Bioinformatics, 38(18), 2022, 4395–4402
https://doi.org/10.1093/bioinformatics/btac533
Advance Access Publication Date: 26 July 2022
Original Paper

OXFORD

Bioimage informatics

Automatic DNA replication tract measurement to assess replication and repair dynamics at the single-molecule level

Longjie Li[1,2,†], Arun Mouli Kolinjivadi[3,†], Kok Haur Ong[1,2], David M. Young[1,4], Gabriel Pik Liang Marini[2], Sock Hoai Chan[5,6], Siao Ting Chong[3], Ee Ling Chew[5], Haoda Lu[2], Laurent Gole[1,*], Weimiao Yu [1,2,*] and Joanne Ngeow[3,5,6,*]

[1]Institute of Molecular and Cell Biology, A*STAR, Singapore 138673, Singapore, [2]Bioinformatics Institute, A*STAR, Singapore 138671, Singapore, [3]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore, [4]Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA 94158, USA, [5]Cancer Genetics Service, Division of Medical Oncology, National Cancer Centre, Singapore 169610, Singapore and [6]Oncology Academic Clinical Program, Duke-NUS Medical School Singapore, Singapore 169857, Singapore

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Hanchuan Peng

Received on February 28, 2022; revised on June 23, 2022; editorial decision on July 12, 2022; accepted on July 22, 2022

Abstract

Motivation: DNA fibre assay has a potential application in genomic medicine, cancer and stem cell research at the single-molecule level. A major challenge for the clinical and research implementation of DNA fibre assays is the slow speed in which manual analysis takes place as it limits the clinical actionability. While automatic detection of DNA fibres speeds up this process considerably, current publicly available software have limited features in terms of their user interface for manual correction of results, which in turn limit their accuracy and ability to account for atypical structures that may be important in diagnosis or investigative studies. We recognize that core improvements can be made to the GUI to allow for direct interaction with automatic results to preserve accuracy as well as enhance the versatility of automatic DNA fibre detection for use in variety of situations.

Results: To address the unmet needs of diverse DNA fibre analysis investigations, we propose *DNA Stranding*, an open-source software that is able to perform accurate fibre length quantification (13.22% mean relative error) and fibre pattern recognition ($R > 0.93$) with up to six fibre patterns supported. With the graphical interface, we developed, user can conduct semi-automatic analyses which benefits from the advantages of both automatic and manual processes to improve workflow efficiency without compromising accuracy.

Availability and implementation: The software package is available at https://github.com/lgole/DNAStranding.

Contact: yu_weimiao@bii.a-star.edu.sg or lgole@imcb.a-star.edu.sg or joanne.ngeow@ntu.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Fluorescent-based DNA fibre assays are a type of biophysical technique developed as a means to visualize individual DNA replication fork (RF) dynamics at the molecular level (Nieminuszczy *et al.*, 2016). This involves the labelling of newly synthesized DNA strands *in vitro* using nucleotide analogues, which are stained with fluorophore conjugated antibodies as shown in Figure 1A. The results of this are fluorescently stained daughter strands that can be visualized using a variety of fluorescence imaging techniques, such as fluorescence microscopy (Green *et al.*, 2015) based on the Stokes shift (which represents the difference in absorption and emission wavelength maxima) of the specific fluorophores used (Santos *et al.*, 2021), as shown in Figure 1B and C. The images obtained from visualization are then analysed and individual RF patterns are characterized (Fig. 1D), counted or rejected according to experimental design.

Single-molecule visualization of DNA strands allows for the measurement and study of replication structures, such as RFs, which are the points of separation between single strands of DNA undergoing replication (Halliwell *et al.*, 2020). Such fluorescence-based DNA fibre assays have advanced our basic understanding of DNA

