FORK-seq: Single-Molecule Profiling of DNA Replication

Magali Hennion, Bertrand Theulot, Jean-Michel Arbona, Benjamin Audit, Olivier Hyrien

FORK-seq: Single molecule profiling of DNA replication

Magali Hennion, Bertrand Theulot, Jean-Michel Arbona, Benjamin Audit, Olivier Hyrien

Affiliations

Magali Hennion

Epigenetics and Cell Fate Centre, UMR7216 CNRS, University of Paris, Paris, 75013, France, e-mail: magali.hennion@u-paris.fr

Bertrand Theulot

Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, Paris, 75005, France, e-mail: theulot@bio.ens.psl.eu

Jean-Michel Arbona

Laboratory of Biology and Modelling of the Cell, Université de Lyon, ENS de Lyon, Université Claude Bernard, CNRS UMR 5239, INSERM U1210, 46 Allée d'Italie, Site Jacques Monod, Lyon, 69007, France, e-mail: jeanmichel.arbona@ens-lyon.fr

Benjamin Audit

Université de Lyon, ENS de Lyon, Université Claude Bernard Lyon 1, CNRS, Laboratoire de Physique, F-69342 Lyon, France, e-mail: benjamin.audit@ens-lyon.fr

Olivier Hyrien

Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, Paris, 75005, France, e-mail: hyrien@bio.ens.psl.eu

Corresponding author : hyrien@bio.ens.psl.eu

Running head

FORK-seq: Single molecule profiling of DNA replication

Abstract

Most genome replication mapping methods profile cell populations, masking cell-to-cell heterogeneity. Here, we describe FORK-seq, a nanopore sequencing method to map replication of single DNA molecules at 200 nucleotide resolution using a nanopore current interpretation tool allowing the quantification of BrdU incorporation. Along pulse-chased replication intermediates from *Saccharomyces cerevisiae*, we can orient replication tracks and reproduce population-based replication directionality profiles. Additionally, we can map individual initiation and termination events. Thus, FORK-seq reveals the full extent of cell-to-cell heterogeneity in DNA replication.

Key Words

DNA replication, whole-genome, single-molecule analysis, nanopore sequencing, convolutional neural network, replication fork direction, replication origins, termination sites

1. Introduction

Eukaryotes replicate DNA by stochastically activating numerous replication origins ***(1)***. Cell-population methods provide only an average picture where inter-cellular variability and rare events are masked. Solving the nature of origins and analyzing fork progression thus requires high-throughput analyses at the single molecule level. Current single molecule methods such as DNA combing can reveal fork progression and spacing by monitoring non-standard nucleotide incorporation with antibodies ***(2)***. However, they provide no sequence information unless they are combined with DNA probes ***(3)***. This is laborious, low resolution and low throughput (typically 6 months of work to study one locus of 1 Mb). Hence, high-throughput single molecule genomic profiling of DNA replication is transformative to replication studies ***(4;5)***.

Here, we describe FORK-seq, one such high-throughput genomic protocol to characterize DNA replication on single DNA molecules. As in many previous replication studies, the progression of replication forks is revealed by monitoring the incorporation of BrdU, a thymidine analog. Originally, we use the nanopore sequencing technology from Oxford Nanopore Technologies (ONT) to obtain the position of each analyzed DNA molecule along the reference genome as well as its BrdU incorporation profile. The former relies on sequence alignment thanks to the robustness of ONT basecalling tools with respect to thymidine substitution by BrdU, while the latter can be achieve using one of two new nanopore current processing pipelines we integrated in the RepNano software suite. Both pipelines use the series of current shifts aligned to the DNA sequence to either (i) assess which thymidine sites are BrdU substituted based on tabulated current shifts matrices (TM) or (ii) directly estimate the proportion of substituted sites over 96 bp windows using a trained neural network (CNN) ***(4)***.

