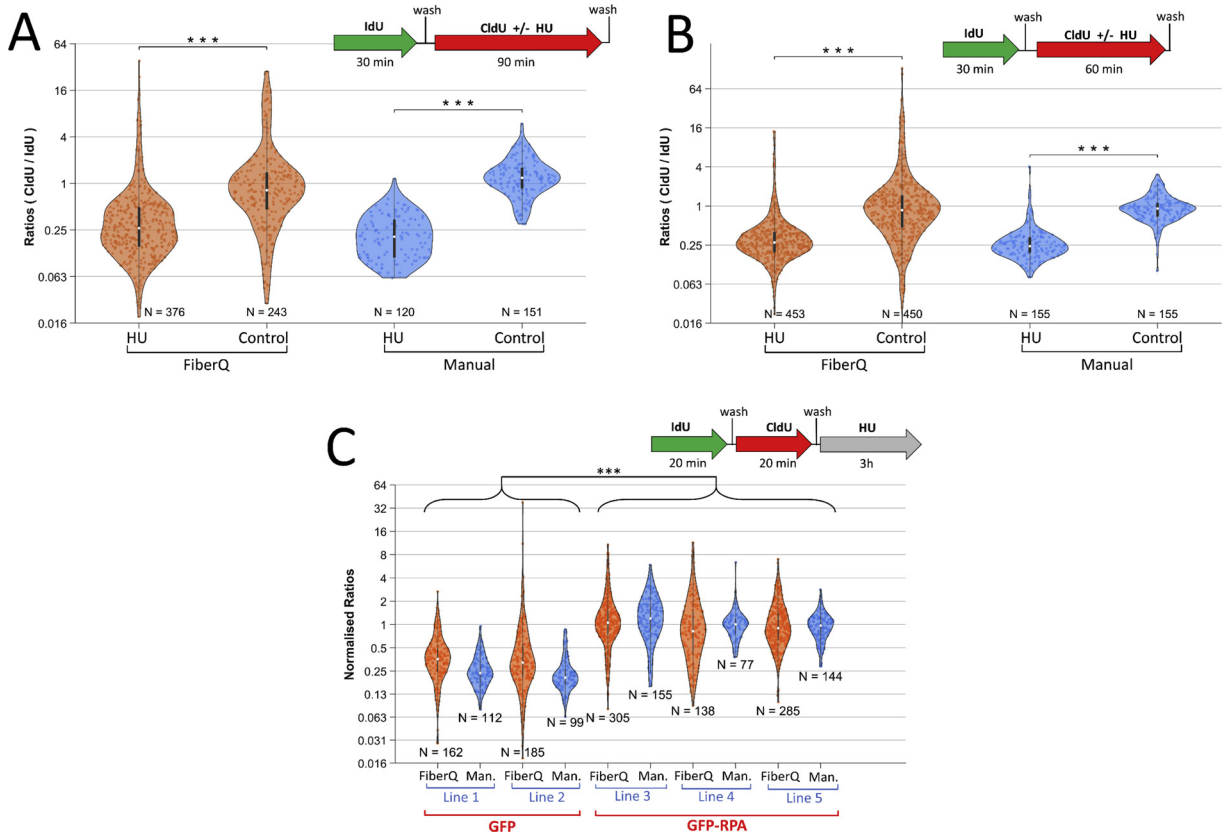

Fig. 5. Comparison between control and HU-treated samples. A, B- HU is incorporated during the second pulse. Orange and Blue violins are respectively the measures of our algorithm (FiberQ) and a manual user. Experiments in A and B only differ in the incubation time of CldU (90 min vs 60 min). ***: $p < 10^{-31}$ (Mann Whitney test). N: number of quantified fibers. C- Comparison between OV1946 cells overexpressing RPA vs GFP. Cells are incubated in HU-containing medium for 3 h after the CldU pulse. Distributions of ratios of the HU-treated sample normalized by the median of the control sample are displayed. FiberQ quantification is in orange, manual quantification in blue. ***: $p < 10^{-15}$ (Mann-Whitney test). N: number of quantified fibers.

ellipse that has the same normalized second central moments are larger than such threshold.

3.2.3. Edge detection method

The edge detection method aims at roughly segmenting DNA fibers. It is applied twice in our whole algorithm: the first time to compute I_{gray_1} during the pre-processing step with a default PSF (PSF = 2 pixels), and the second time to obtain I_{gray_2} during the *first fiber segmentation* step using the measured value of the PSF.

This edge detection method is made up of two parts. Briefly, we first use an edge detection that produces many false positives and then use a very selective edge detection that yields numerous false negatives. The output image is a combination between those two edge detection methods.