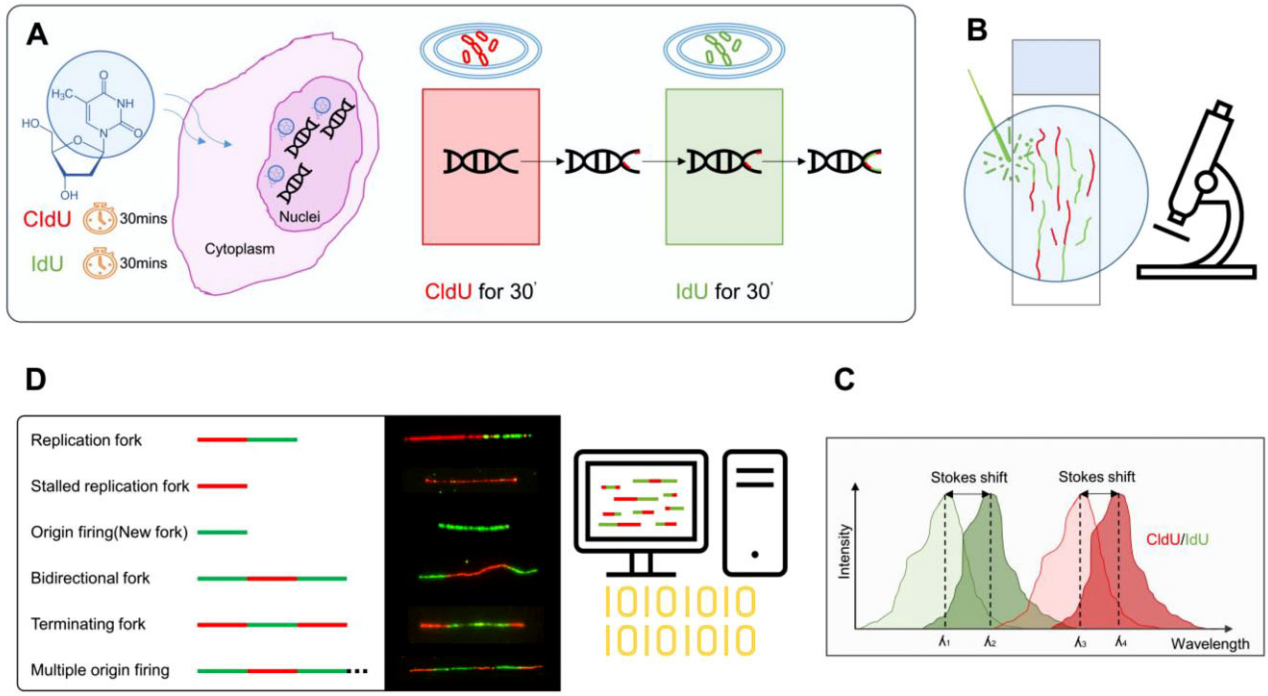

Fig. 1. (**A**) Illustration of the DNA labelling process used in this article. LCLs are incubated with thymidine analogues CIdU and IdU consecutively for 30 min each. During the incubation period, CldU and IdU are incorporated into newly synthesized daughter strands in cells undergoing DNA replication. Cells are lysed and genomic DNA is spread on the slide and stained with antibodies conjugated with fluorophores of different excitation and emission wavelengths. (**B**) Illustration of the fluorescence visualization process. An excitation light, which is channelled into a specific wavelength, is used to reveal DNA fibres that have been fluorescently labelled. (**C**) Depicts the detection of the signal from the fluorophores conjugated to the antibodies that are bound to CIdU (detected at 594 nm) and IdU (488 nm), respectively. The Stokes shift represents the difference between the absorption and the emission wavelengths. (**D**) Illustration of the computational process where the images captured from fluorescence visualization are processed by DNA Stranding software. Individual DNA fibres are automatically detected and categorized to one of six patterns, which are indicative of different structural features of the replicating DNA fibre as shown

replication dynamics and sheds light on chemotherapeutic responses. Analysing the newly synthesized DNA at single-molecule resolution has been previously proposed (Lopes, 2009) to have potential clinical applications, such as in predicting chemotherapeutic response in homologous recombination (HR) deficient cancers using organoid model systems (Hill *et al.*, 2018; Lopes, 2009; Quinet *et al.*, 2017). Visualization of RFs upon exposure to genotoxic agents can help elucidate DNA repair mechanism at the molecular level by quantification of RF speed/velocity, frequency of replication origins initiation and frequency of stalled RFs. Another potential application of DNA fibre assays is to quantify the functional competency of DNA repair in cells harbouring inherent genetic mutations in DNA replication and repair genes (Kolinjivadi *et al.*, 2020). With the rapid adoption of next-generation sequencing for clinical genetic testing as well as human genomics research, large numbers of novel variants with uncertain significance are being discovered. This represents a major challenge for the clinical implementation of genomic medicine in the clinic as it limits the clinical actionability and medical management (Federici and Soddu, 2020). Presently, functional evaluation of HR gene variants relies predominantly on double-strand break repair reporter and DNA damage sensitivity cell survival assays. DNA fibre assays complement these existing assays by characterizing the functional impact of novel variants. However, the existing automated solution for DNA fibre assay quantification, FibreQ and FibreAI (Ghesquière *et al.*, 2019; Mohsin *et al.*, 2020) are unable to account for more than two fibre patterns, leaving manual characterization the only viable way to quantify those assays.

DNA fibre assays have seen widespread use in understanding the role of DNA replication and repair factors (Kolinjivadi *et al.*, 2017, 2020; Ray Chaudhuri *et al.*, 2016; Zeman and Cimprich, 2014). This technique has potential application towards genomic medicine, such as in the characterization of mitochondrial diseases (Zereg *et al.*, 2020) or analysis in human pluripotent stem cell research (Halliwell *et al.*, 2020). The technique has garnered specific attention in recent years with the emergence of replication stress dynamics as a central theme in the understanding the pathogenicity of chromosome instability disorders (Quinet *et al.*, 2017). In this work, we perform a series of contemporary experiments in the investigation of RF speed analyses and efficiency of stalled fork recovery to validate our proposed software, *DNA Stranding*, and demonstrate the effectiveness of the semi-automatic method to enhance the workflow of DNA fibre analysis.

Automated DNA fibre detection makes use of the design parameters (e.g. minimum fibre length) to assist in the analysis stage by counting strands that fit within the experimental objective to some certain specificity. Depending on the size of a study and operator skill in the erstwhile manual analysis, automated fibre detection can save anywhere between days to weeks of time in the characterization process (Ghesquière *et al.*, 2019); there is also an elimination of operator-dependent bias, which improves the consistency of results. However, relying fully on automated detection is not without certain caveats, that is the rigidity of parameters used for detection, which may fail to count a certain number of fibres that a researcher may deem useful to the investigation or fail to reject ones that do not fit the objective; as such, the total count of valid fibres obtained from an image by existing automated programmes are often less in comparison with a manual operator and moreover, the count summaries belie the number of false positive results therein which remain unknown unless a manual characterization and count is performed. Although the loss of human input confers with it certain advantages, such as the elimination of operator-dependent bias (Ghesquière *et al.*, 2019; Mohsin *et al.*, 2020), it also eliminates the benefits of human judgement to identify novel features, which are useful in studies, such as in Zereg *et al.* (2020) and Stanojcic *et al.* (2016), or rule out fibres detected that are incompatible with the objective of a specific research. In this work, we show that existing automated fibre detection methods may introduce false positive results that are noticeable only if a manual comparison is made.

This defeats the purpose of automatic detection to begin with. As DNA fibre analysis is a general applied research method, the introduced type 1 error rate and trade-off with sensitivity by fully automatic detection may preclude its use in applications where the highest possible accuracy of results is sought, and a meticulous manual quantification is preferred.

To address the issues inherent with fully automatic detection, we introduce an enhanced automatic workflow allowing human interaction and have developed *DNA Stranding*. *DNA Stranding* is a computational software that builds upon the effectiveness of automated detection by allowing the operator to amend the inaccuracies generated from the fully automatic tract measurement. The inaccuracies accounts for the outlying fibres, as well as the fibres that goes undetected by the automatic detection. Unlike previous fibre detection algorithms, *DNA Stranding* is able to recognize and categorize up to six DNA fibre patterns that are useful in RF analyses in various DNA fibre assay investigations (Quinet *et al.*, 2017; Stanojcic *et al.*, 2016; Zereg *et al.*, 2020). Further improvements have been made to cater towards the needs of current research in DNA fibre analysis by enhancing the accuracy of fibre length measurements in relation to manual annotations serving as ground truth. In this work, we detail a complete analysis pipeline using a relevant demonstration and compare its effectiveness with both the manual process as well as an existing open-source automatic fibre detection software, FiberQ (Ghesquière *et al.*, 2019).