FORK-seq was successfully applied to a *Saccharomyces cerevisiae* strain whose growth fully depends on the external supply of thymidine or BrdU. The protocol consists in (i) growing yeasts exposing them to a short (contributing to the high resolution of the protocol) BrdU pulse followed by a thymidine chase, (ii) extracting

genomic DNA and preparing sequencing libraries, (iii) sequencing using the MinION from ONT and basecalling the reads, and (iv) aligning the reads to the reference genome and quantifying their BrdU incorporation profiles. BrdU incorporation profiles present localized asymmetric motifs where a sharp transition from low to high values is followed by a shallower decrease as the signature of replication forks progression at the time of the BrdU pulse and during the chase, respectively (*see* **Figure 1**).

Dedicated post-processing tools from the RepNano software suite can then be used to automatically determine replication fork locations and orientations. When two diverging (resp. converging) forks are detected on the same nanopore read, the location of an initiation (resp. termination) event can also be deduced ***(4)***. All the steps from yeast culture to final lists of forks as well as initiation and termination events on their genomic location are described in this protocol that can be fully completed within one week.

Thus, we reported over 1.4 million nanopore reads the identification of 58 651 BrdU content motifs associated to oriented-forks and the location of 4 964 and 4 485 individual initiation and termination events, respectively ***(4)*** (*see* **Figure 1**). Bioinformatic analyses showed that replication directionality profile derived from single forks is fully compatible with population-based profiles, and that most events cluster at known origins and fork merging zones, but that 9% and 18% of initiation and termination events, respectively, occur at previously unreported locations. This illustrates that FORK-seq protocol can quantify DNA replication program stochasticity over a complete genome.

[Figure 1 near here]

2. Materials

2.1 Yeast strain

MCM869 genotype is *MATa ade2-1 trp1-1 can1-100 leu2-3 his3-11,15 ura3-1::URA3-GPD-TK7x aur1::AUR1-C-ADH-hENT1 bar1::LEU2 cdc21::kanMX* ***(6)*** (*see* **Note 1**).

2.2 Yeast culture and labelling

1. Liquid YPD medium: Suspend 100 g of YPD powder in 1.8 L of distilled water (*see* **Note 2**). Make up to 2 L with water. Fill five 500 mL-bottles with 400 mL of medium. Autoclave for 15 minutes at 110°C. Store at RT protected from light.

2. 100 mM thymidine solution: Dissolve 1 g of thymidine (MW 242.23 $g.mol^{-1}$) into 41.3 mL of ultrapure water. Filter-sterilise with 0.22 µm filter. Store in 1 mL aliquots at -20°C.

3. 100 mM BrdU solution: Dissolve 1 g of BrdU (MW 307.1 $g.mol^{-1}$) into 32.6 mL of DMSO. Store in 1 mL aliquots at -20°C.

4. YPD thymidine plates: Suspend 67 g of YPD agar in 1 L of distilled water. Heat to boiling while stirring to dissolve all ingredients. Autoclave for 15 minutes at 110°C. Let the mixture cool down to ~55°C. Add 1 mL of 100 mM thymidine solution. Mix well and pour into sterile, labeled petri dishes. Air bubbles should be removed by slow orbital sloshing.

5. Thymidine medium: The day of the experiment, add 1/1000 (v/v) of 100 mM thymidine solution to YPD medium. Prewarm to 30°C before adding it to the cells.