First, the input image (I_{gray_i} with $i = 1 \, or \, 2$) is convolved with a Laplacian of Gaussian (LoG) filter, whose standard deviation is set to EOP_4 PSF. Edge pixels, defined as zero-crossing pixels, are closed 8-connected contours delimiting fibers that we fill. The resulting binary image (Fig. 8B) called I_{LoG} contains a high number of false positive contours originated from noise and debris.

Next, a more selective contour detection is performed. We compute two smooth gradients of I_{gray_i} in x and y direction: $\nabla_x I_{gray_i}$, $\nabla_y I_{gray_i}$. Those smooth gradients are obtained by convolving I_{gray_i} with a 1D-derivative of gaussian ($\sigma = PSF$). Then, the gradient modulus M is calculated:

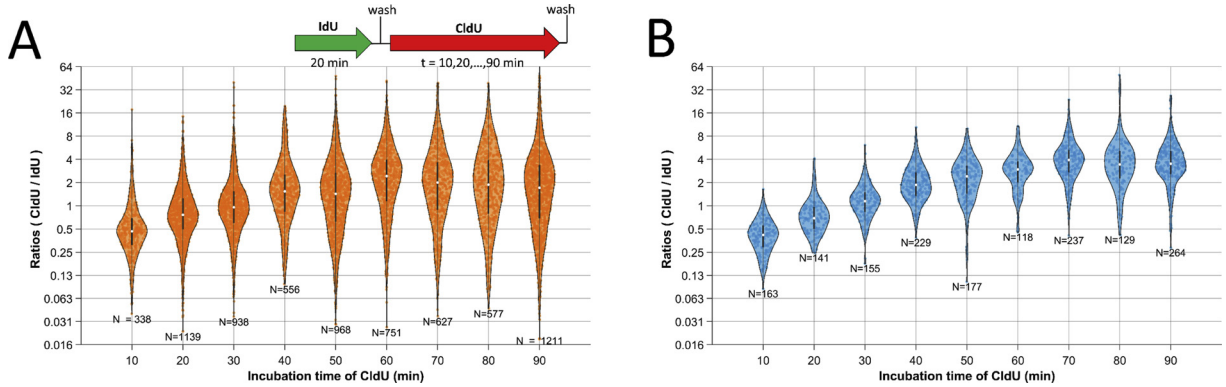

Fig. 6. Variation of the incubation time of CldU A- FiberQ quantification for CldU incubation time varying from 10 to 90 min. N: number of quantified fibers. B- Manual quantification for CldU incubation time varying from 10 to 90 min. N: number of quantified fibers.