2 Materials and methods

Accurately measuring fibre lengths is critical in a variety of investigations involving DNA fibre analyses, such as in RF speed analyses, efficiency of stalled fork recovery and measurement of the inter-origin distance, etc. We developed the DNA fibre assays using Lymphoblastoid Cell Lines (LCLs) for our study. The details of our analysis pipeline and software developments are provided in this section. A graphical user interface (GUI) enabling manual correction is also implemented.

2.1 DNA fibre assay preparation

2.1.1 LCLs generation

LCLs derived from human peripheral blood B cells were generated using the Epstein–Barr Virus (EBV)-immortalization method as described in Frisan *et al.* (2001). To isolate peripheral blood mononuclear cell (PBMC) by density gradient centrifugation, 7 ml of human peripheral blood was diluted with phosphate buffered saline (PBS) and slowly layered above 12 ml Ficoll-Paque Plus (Cytiva, 17144003) followed by centrifugation at 450×*g* for 25 min at room temperature (RT) with zero brake. The buffy coat containing PBMC was carefully aspirated and washed twice with PBS by centrifugation at 300×*g* for 10 min at RT. The PBMC was resuspended in 5 ml EBV media and transferred into a T25 flask followed by adding 2 ml of RPMI 1640 media containing Cyclosporine-A (Sigma-Aldrich, C1832). The cells were maintained at 37°C in a humidified 5% CO_2 atmosphere. Feeding cells with complete RPMI 1640 media supplemented with 20% Foetal Bovine Serum and Antibiotic-Antimycotic was done every 7 days for four cycles before the cells were ready for experiments.

2.1.2 RF analysis by DNA fibre assay and image acquisition

For each LCL sample, ∼100 000 cells were labelled with thymidine analogues, 20 µM 5-chloro-2′-deoxyuridine (CldU, Sigma-Aldrich, C6891) for 30 min followed by 200 µM 5-Iodo-2′-deoxyuridine (IdU, Sigma-Aldrich, I7125) for 30 min and subsequently washed thrice with PBS. Labelled cells were mixed with lysis buffer in 1:1 ratio before mounted on SuperFrost Plus slides (Epredia, J1800AMNZ). Cellular DNA were spread by slanting the slides and fixed in 3:1 methanol/acetic acid for 30 min, followed by denaturation with 2.5 N hydrochloric acid for 40 min and subsequently blocked with 2% BSA in PBS-0.1% Tween20 for 40 min at RT. Slides were washed five times with PBS followed by 2 h incubation with 1:200 anti-BrdU/IdU (Becton Dickinson, 347580, mouse) and 1:200 anti-BrdU/CldU (Abcam, ab6326, rat) antibodies. Subsequently, slides were washed five times with PBS and finally stained with anti-mouse AlexaFluor-488 (Abcam, 150157) and anti-rat AlexaFluor-594 (Abcam, 150116) conjugated secondary antibodies for 1 h at RT.

Labelled DNA fibres were imaged with 63× objective (0.103 um/pixel) using a Zeiss Inverted Fluorescence Live Cell Microscope AO7. The experiments were performed in the presence of replication stress inducing agents, such as HydroxyUrea (HU, Sigma-Aldrich, H8627) and PARP inhibitor (PARPi) (Talazoparib BMN-673, MedChem Express), as controls to assess the differences in RF speed.

2.2 DNA Stranding and fibre analysis

Figure 2 describes the DNA Stranding analysis pipeline (Fig. 2A) and the accompanying relationships with image processing flow (Fig. 2B–F). The overall DNA Stranding analysis pipeline consists of four steps: (i) a signal enhancement process is first applied to the DNA fluorescence image for fibre detection; (ii) the fibres are extracted by a dedicated two-step fibre detection scheme (Frangi filtering followed by Gaussian mixed model-based segmentation); (iii) a gap-bridging process is deployed to reconstruct the fragments from a long fibre; and (iv) a robust colour encoding scheme is designed to convert the fibre intensity to a colour code (nucleotide analogue, CldU or IdU). A detailed description of each step is provided as follows.

2.2.1 Signal enhancement

As shown in Figure 2B and C, a signal enhancement (normalization) process was used to boost automatic fibre detection. Let I_R and I_G denote the channel corresponding to the first nucleotide analogue (CldU, red) and second nucleotide analogue (IdU, green), respectively. Based on qualitative assessment, the pixel intensity spectrum of image I_R is linearly stretched by saturating 0.02% of the top (brightest) pixels and suppressing 3% of the bottom (darkest) pixels to zero. The enhanced image is called I_{RE}. Then a histogram transformation process is applied to I_G to produce a new image I_{GE} whose histogram is approximately matched with the histogram of I_{RE}. The enhanced images I_{RE} and I_{GE} are then combined to form the enhanced true colour fibre image (Fig. 2C and D).