6. Autoclaved glass Erlenmeyer flasks (50 and 500 mL).

7. Shaking incubator at 30°C.

8. Laminar flow hood.

2.3 DNA extraction

1. Nuclease free Milli-Q water
2. 10 % Sodium dodecylsulfate (SDS)
3. TE buffer : 10 mM Tris-HCl pH 8.0, 1 mM EDTA, filtered
4. Isopropanol
5. 70% Ethanol
6. 40 mg/mL RNase A solution
7. 5 M potassium acetate (KAc), pH 7.5
8. Y1 buffer : 1 M sorbitol, 100 mM EDTA pH 8.0, 14 mM β-mercaptoethanol. For a 1-liter solution, dissolve 182.2 g sorbitol in 600 mL distilled water. Add 200 mL of 0.5 M EDTA. Under a chemical hood, add 1 mL of β-mercaptoethanol (14.3 M). Adjust the volume to 1 L with distilled water. Store at 4°C.
9. 1 U/µL Zymolyase. Prepare a master solution for all your samples just before use by diluting in Y1 buffer to 1 U/µL. Using Zymolyase 20T (20 U/mg), this corresponds to a 50 mg/mL solution.
10. Lysis buffer: 114 mM Tris-HCl pH 8.0, 115 mM EDTA, 571 mM NaCl, 1.14% PVP40. Prepare a fresh master solution for all your samples. Per sample, mix 500 µL of 1 M Tris-HCl pH 8.0, 500 µL of 0.5 M EDTA and 500 µL of 5M NaCl. Add 50 mg of PVP 40, and complete with 2.875 mL of water. Warm up 30 minutes at 65°C and filter sterilize with 0.45 µm filter.
11. Qubit fluorometer
12. Qubit dsDNA BR Assay Kit
13. Pulsed-field gel electrophoresis system or Agilent TapeStation
14. 50°C water bath

15. NanoDrop (Thermofisher)

2.4 Sequencing libraries

1. Ligation Sequencing Kit (ONT)
2. NEBNext FFPE Repair Mix
3. NEBNext Ultra II End repair/dA-tailing Module
4. If multiplexing, Native Barcoding Expansion 1-12 (ONT) or/and 13-24 (ONT)
5. If multiplexing, NEB Blunt/TA Ligase Master Mix
6. NEBNext Quick Ligation Module
7. Agencourt AMPure XP beads
8. Tris buffer: 10 mM Tris-HCl pH 8.0, filtered
9. Freshly prepared 70% ethanol in Tris buffer
10. 1.5 mL Eppendorf DNA LoBind tubes
11. 0.2 mL PCR tubes
12. Magnetic rack
13. Benchtop centrifuge
14. Thermal cycler or thermoblock
15. 37°C water bath
16. Hula mixer

2.5 Sequencing

1. R9 flow cell (ONT)
2. Flow Cell Priming Kit (ONT)

3. MinION Mk1B or MinION Mk1C or GridION (ONT) with adapted computational resources

3. Methods

All yeast manipulations are done under sterile conditions.

3.1 Yeast culture

1. Streak frozen MCM869 cells with a sterile tip onto a YPD thymidine plate. Let grow in the dark at 30°C for 48h to 72h until you see small colonies. The plate can be stored at 4°C for a maximum of two weeks.

2. The day before the labelling, pick one colony from the plate using a sterile tip and transfer it into 10 mL thymidine medium.

3. Measure OD_{600} and dilute the cells into 100 mL fresh thymidine medium adjusting the volume of cells so that the OD_{600} would reach 1 the next morning (beginning of the exponential phase, *see* **Note 3**). Grow cells in a 500 mL Erlenmeyer flask at 30°C in the dark under constant agitation (typically 200 rpm).

3.2 Replication labelling

1. The day of the experiment, check that OD_{600} is around 1. Dilute the cells to 0.1 OD_{600} in 100 mL of fresh thymidine medium. Let grow the cells for 8h (OD_{600} ~0.8).

2. Transfer the cells to two 50 mL conical tubes and centrifuge in a swinging bucket at 30°C for 2 min at 3000 g. Remove carefully the supernatant by pouring or by pipetting, taking care of not removing part of the pellet.

3. Wash with 100 mL of YPD medium (without thymidine), repeating the centrifugation and the supernatant removal.

4. Resuspend into 100 mL YPD medium (without thymidine) and put the cells back into the incubator for 30 min at 30°C (*see* **Note 4**).