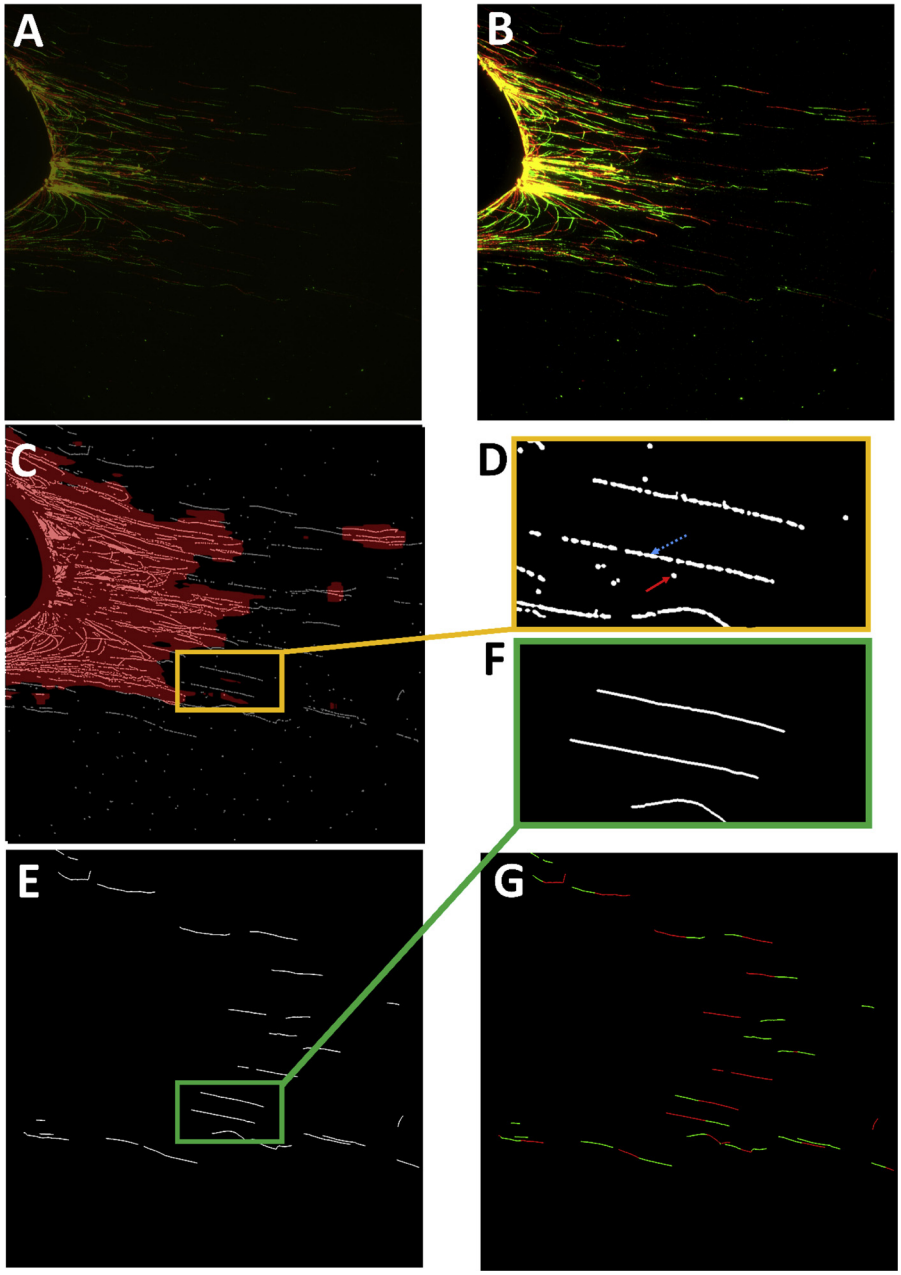

Fig. 7. From the raw image to the final segmentation. A- Raw image [I_1 : red channel (CldU), I_2 : green channel (IdU)]. B- Color enhancement after the pre-processing step [I_{1N}, I_{2N}]. C- Superposition of BW_{DNA2} in white (segmentation of DNA fibers with the tuned edge detection method) and unexploitable high density areas in red obtained by thresholding the local white pixel density ($d(x,y) < EOP_2$). D- Zoom of the yellow rectangle in C. DNA Fibers are fragmented. An example of blob and an example of strand are respectively flagged by a red full line arrow and a blue dotted arrow. E- $SKEL_{DNA}$: Result of the splicing method. F- Zoom of the green rectangle in E. Fragmented fibers have been spliced. G– Color assignment: analysis of the red and green channels of the color enhanced image (B).

$M = \sqrt{\nabla_x I^2_{gray_i} + \nabla_y I^2_{gray_i}}$. Finally, a pixel P of I_{gray_i} is a contour if it fulfills two conditions: (i) its gradient magnitude M(P) is bigger than Otsu's threshold t_{otsu} applied on M, and (ii) P is a maximum in the gradient direction $\theta = arctan(\frac{\nabla_y I_{gray}}{\nabla_x I_{gray}})$. Note that this selective contour detection is equivalent to Canny edge detection with both thresholds equal to t_{otsu}. The contours obtained with this selective method are morphologically dilated with a disk of diameter $\frac{PSF}{EOP_5}$. The resulting binary image is called I_{Canny}.

We combine these two binary images (I_{Canny} and I_{LoG}) by deleting all objects from I_{LoG} that have no intersection with I_{Canny}. The resulting image (BW_{DNAi}) is the output of the edge detection method.

3.2.4. Fiber splicing

The main drawback of this first segmentation (BW_{DNA2}) is the frequent fragmentation of DNA fibers (Fig. 7C, D). A splicing method is thereby necessary to reconnect portions of the same DNA fiber (Fig. 7E, F). Briefly, large objects of BW_{DNA2} are successively spliced with nearby objects if several continuity criteria (based on distance and fiber orientation) are fulfilled.

All objects of BW_{DNA2} are classed in two different groups: *Blobs* and *Strands* (Fig. 7D). A *Blob* is a small and compact object which can be modeled by an ellipse with a relatively low eccentricity. In BW_{DNA2}, even though many *blobs* are the consequence of noise in the original image, some of them are portions of a longer fiber. On the other hand, a *strand* is a longer curvilinear object that is a fraction or the totality of a DNA fiber. Practically, a *strand* fulfills two criteria:

(i) High eccentricity e: $e > \frac{\sqrt{15}}{4} \approx 0.968$ (ie. the ratio of the large axis of the ellipse over the small axis has to be higher than 4)
(ii) Low solidity s: $s = \frac{object\ Area}{area\ of\ its\ convex\ hull} < 0.7$.
(iii) Minimum length l: $l > EOP_6\ PSF$

All other objects are *blobs*.