2.2.2 Fibre detection

The greyscale fibre image used for fibre detection is obtained by calculating the mean value of I_{RE} and I_{GE} ($I_{RE}/2 + I_{GE}/2$). The greyscale fibre image is made up of three parts: background, bright fibre and dark fibre. The background is the dark region of the image, consisting of pixels having the lowest intensity values. The bright fibre refers to the tubular structure having the highest intensities. And the

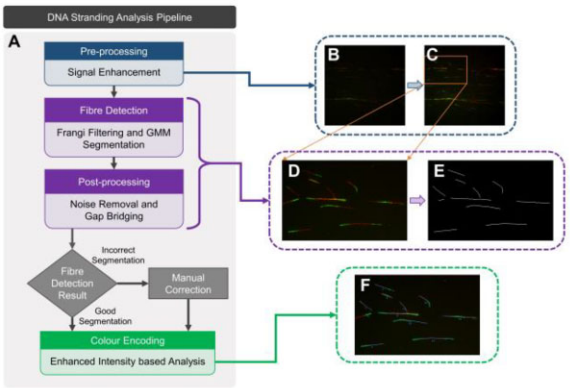

Fig. 2. (**A**) DNA Stranding software workflow. (**B**) Original DNA fluorescence image, where the first channel represents the first nucleotide and the second channel represents the second one. (**C**) Enhanced DNA fluorescence image. (**D**) Zoom-in of the selected region in (C). (**E**) Automatic fibre detection result. (**F**) Fibre detection result overlaid with the enhanced image

dark fibre refers to the tubular structure having relatively low intensities, which makes it difficult to be distinguished from background. To address this, we designed a two-step fibre detection scheme that considers both morphology and intensity features.

First, the multiscale filter proposed in Frangi *et al.* (1998) is applied to greyscale fibre image to enhance the fibre structure in it. Second, the pixel values of the output fibre-enhanced image are modelled as a combination of three groups of Gaussian distributed data representing background, dark fibre and bright fibre, respectively. A Gaussian mixture distribution model with three components is then fit to the pixel values to obtain three Gaussian distributions $(\mu_i, \sigma_i^2, i = 1, 2, 3)$. Let μ_1 denote the smallest mean value among all three Gaussian distributions, the primary fibre detection result is given by assigning 0 (black) to the pixels smaller than μ_1 and 1 (white) to those larger than μ_1 in fibre-enhanced image, respectively.

2.2.3 Noise removal and gap bridging

The primary detection result is consisted of two types of objects, qualified fibre and noise. A qualified fibre is defined as an object having either high eccentricity (for elongated fibre) or low solidity (for zigzag fibre). All other objects are regarded as noise. An eccentricity and solidity filtering process is applied to the primary detection result to eliminate noise. All the fibres shorter than the minimum fibre length threshold $L_{\min}$ are removed as well. The fibre length threshold is set to 20 pixels (~2 um based on our image resolution) but it is adjustable to adapt to various experiment conditions and image resolutions. A dilation process is then deployed to fill the gap between the fragments of a long fibre. Two fragments will be connected if the distance between them is smaller than the maximum gap tolerance $G_{\max}$, which is empirically set to 5 pixels but can be tuned if necessary.

As shown in Figure 2E, the output of the gap-bridging process will be the final fibre detection result. The final fibre image is further skeletonized to obtain the fibre skeleton image, where the fibre is reduced to a 1-pixel-wide skeleton while its topological structures are preserved. The fibre skeleton is then overlaid with the enhanced true colour fibre image for subsequent colour encoding process, as shown in Figure 2F.

2.2.4 Colour encoding

The objective of the colour encoding process is to convert fibre pixel intensities to a colour code (CldU, red or IdU, green). Every pixel along the fibre skeleton will be assigned a colour by comparing the mean intensity of itself and its 3-by-3 surrounding pixels between channels I_{RE} and I_{GE}. The intensity values of the skeleton are extracted pixel-by-pixel from one endpoint to another and then form two intensity arrays, Int_R and Int_G (corresponding to I_{RE} and I_{GE}, respectively). A moving average filter is applied to both arrays to generate two smoothed intensity arrays, called Int_{RS} and Int_{GS}. The normalized intensity difference array ΔInt is calculated as $\Delta\mathrm{Int} = 2 \times |\mathrm{Int}_{\mathrm{RS}} - \mathrm{Int}_{\mathrm{GS}}|/(\mathrm{Int}_{\mathrm{RS}} + \mathrm{Int}_{\mathrm{GS}}) \times 100\%$. Inspired by the colour assignment scheme proposed in Ghesquière *et al.* (2019), a similar approach is used here:

1. For one given skeleton (fibre), all the pixels whose ΔInt is larger than 10% are assigned the colour having higher intensity.
2. All the pixels are grouped into separate segments, a segment is defined as contiguous pixels for which have the same colour or no colour has been assigned yet. After this step there will be three types of segments: red segment, green segment and blank segment (no colour has been assigned yet).

   For the remaining blank segments:
3. If the blank segment has two neighbouring segments having the same colour, it will be assigned the same colour as its neighbours.
4. If the blank segment has two neighbouring segments having different colours, it will be split to two halves, and each half is assigned the colour that matched its neighbour.
5. If the blank segment has only one neighbouring segment (i.e. it is located at the end of fibre) then its neighbour's colour will be assigned to it.

Once all the pixels of a skeleton are assigned a certain colour, all segments shorter than 5 pixels are considered as erroneous, and their colour will be inverted so that they will be merged with their neighbouring segments. The actual DNA fibre image and its colour encoding result (pseudo fibre) are shown in the inset of Figure 3A–F. After the colour encoding result is obtained, all fibres will be classified as one of six patterns according to their colours and number of segments.

2.3 DNA Stranding GUI

Imaging DNA fibre is challenging because of the fragility of structures, and image quality can vary greatly from lab to lab. The variation in fibre morphology, fibre branching (crossing) and clustering as well as the potential non-specific staining also raise great challenge for the development of automatic fibre detection and segmentation algorithm. A GUI, which allows users to adjust the automatic fibre segmentation and produce the final analytic results is hence developed. The user guide of the proposed user interface is provided in Supplementary File S1.

Briefly, a batch of images can be loaded from a specified folder, providing an image list from which individual images can be enhanced by a simple click. An additional contrast adjustment tool is provided for display and easier annotation. A toolbar on the top right corner of the image consists of standard zoom and pan tools along with dedicated annotation tools denoted as lock, draw and erase.

Drawing tools allow user to manually annotate the image via a freehand drawing, polygon drawing or enhanced freehand drawing with snapping to the underlying intensity image. A key feature of the manual annotation tool is that it allows drawing of crossing fibres. The lock tool allows the user to access and edit the location of each individual fibre 'waypoint' as well as an object-based eraser mode to erase complete objects for faster editing. The eraser can also be set to an area-based mode (i.e. delete any fibre within the freehand drawn region). The GUI also provides the flexibility of either individual or batch fibre analysis. Therefore, manual, semi-automatic and fully automatic quantification of DNA RFs are implemented in *DNA Stranding* software package.