5. Add 100 µL of 100 mM BrdU solution with constant shaking, processing quickly to avoid medium cooling. After 4 min at 30°C (*see* **Note 5**), add 1 mL of 100 mM thymidine and incubate for 45 min at 30°C.

6. Centrifuge the cells for 5 min at 5000 g (room temperature). Remove the supernatant, wash the pellet with water and put the tubes with cell pellets at -20°C. They can be stored for several months.

3.3 DNA extraction

This protocol is derived from ***(7)*** with minor changes. It is optimized for a yeast pellet corresponding to 100 mL of OD_{600} ~1 suspension, for an expected yield of 10-20 µg of DNA. In the original paper ***(4)***, we used a different protocol described in **Note 6**.

3.3.1 Cell wall lysis

1. Resuspend the pellet in 4 mL of Y1 buffer, and add 250 µL of zymolyase solution in a 50 mL tube. Make sure the pellet is well resuspended.

2. Incubate 30 minutes at 37°C with gentle shaking (300 rpm). Centrifuge for 3 minutes at 3000 g. Discard the supernatant.

3.3.2 Spheroplasts lysis and protein precipitation

1. Resuspend the pellet in 3.5 mL of Lysis buffer using a wide bore tip. It is crucial that spheroplasts are well resuspended for proper lysis.

2. Add 500 µL of SDS 10% and 15 µL of RNAse A solution (*see* **Note 7**). Flip gently 10 times to homogenize. Incubate 30 minutes at 50°C with moderate shaking (300 rpm).

3. Put the lysate tube on ice. Add 10 mL of TE buffer and 5 mL of 5M KAc, pH 7.5. Invert with precaution 20 times. A white precipitate should appear at this stage. Incubate on ice for at least 5 minutes.

4. Centrifuge for 10 min at 5000 g (4°C). Transfer the supernatant to a new 50 mL tube. If some white precipitate remains, repeat the previous centrifugation.

3.3.3 DNA precipitation

1. Add 1 volume of isopropanol to the supernatant. Invert gently 30 times so that the solution is well homogenized. Bad homogenization will result in low yield. It is possible to see some DNA filaments at this stage.

2. Centrifuge for 20 minutes at 9,500 g at 4°C. Discard the supernatant.

3. Wash the DNA pellet with 5 mL of 70% ethanol. Centrifuge 5 minutes at 1000 g. Remove as much ethanol as possible. Repeat the previous centrifugation to remove completely the ethanol. Do not allow the pellet to dry as it will make it harder to resuspend.

3.3.4 Elution, quantification and quality control

1. Add 100 µL of TE. Incubate the pellet for 30 minutes at 50°C in a water bath to resuspend it.

2. Transfer to a 1.5 mL tube using a wide bore tip.

3. Quantify the amount of DNA with Qubit fluorometer (dsDNA BR Assay Kit, *see* **Note 8**). The final concentration should not be lower than 40 ng/µL to prepare sequencing libraries.

4. Measure A260/280 and A260/230 purity ratios using NanoDrop. Both ratios should be higher than 1.8 (*see* **Note 9**).

5. The size of the DNA is determined using pulsed-field gel electrophoresis or automatized electrophoresis system such as Agilent TapeStation (*see* **Note 10**). The majority of the DNA should be larger than 50-100 kb (*see* **Note 11**).

3.4 Nanopore sequencing libraries

This part follows ONT protocols with minor changes in incubation times, amount of beads and of DNA used at the different steps. Please refer to official documentation at https://nanoporetech.com/resource-centre/protocols for extra details and tips.