Strands are skeletonized by applying successive morphological erosions. For each *strand*, we store their pixel coordinates (x, y), the position of their two endpoints (EP) and the orientation of the tangents at

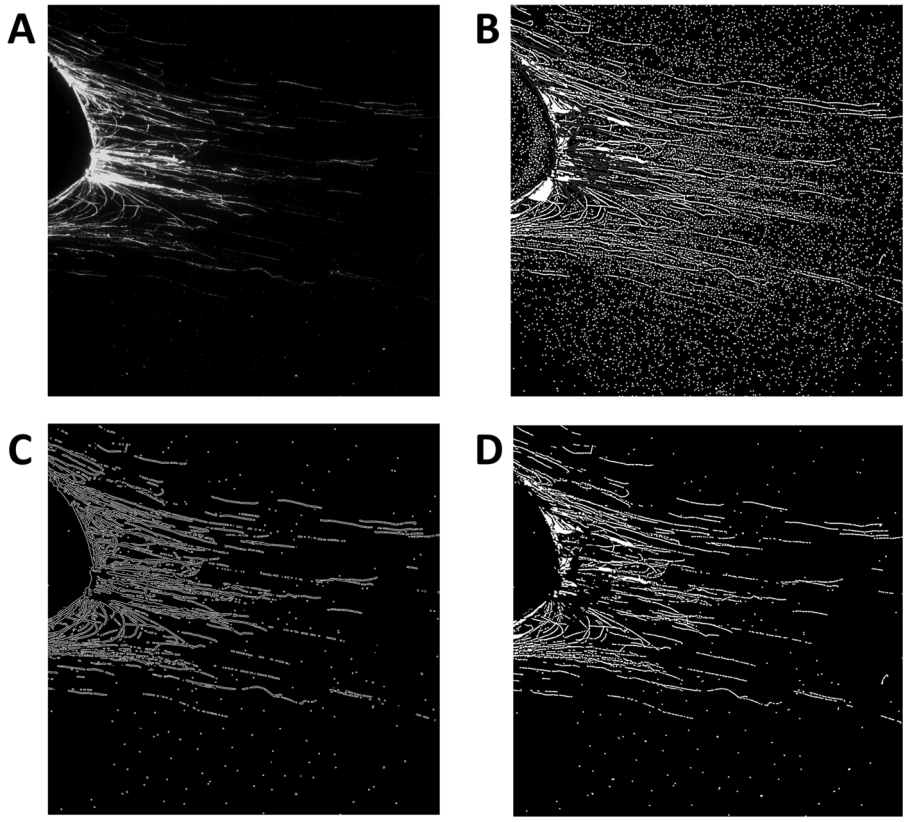

Fig. 8. Edge detection method. A– I_{gray2} : combination of the normalized intensity channel : $I_{gray2} = (I_{1N} + I_{2N})/2$. B- I_{LoG} : First edge detection [Laplacian of Gaussian (LoG)]. C– I_{Canny} : Second edge detection. D– BW_{DNA2} : Combination of I_{LoG} and I_{Canny} : Objects of I_{LoG} that intersect a white pixel of I_{Canny} are kept.