As described earlier, Figure 3 represents the fibre patterns that are frequently analysed for interpreting DNA replication and repair using the DNA fibre-combing technique at the single-molecule level. Depending on the question to be addressed and the experiments performed, the user decides to interpret the results for each condition based on the fibre patterns detected by the software. Therefore, precise identification of DNA fibre patterns is critical to fasten and accurately obtain the overall understanding on DNA replication and repair using the DNA fibre analyses. Here, with the software, we developed, user is able to automatically detect and measure all common-used fibre patterns, for interpreting the fibre images. An example of the fibre image and the software interpretation of the fibre patterns and their relevance to DNA replication and repair processes are presented in Supplementary File S2, frequently asked questions.

3 Results

3.1 Enhanced semi-automatic user interface of DNA Stranding allows for greater flexibility in the correction of automatic fibre detection results

Variations in sample preparation and imaging, the complexity of fibre morphology and fibre clustering together raise a great challenge to the development of precise fibre detection algorithms. Here, we show two examples of the semi-automatic workflow where the user can correct the incorrect detection results using the annotation tool provided by *DNA Stranding*.

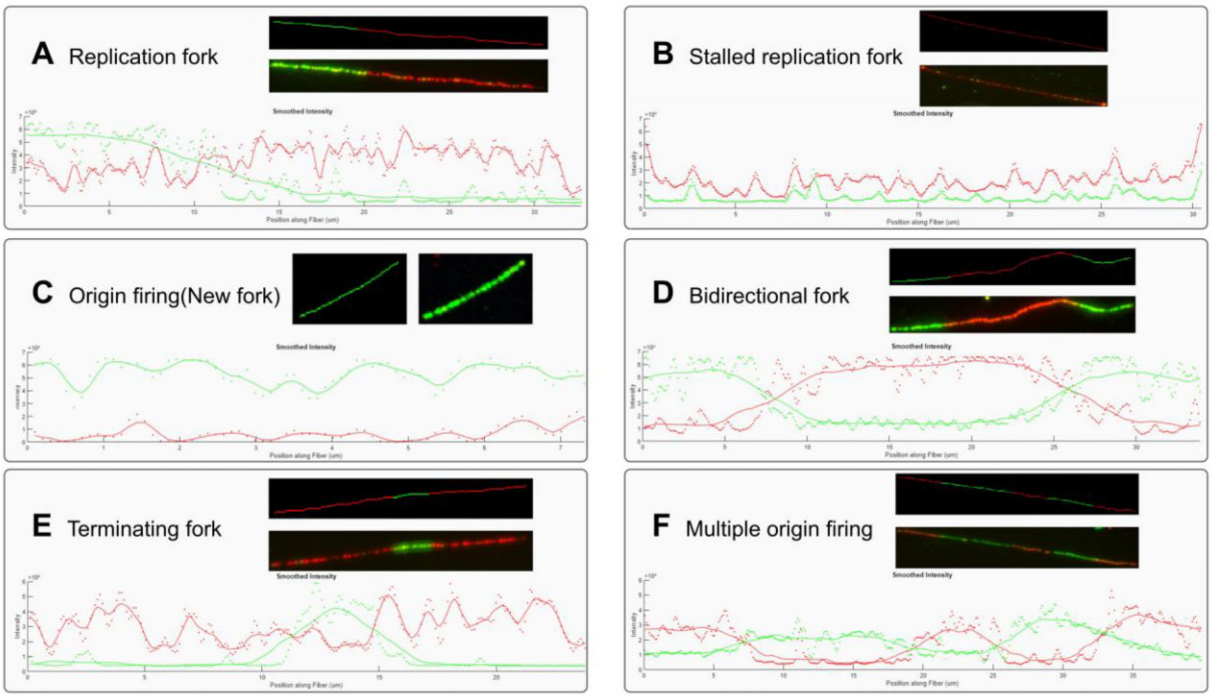

Fig. 3. Intensity profile of DNA fibre. The upper part of each sub-figure shows the actual and pseudo DNA fibre while the lower part shows the original (dotted line) and smoothed (solid line) intensity of both channels of DNA fibre. (**A**) RF. (**B**) Stalled RF. (**C**) Origin firing/New fork. (**D**) Bidirectional fork. (**E**) Terminating fork. (**F**) Multiple origin firing

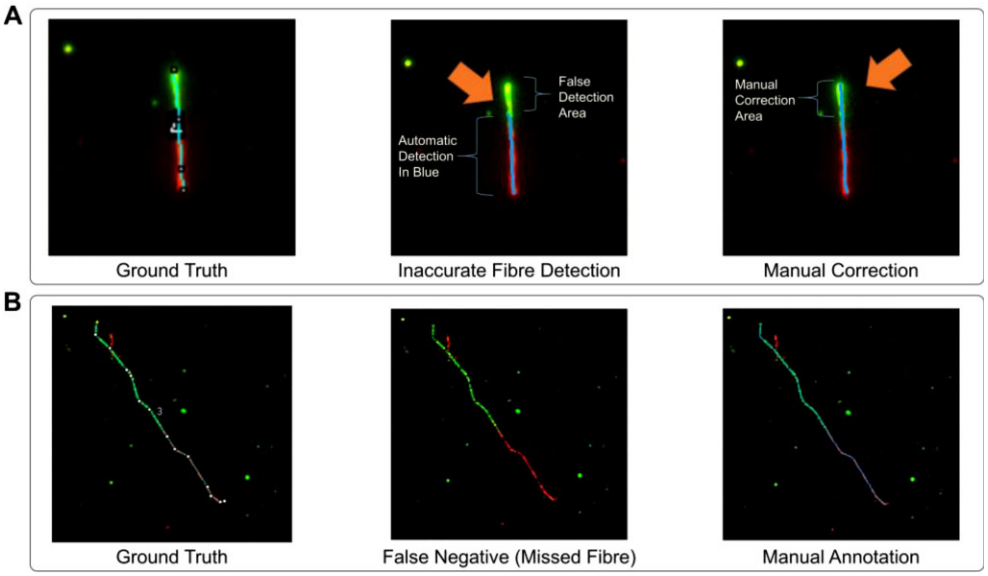

Fig. 4. Two examples showing that how user can modify/correct the automatic fibre detection result using the annotation tools provided by DNA Stranding. (**A**) Left panel: ground truth. Middle panel: automatic detection only identified part of the fibre. Right panel: user can drag the waypoint of fibre to correct the inaccurate result. (**B**) Left panel: ground truth. Middle panel: automatic detection failed to identify this fibre. Right panel: user manually annotates this fibre

The left panel of Figure 4A shows a fibre annotated by user using another image manipulation software (without any involvement of *DNA Stranding*) and it serves as the ground truth. The middle panel of Figure 4A shows the incorrect detection of this fibre given by the algorithm, in which only part of the fibre was identified. The right panel of Figure 4A shows the manually corrected fibre. In this case, the user stretched this fibre by dragging its waypoint, alternatively the user can also choose to delete the incorrect fibre and manually annotate a new fibre. Similarly, the left panel of Figure 4B shows the manually annotated ground truth fibre. The middle panel of Figure 4B shows an example of a missed fibre, i.e. false negative. In the right panel of Figure 4B, the user drew a new fibre using the enhanced freehand draw tool. The enhanced semi-automatic pipeline provided by *DNA Stranding* has given users greater capability and flexibility in the correction of automatic fibre detection results.

3.2 Computer-aided automatic fibre length measurement

We evaluated the accuracy of our fibre length detection using multiple images from different experiment conditions described in Section 2 and compared the manual measurement results (which we use as the ground truth) with the automatic measurements of *DNA Stranding* along with another existing open-source fibre analysis software, FiberQ (Ghesquière *et al.*, 2019). The performance of automatic fibre detection of another open-source software FiberAI (Mohsin *et al.*, 2020) is poor so that we were not able to conduct the comparison between FiberAI and ground truth. As shown in Figure 5A, a good correlation ($R = 0.81$ for *DNA Stranding* and $R = 0.82$ for FiberQ, Pearson correlation coefficient) between the fibre length given by manual measurement and both automatic approaches has been observed. Among all 182 manual-annotated fibres, *DNA Stranding* successfully identified 149 (81.87%) of them, while FiberQ identified 130 (71.43%). For both software, most of the missed fibres are located at very dense regions or at crossings with other fibres. The mean relative error of fibre length measurement of *DNA Stranding* and FiberQ is 13.22% and 16.58%, respectively, as shown in Figure 5B. Our analyses thus revealed a close overall association between both software packages and ground truth.

We have also conducted a replication stress experiment to validate the capability of fibre length measurement as used in a contemporary DNA fibre analysis investigation. LCLs were variably pre-treated with HU and PARPis and then subjected to CldU and IdU labelling for 30 min each. The fibre length and replication speed were measured by *DNA Stranding* for both treatment and control groups. As shown in Figure 5C, consistent with prior reported results, HU (Wilhelm *et al.*, 2014) and PARPi (Maya-Mendoza *et al.*, 2018) significantly altered the replication speed (Control versus. HU, $P < 1e{-}15$, Control versus PARPi, $P < 1e{-}6$, two-sample *t*-test) as indicated by their respective fibre lengths recorded.

3.3 Greater fibre pattern identification capability using DNA Stranding

After colour encoding step, each individual fibre will be categorized as one of six classes as shown in Figure 3. This is an improvement over existing fibre analysis software, FiberQ, which only allows for the identification of two patterns or categories of DNA strands. Fibre pattern recognition is useful for calculating the percentage of origin firing/new fork or stalled RFs, etc. To validate the performance of the fibre pattern recognition process, we performed DNA RF stalling experiments to evaluate the level of stalled RFs when replication stress inducing agent was added. The RFs were stalled transiently after the first labelling (CIdU: red, 30 min) with a high dose of HU or Camptothecin (CPT) for 4 h. After 4 h fresh media with the second label (IdU: green) was added and cells were incubated for another 30 min. The red-only fibres indicate stalled RFs while red-green fibres indicate RF restarts. With this experimental setup we measured the percentage of stalled RFs detected by *DNA Stranding*. As shown in Figure 5D, consistent with the effect of HU and CPT reported previously (Vesela *et al.*, 2017; Yousefi and Rowicka, 2019), there is a statistically significant increase (non-treated (NT) versus CPT, $P < 1e{-}3$; NT versus HU, $P = 0.02$, two-sample *t*-test) in the levels of stalled RFs.

We have also compared the performance between semi-automatic and fully automatic fibre pattern analyses using two groups of images from same experimental setup. The RFs were labelled with CldU and IdU for 30 min each and subjected to automatic or semi-automatic analyses. For the semi-automatic analyses, two experienced users were asked to manually annotate a substantial amount of fibres for two groups of images within the software using the annotation tool and perform automatic fibre pattern recognition. For fully automatic analyses, the images were directly subjected to fibre detection and fibre pattern recognition using the software. Both manually annotated and automatically detected fibres are categorized into one of six patterns and the percentage of each class is shown in Figure 6A (Group 1) and 6B (Group 2). In Group 1, although User 1 annotated almost 20% more fibres than User 2, the proportion of different fibre patterns are similar in their annotation (User 1 versus User 2, $R = 0.99$, Pearson correlation coefficient). A strong correlation between fibre pattern proportion given by semi-automatic and fully automatic analyses has also been observed (Auto. versus User 1, $R = 0.96$; Auto. versus User 2, $R = 0.93$, Pearson correlation coefficient). In Group 2, a very close correlation between two human users as well as automatic analyses and human users with regards to fibre pattern percentage has been