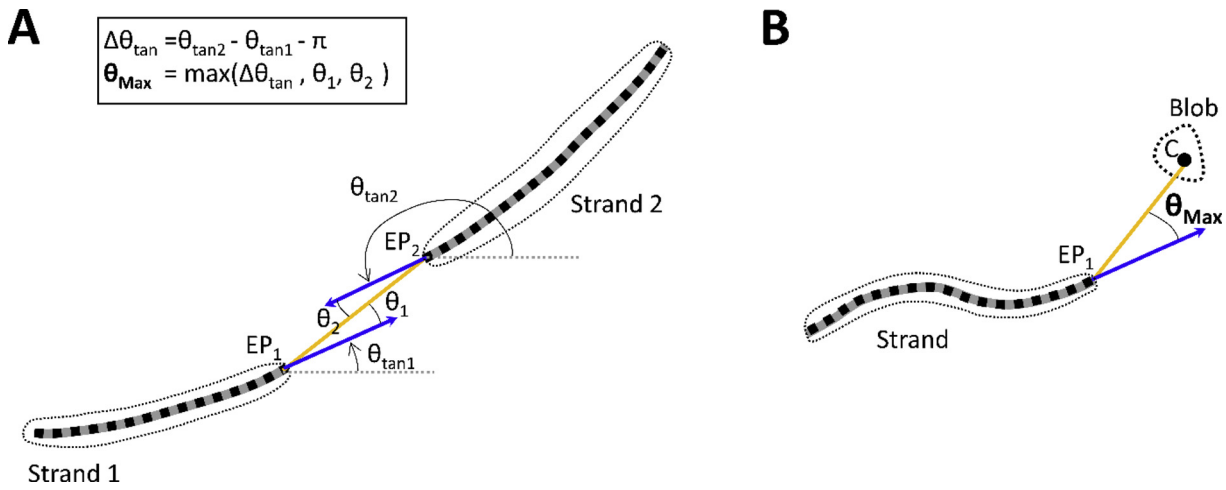

Fig. 9. Splicing parameters. A- Connection between 2 strands. θ_{tan1} and θ_{tan2} are the orientation of the tangents at each endpoint (EP1, EP2) for each strand. θ_1 and θ_2 are the angles between the tangents and the connection segment [EP1, EP2] in yellow. θ_{Max} is the maximum of $\Delta\theta_{tan}$, θ_1, θ_2. B– Connection between a strand and a blob. θ_{Max} is the angle between the tangent and the connection segment [EP1,C] in yellow.

each one of the two endpoints (Fig. 9A). Such two tangents are computed after smoothing the coordinates (x,y) using a local least square regression with a 1st degree polynomial model spanning a length of EOP_7 *PSF*.

For each *blob,* only the coordinates (x_c, y_c) of its centroid are stored.

With all this data, a graph G is built (Fig. 10) where each object (*strand* or *blob*) is a node. The edges of G link objects whose EP are separated by a distance *D* inferior to D_{Max} ($D_{Max} = EOP_8$ *PSF*). Actually, an edge symbolises a *potential* connection between the EPs of two different objects. Each edge is characterised by a doublet *edge* = (*s*, *D*) where *s* is a score based on the angular continuity of the *potential* connection and *d* is the distance between the two objects. Those two parameters (*s* and *D*) will be used to rank all potential connections from a given strand.

Calculation of the score *s* :

- Connection between two *strands*:

Fig. 9A shows the different parameters used in the calculation of the score *s*. θ_{EP1} and θ_{EP2} are the orientations of the tangents for both endpoints EP. θ_1 and θ_2 are the angles between the tangents and the segment connecting EP1 to EP2. We define $\Delta\theta_{tan} = \theta_{tan2} - \theta_{tan1} - \pi$. If the connection is continuous, $\Delta\theta_{tan}$, θ_1 and θ_2 should be minimal.

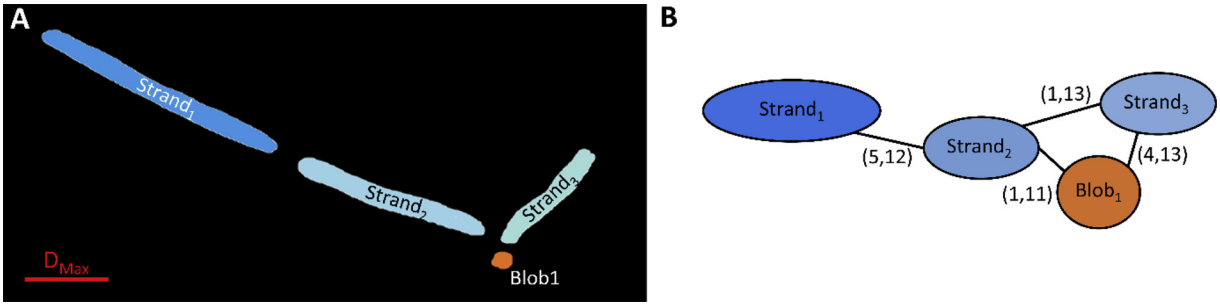

Fig. 10. Graph of Strands and Blobs. A - Image of 3 strands (in blue) and one blob (in orange) B— Objects whose endpoints are separated by less than d_{Max} are linked by an edge (s,D). The angular score s is an integer between 1 and 5. The parameter D is the distance between the endpoints.

Table 4
Score s with respect to θ_{Max}. The lower is θ_{Max}, the higher the angular score s is.

Value of θ_{Max} (in degree)	Angular score s
$\theta_{Max} \leq 2$	5
$2 < \theta_{Max} \leq 5$	4
$5 < \theta_{Max} \leq 10$	3
$10 < \theta_{Max} \leq 15$	2
$15 < \theta_{Max}$	1

Table 5
List of Experimentally Optimized Parameter. All these parameters are set as multiplication factors of the PSF of the imaging system.

Parameter name	Value	Description
EOP_1	8	Sets the size of the area used to calculate local fiber density
EOP_2	0.21	Density threshold
EOP_3	3.6	Sets the maximum fiber width
EOP_4	1	Sets the kernel size of the Laplacian of Gaussian filter
EOP_5	2	Sets the length of the structural element used for dilating Canny's edges
EOP_6	9	Set the minimum length for a strand
EOP_7	6	Sets the spanning length of the 1st degree polynomial for smoothing fiber skeletons
EOP_8	11	Sets the maximum splicing distance
EOP_9	15	Sets the minimum length for fibers after splicing
EOP_{10}	6	Sets the minimum length of a pulse section within a fiber

We also define $\theta_{Max} = \max(\Delta\theta_{tan}, \theta_1, \theta_2)$. The value of s from 1 to 5 is given according to the value of θ_{Max} (see Table 4). The advantage of splitting the connections into five discrete classes rather than assigning a continuous score value is that potential connections with similar θ_{Max} have equal scores. The distance parameter D will be used to rank the connections with the same score s.

- Connection between a *strand* and a *blob*:
  Fig. 9B illustrates this configuration. Here, θ_{Max} is the angle between the tangent of the strand and the segment that links the endpoint and the centroid of the *blob*.
- Connection between two *blobs*:
  As the splicing process begins from a *strand* and iteratively merges nearby objects (*blobs* or *strands*), we never consider *blob-blob* edges.

The following splicing procedure is applied iteratively from the longest to the smallest strand of the graph. Let str_i be the i^{th} strand processed. First, all edges connected to str_i are sorted in decreasing order of score s. Strands with equal s are ranked in decreasing order of distance D. Let's note $edge_1 = (s_1, D_1)$ the first edge of the ranking. This edge links str_i to another object that we call obj_1. We splice the strand str_i to obj_1 if two conditions are fulfilled: (i) $s_1 \leq 4$, (ii) obj_1 is not connected to a better edge (i.e. an edge whose score s is strictly higher than s_1, or $s = s_1$ and $D < D_1$). If those two conditions are not fulfilled for $edge_1$, we try with the following edges in the ranking.

If a candidate obj_k meeting those requirements is found, str_i merges with obj_k ($str_i \longleftarrow str_i \cup obj_k$): (i) str_i is linked to obj_k by a straight line, (ii) the node of str_i in the graph merges with the node of obj_k. The new node preserves the links of str_i and obj_k (except the link between both objects) but with updated value (s, D).

At the end of this splicing step, we obtain a binary image containing the skeletons of DNA fibers: $SKEL_{DNA}$ (Fig. 7E). All skeletons with pixels in high-density areas (as defined before in the *first fiber segmentation*) are deleted. We also remove skeletons whose length is inferior to *lmin* ($l_{min} = EOP_9\ PSF$) because they often are artefacts due to debris or non-specific staining.

3.2.5. Color assignment

Once DNA fibers are segmented and skeletonized, we estimate the color at each pixel of the skeletonized fibers by comparing intensities of each channel: I_{1N} and I_{2N}. The objective is to convert *color intensities* (doubles between 0 and 1) to a *color label* (e.g. IdU or CldU).

Each foreground pixel of $SKEL_{DNA}$ is assigned a pair of intensities referring to the normalised fluorescence of the two nucleotide analogs in the vicinity of the pixel. To do so, each one of the two normalized intensity images I_{1N} and I_{2N} is convolved with a Gaussian kernel of standard deviation PSF, and multiplied element-wise by $SKEL_{DNA}$. Following the pixels of each skeleton from one endpoint to the other, it is possible to define two color intensity profiles: S_1 (intensity profile of the first nucleotide analog) and S_2 (intensity profile of the second one).

We compute the color difference function as $\Delta S = S_1 - S_2$, which can be interpreted as the difference between the normalized fluorescence of both channels. The zero-crossing of the function ΔS are computed to partition fibers into segments of a predominant nucleotide analog. Colors are assigned in two steps:

1 All Segment where $|\text{mean}(\Delta S)|$ is larger than a threshold (empirically set to 7%), are assigned to the predominant nucleotide analog color.
2 Each remaining section is assigned as follows:
3 When the segment is surrounded by neighbours of the same color, the same color is assigned.
4 When the segment is surrounded by neighbours of different colors, the segment is split in halves and colors are assigned to match the color of neighbours.
5 When the segment is located at the end of a fiber, the neighbour segment color is assigned.

Thus, each skeleton is partitioned in sections of different colors. If some sections are too small ($length < EOP_{10}.\ PSF$) they are considered as mistakes: the *color label* of those sections is inverted so that they are merged with their two neighbours.