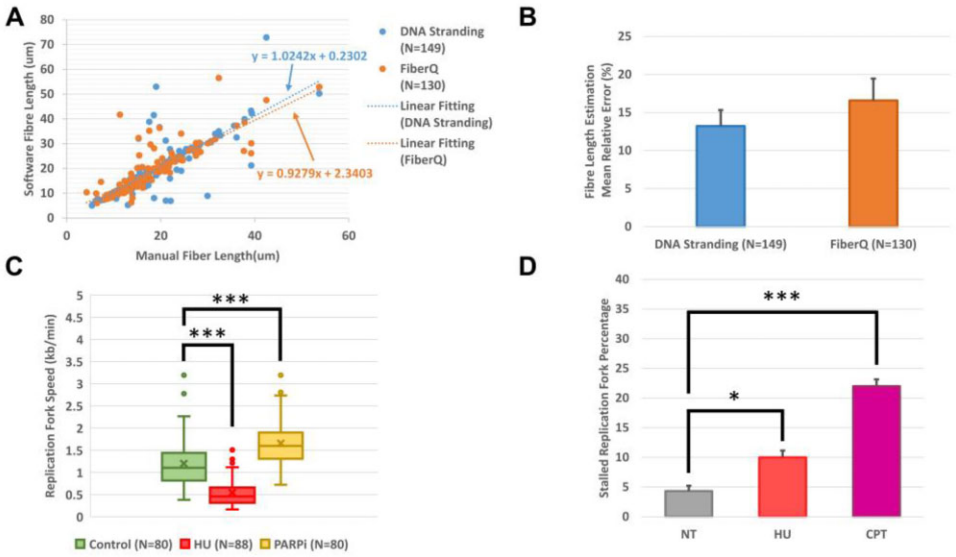

Fig. 5. (**A**) Comparison of fibre length between manual annotation and software estimation, a good correlation between the fibre length given by manual annotation and both software ($R = 0.81$ for DNA Stranding and $R = 0.82$ for FiberQ, Pearson correlation coefficient) has been observed. (**B**) Comparison of mean relative error of fibre length estimation between two software. DNA Stranding, 13.22%; FiberQ, 16.58%. (**C**) HU and PARPi significantly alter RF speed (Control versus HU, $P < 1e{-}15$; Control versus PARPi, $P < 1e{-}6$, two-sample *t*-test). Cells were optionally pre-treated with HU and PARPis, then subjected to CIdU and IdU labelling 30 min each. The RF speed is measured automatically using DNA Stranding. (**D**) Stalled RF percentage aberrantly increases under the replication stress introduced by high dose of HU and CPT (NT versus HU, $P = 0.02$; NT versus CPT, $P < 1e{-}3$, two-sample *t*-test). The RFs were stalled transiently after the first labelling (CIdU, 30 min) with a high dose of HU or CPT for 4 h and then subject to second labelling (IdU) for another 30 min

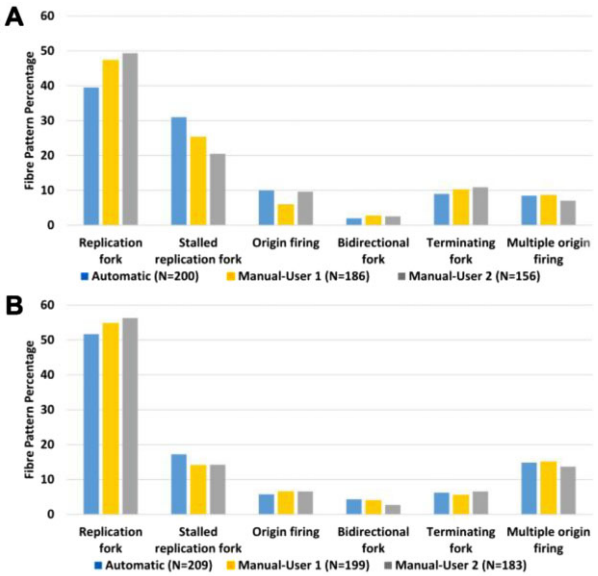

Fig. 6. Percentage of different fibre patterns identified by software and human using two groups of images from same experimental setup. The RFs were labelled with CldU and IdU for 30 min each and subjected to automatic or semi-automatic analyses. (**A**) Group 1, Auto. versus User 1, $R = 0.96$; Auto. versus User 2, $R = 0.93$, User 1 versus User 2, $R = 0.99$, Pearson correlation coefficient. (**B**) Group 2, Auto. versus User 1, $R = 0.99$; Auto. versus User 2, $R = 0.99$, User 1 versus User 2, $R = 0.99$, Pearson correlation coefficient

observed (Auto. versus User 1, $R = 0.99$; Auto. versus User 2, $R = 0.99$; User 1 versus User 2, $R = 0.99$, Pearson correlation coefficient). These results were able to be obtained by the ability of *DNA Stranding* to identify and categorize different fibre patterns.

4 Conclusion and discussion

Fluorescent-based DNA fibre assays are widely used in the studies of DNA replication dynamics. The analyses of DNA fibres are usually conducted manually, which is laborious and time-consuming. The development of an automatic fibre quantification solution is hence needed. Some previous studies had been reported to address the unmet needs, such as CASA (Wang *et al.*, 2011) and FiberQ (Ghesquière *et al.*, 2019). FiberQ is open-sourced while CASA is not. However, the fibre detection process of both FiberQ and CASA are purely automatic, and solely relying on automatic detection is not a perfect solution since inaccurate fibre detections and false negatives are almost inevitable. The open-source software FiberAI (Mohsin *et al.*, 2020) also introduced the concept of semi-automatic pipeline, but their software was not well implemented. FiberAI uses object detection and instance segmentation algorithm to predict the location of fibres. If there are more than one predicted objects in one bounding box (which is the case sometimes), FiberAI does not connect them but adds the length of each object together as 'fibre length'.

We hence introduced an enhanced semi-automatic fibre analysis pipeline that is integrated with our software, named *DNA Stranding*. Users can first apply automatic detection to identify most of the fibres and then utilize the annotation tool provided by the software to rectify any incorrect detections and/or missed fibres.