4. Discussion

Fluorescent imaging of DNA fibers is widely used to study the dynamics of RF progression, as a proxy for several aspects of genomic stability and replicative stress. The vast majority of studies using this technology are based on manual segmentation of fibers, using simple image processing software to facilitate record keeping and annotation. Our results demonstrate a significant degree of variability when manually measuring DNA fibers, which compromises data reproducibility and renders analyses prone to bias. The automated segmentation method that we present here (FiberQ) is devoid of subjectivity and enables rapid analysis of large image databases, thereby increasing statistical power.

The open-source implementation we provide was programmed using Matlab. We also made available a free compiled version of the code which can be used by investigators without programming experience or access to this commercial language. The output of FiberQ consists of four images outlining the results and a spreadsheet with single fiber details. Images show 1) which of the fibers where chosen for analysis, 2) skeleton versions of such fibers with a tag that allows identification in the spreadsheet, 3) the ratio (second/first pulse) for each one of them and 4) high density areas not considered in the analysis. The spreadsheet contains all necessary information for statistical analysis: the file name of origin, the fiber tag, the combination of colors found, the length of each nucleotide label.

The proposed framework is robust and can be adapted to various experimental conditions. Indeed, the use of the Point Spread Function, which is computed for each image using fiber width auto-calibrates free parameters of the algorithm. In addition, this characteristic metric enables removal from the analysis of anomalous objects such as small nonspecific staining spots, unusually large fibers and areas characterized by excessively high fiber density. Moreover, even if the calibration of all parameters (thresholds, splicing distance, filter size) is done automatically, users can also fine tune them manually. The three most important parameters that may need adjustment to experimental conditions are: the maximum splicing distance used to connect fragmented fibers (EOP_8), the fiber density threshold used to remove unexploitable clusters of fibers (EOP_2) and the maximum fiber length after splicing. We also provide some useful confidence metrics available in the output spreadsheet for each segmented fiber, which consider their density, maximum splicing distances and maximum splicing angle.

FiberQ was tested on a database of different images of varying quality, signal to noise ratio, or fiber length which were acquired from two different microscopes. Segmenting fibers in such images represents a complex task because (i) fiber fluorescence is not homogeneous, (ii) fibers are split in several segments, tangled in clusters, and/or mixed with nonspecific stained objects. We note that our algorithm was implemented specifically for the analysis of DNA fibers and was not optimized for DNA combing images. In our experience, the latter generally present a much larger number of split DNA segments, which renders adequate joining and segmentation of labeled tracks challenging. Further optimization of the FiberQ algorithm will therefore be required to make it reliable for DNA combing analysis.

We showed in Table 3 that 27% of fibers segmented by FiberQ were not segmented by any other user. After close examination, we found that this set of fibers contains good fibers ignored by users, fibers whose configuration is too ambiguous to be taken into account, and erroneous fibers. It is possible to reduce the number of ambiguous and erroneous fibers by imposing more stringent constraints on the algorithm, i.e. lowering the density threshold, the maximum splicing angle and removing fibers with too high splicing distances. Since these three parameters are provided in the output spreadsheet, users can simply filter unwanted fibers *a posteriori*.

We also observed that in the case of very long fibers split into a high number of segments or blobs, our algorithm can yield aberrant results. This problem was highlighted in the experiment where the incubation time is varied from 10 to 90 min, for which both manual and automatic segmentation fail to detect the extension of fibers for long incubation times (i.e. when fibers are very long; Fig. 6). Examination of the images revealed that in the case of FiberQ, the above problem is the consequence of mistakes during the splicing stage of the algorithm which fails to connect all strands and blobs belonging to very long fibers. As second pulse tracks are expected to be much longer than contiguous first pulse ones, their ratio is often underestimated. While our results suggest that such images also cannot be reliably quantified manually, artifacts and mistakes appear to be worse with FiberQ. We also note that we observed larger standard deviation in FiberQ measurements as compared to manual quantification.

To improve accuracy, we propose that two main experimental parameters should be optimized for automated analysis: (i) incubation duration should be kept relatively short to reduce fiber length, and ii) dilution of fibers on the slide might reduce clustering of fluorescent DNA molecules.

In summary, FiberQ is an algorithm that greatly facilitates investigations on the dynamics of DNA replication by automatically measuring the length of fluorescently labelled DNA fibers. Contrary to manual techniques, FiberQ is devoid of inter/intra user variability. Our algorithm should therefore be useful to reduce the tediousness, bias, and poor reproducibility associated with manual quantification of DNA fiber length.