In our experiment, both FiberQ and *DNA Stranding* have missed a certain portion of ground truth fibres. After thorough examination, we find that almost all missed fibres are located at very dense regions or are crossed with other fibres. We therefore suggest that users pay more attention to fibre dense regions and visually inspect for any potential inaccurate or missed detections or to otherwise select fibre detection results only from sparser regions.

Unlike previous automatic fibre detection solutions, *DNA Stranding* is able to recognize up to six categories of fibres. Aside from normal replication speed and CldU/IdU ratio analyses, *DNA Stranding* also enables users to perform additional various analyses such as assessing the level of stalled forks/new forks, measuring the inter-origin distance, etc. These analyses would otherwise have to be performed manually since existing software do not have the capability to identify novel features in patterns. *DNA Stranding* therefore enables a greater range of research on DNA fibres to be performed with the incorporation of automatic detection to save time.

Overall, we developed an automatic DNA fibre analysis algorithm that can perform a more accurate rapid detection and quantification of DNA replication tracts over previous existing software FiberQ and FiberAI. We incorporated this enhanced automatic detection algorithm into a semi-automatic fibre analysis workflow, which is integrated in our *DNA Stranding* software. Users can utilize the automatic detection to obtain most of the fibres and then take advantage of the annotation tool to perform minor corrections. Compared to conventional manual approach, the semi-automatic analysis results in a significant reduction in processing time while maintaining a very high accuracy of results comparable to human analysis. Our software hence has great potential to be applied in DNA fibre analysis studies.

Funding

This work was supported by National Research Foundation Singapore, Clinician Scientist Award [NMRC/CSA-INV/0017/2017, MOH-000654] and administered by the Singapore Ministry of Health's National Medical Research Council; and the Ministry of Education, Singapore, Academic Research Fund Tier 1 [2019-T1-001-018]; the National Cancer Centre Research Fund Terry Fox Grant [NCCRF-YR2018-NOV-1]; and the Nanyang Technological University Start-Up Grant (to J.N.). This work was jointly supported by BII and IMCB, BMRC, A*STAR research funding, and A*STAR BMRC ATR Grant.

Conflict of Interest: none declared.

Data availability

The data underlying this article will be shared upon reasonable request to the corresponding author.

References

Federici,G. and Soddu,S. (2020) Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J. Exp. Clin. Cancer Res.*, **39**, 1–12.

Frangi,A.F. *et al.* (1998) Multiscale vessel enhancement filtering. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, pp. 130–137.

Frisan,T. *et al.* (2001) Generation of lymphoblastoid cell lines (LCLs). *Methods Mol. Biol.*, **174**, 125–127.

Ghesquière,P. *et al.* (2019) An open-source algorithm for rapid unbiased determination of DNA fiber length. *DNA Repair (Amst.)*, **74**, 26–37.

Green,M.D. *et al.* (2015) Microscopy techniques to examine DNA replication in fission yeast. *Methods Mol. Biol.*, **1300**, 13–41.

Halliwell,J.A. *et al.* (2020) DNA fiber assay for the analysis of DNA replication progression in human pluripotent stem cells. *Curr. Protoc. Stem Cell Biol.*, **54**, e115.

Hill,S.J. *et al.* (2018) Prediction of DNA repair inhibitor response in Short-Term Patient-Derived ovarian cancer organoids. *Cancer Discov.*, **8**, 1404–1421.

Kolinjivadi,A.M. *et al.* (2017) Moonlighting at replication forks – a new life for homologous recombination proteins BRCA 1, BRCA 2 and RAD 51. *FEBS Lett.*, **591**, 1083–1100.

Kolinjivadi,A.M. *et al.* (2020) Emerging functions of Fanconi anemia genes in replication fork protection pathways. *Hum. Mol. Genet.*, **29**, R158–R164.

Lopes,M. (2009) Electron microscopy methods for studying in vivo DNA replication intermediates. *Methods Mol. Biol.*, **521**, 605–631.

Maya-Mendoza,A. *et al.* (2018) High speed of fork progression induces DNA replication stress and genomic instability. *Nature*, **559**, 279–284.

Mohsin,A. *et al.* (2020) FiberAI: a deep learning model for automated analysis of nascent DNA fibers. BioRxiv. https://doi.org/10.1101/2020.11.28.397430.

Nieminuszczy,J. *et al.* (2016) The DNA fibre technique–tracking helicases at work. *Methods*, **108**, 92–98.

Quinet,A. *et al.* (2017) DNA fiber analysis: mind the gap! *Methods Enzymol.*, **591**, 55–82.

Ray Chaudhuri,A. *et al.* (2016) Replication fork stability confers chemoresistance in BRCA-deficient cells. *Nature*, **535**, 382–387.

Santos,E.M. *et al.* (2021) Design of large stokes shift fluorescent proteins based on excited state proton transfer of an engineered photobase. *J. Am. Chem. Soc.*, **143**, 15091–15102.

Stanojcic,S. *et al.* (2016) Single-molecule analysis of DNA replication reveals novel features in the divergent eukaryotes Leishmania and *Trypanosoma brucei* versus mammalian cells. *Sci. Rep.*, **6**, 23142.

Vesela,E. *et al.* (2017) Common chemical inductors of replication stress: focus on cell-based studies. *Biomolecules*, **7**, 19.

Wang,Y. *et al.* (2011) Automated DNA fiber tracking and measurement. In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, Chicago, IL, pp. 1349–1352.

Wilhelm,T. *et al.* (2014) Spontaneous slow replication fork progression elicits mitosis alterations in homologous recombination-deficient mammalian cells. *Proc. Natl. Acad. Sci. USA*, **111**, 763–768.

Yousefi,R. and Rowicka,M. (2019) Stochasticity of replication forks' speeds plays a key role in the dynamics of DNA replication. *PLoS Comput. Biol.*, **15**, e1007519.

Zeman,M.K. and Cimprich,K.A. (2014) Causes and consequences of replication stress. *Nat. Cell Biol.*, **16**, 2–9.

Zereg,E. *et al.* (2020) Single-fiber studies for assigning pathogenicity of eight mitochondrial DNA variants associated with mitochondrial diseases. *Hum. Mutat.*, **41**, 1394–1406.