Availability

FiberQ is an open-source collaborative initiative available in the GitHub repository (https://github.com/pierreghesqui/FiberQ).

Funding

This work was supported by grants from the Canadian Institutes of Health Research and the Natural Science and Engineering Research Council of Canada to SC, HW and ED. SC and HW hold salary awards from the Fonds de Recherche du Québec – Santé.

Author contributions

Conceptualization: PG. and SC. Software: PG. Materials, experiments and resources: AE, EF, MMQ, JM, FB, FC, ED and HW. Writing, review and editing: PG, ED, HW and SC.

Acknowledgements

The authors thank Grant Brown and David Gallo from University of Toronto for images provided for testing.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.dnarep.2019.01.003.

References

[1] M.K. Zeman, K.A. Cimprich, Causes and consequences of replication stress, Nat. Cell Biol. 16 (2014) 2–9.
[2] J.L. Alexander, T.L. Orr-Weaver, Replication fork instability and the consequences of fork collisions from rereplication, Genes Dev. 30 (2016) 2241–2252.
[3] C. Ribeyre, J. Lopes, J.B. Boule, A. Piazza, A. Guedin, V.A. Zakian, J.L. Mergny, A. Nicolas, The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo, PLoS Genet. 5 (2009) e1000475.
[4] W. Gan, Z. Guan, J. Liu, T. Gui, K. Shen, J.L. Manley, X. Li, R-loop-mediated genomic instability is caused by impairment of replication fork progression, Genes Dev. 25 (2011) 2041–2056.
[5] B. Liu, M.L. Wong, R.L. Tinker, E.P. Geiduschek, B.M. Alberts, The DNA replication fork can pass RNA polymerase without displacing the nascent transcript, Nature 366 (1993) 33–39.
[6] J. Poli, O. Tsaponina, L. Crabbe, A. Keszthelyi, V. Pantesco, A. Chabes,

A. Lengronne, P. Pasero, dNTP pools determine fork progression and origin usage under replication stress, EMBO J. 31 (2012) 883–894.
[7] M. Xie, Y. Yen, T.K. Owonikoko, S.S. Ramalingam, F.R. Khuri, W.J. Curran, P.W. Doetsch, X. Deng, Bcl2 induces DNA replication stress by inhibiting ribonucleotide reductase, Cancer Res. 74 (2014) 212–223.
[8] T. Hideshima, G. Tonon, K.C. Anderson, The oncogene *MYC* triggers replicative stress and DNA damage in multiple myeloma, Blood 122 (2013) 3114-3114.
[9] S. Negrini, V.G. Gorgoulis, T.D. Halazonetis, Genomic instability–an evolving hallmark of cancer, Nat. Rev. Mol. Cell Biol. 11 (2010) 220–228.
[10] Y. Ruzankina, C. Pinzon-Guzman, A. Asare, T. Ong, L. Pontano, G. Cotsarelis, V.P. Zediak, M. Velez, A. Bhandoola, E.J. Brown, Deletion of the developmentally essential gene ATR in adult mice leads to age-related phenotypes and stem cell loss, Cell Stem Cell 1 (2007) 113–126.
[11] J. Nieminuszczy, R.A. Schwab, W. Niedzwiedz, The DNA fibre technique - tracking helicases at work, Methods 108 (2016) 92–98.
[12] D.A. Jackson, A. Pombo, Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells, J. Cell Biol. 140 (1998) 1285–1295.
[13] A. Ray Chaudhuri, E. Callen, X. Ding, E. Gogola, A.A. Duarte, J.E. Lee, N. Wong, V. Lafarga, J.A. Calvo, N.J. Panzarino, et al., Replication fork stability confers chemoresistance in BRCA-deficient cells, Nature 535 (2016) 382–387.
[14] F. Belanger, E. Fortier, M. Dube, J.F. Lemay, R. Buisson, J.Y. Masson, A. Elsherbiny, S. Costantino, E. Carmona, A.M. Mes-Masson, et al., Replication protein A availability during DNA Replication stress is a major determinant of cisplatin resistance in ovarian cancer cells, Cancer Res. (2018).
[15] Y.P. Wang, P. Chastain, P.T. Yap, J.Z. Cheng, D. Kaufman, L. Guo, D.G. Shen, Automated DNA fiber tracking and measurement, 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, (2011).
[16] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (1986) 679–698.
[17] D. Marr, E. Hildreth, Theory of edge detection, Proc. R. Soc. Lond. B Biol. Sci. 207 (1980) 187–217.
[18] J.W. Yarbro, Mechanism of action of hydroxyurea, Semin. Oncol. 19 (1992) 1–10.