3.4.1 DNA repair and purification

1. Dilute 2 µg of DNA (*see* **Note 12**) with Tris buffer for a final volume of 48 µL in a PCR tube.

2. Add 3.5 µL of NEBNext FFPE Repair Buffer and 3.5 µL of NEBNext Ultra II End-prep Buffer. Complete with 3 µL of Ultra II End-prep enzyme mix and 2 µL of NEBNext FFPE DNA Repair enzyme mix. Flick gently the tube and incubate 20 minutes at 20°C followed by 20 minutes at 65°C using a thermal cycler (*see* **Note 13**).

3. Transfer the mix into a 1.5 mL DNA LoBind tube. Add 0.5 volume (30 µL, *see* **Note 14**) of AMPure XP beads previously intensively resuspended by vortexing. Mix by gently flicking the tube.

4. Incubate on a Hula mixer for 10 minutes at room temperature. Spin down the sample using a mini benchtop centrifuge (full speed for 3 seconds). Place the tube on a magnetic rack and wait until the eluate is clear.

5. Discard the supernatant and wash the beads twice with 300 µL of 70% ethanol. Spin down and put back on the magnet to remove the ethanol. Repeat this step to remove all residual ethanol. Never let the beads' pellet dry as it can perturb the elution.

6. Remove the tube from the magnetic rack and add 25 µL of Tris buffer. Incubate 10 minutes in a 37°C water bath. Put the tube back on the magnet and wait until the eluate is perfectly clear.

7. Transfer all the eluate in a new 1.5 mL DNA LoBind tube.

8. Quantify by Qubit the amount of DNA recovered using 1 µL of the eluate. The recovery aim is around 1.3 µg of DNA.

3.4.2 Native barcode ligation - facultative

If you want to sequence several samples in one flow cell you have to add a specific barcode to each sample. If not, skip this step and proceed directly with adapter ligation.

1. In a new 1.5 mL DNA LoBind tube, dilute 750 ng of previous eluate in Tris buffer to a final volume of 22.5 µL. Add 2.5 µl of the selected Native Barcode and 25 µL of Blunt/TA Ligase Master Mix. Mix by gently flicking the tube and incubate for 1 hour at room temperature.

2. Add 0.5 volume (25 µL) of AMPure XP beads previously intensively resuspended by vortexing. Mix by gently flicking the tube. Incubate on a Hula mixer for 10 minutes at room temperature.

3. Spin down the sample with a benchtop centrifuge. Place on a magnet and wait until eluate is clear. Discard the supernatant. Wash the beads twice with 300 µL of 70% ethanol. Spin down and put back on the magnet to remove the ethanol. Repeat this step to remove all residual ethanol.

4. Remove the tube from the magnetic rack. Add 25 µL of Tris buffer. Incubate 10 minutes at 37°C in a water bath. Put the tube back on the magnet and wait until the eluate is perfectly clear. Transfer the eluate into a new 1.5 mL DNA LoBind tube.

5. Quantify by Qubit the amount of DNA recovered using 1 µL of the eluate. The recovery aim is 500 ng of DNA.

6. In a new DNA LoBind tube, pool the samples you want to sequence in the same flowcell so that you end up with 800 ng of DNA in total. To do so, adjust the volumes to put the same amount of DNA for each sample (or adjust accordingly if you want more reads from specific samples).

3.4.3 Adapter ligation

This step is identical whether you have multiplexed samples or not, except that the Adapter Mix is different. Use Adapter Mix I if you process a single sample without barcode and Adaptater Mix II for multiplexed samples.

1. Dilute 800 ng of previously prepared sample into 65 µL of Tris buffer. Add 5 µL of Adaptater Mix (I or II, see above), 20 µL of NEBNext Quick Ligation Reaction Buffer and 10 µL of Quick T4 DNA Ligase. Mix by gently flicking the tube. Incubate at least 1h at room temperature. We did notice significant increase in read length with longer incubation time.

2. Add 0.5 volume (50 µL) of AMPure XP beads previously intensively resuspended by vortexing. Mix by gently flicking the tube. Incubate on a Hula mixer for 10 minutes at room temperature. Spin down the sample with a benchtop centrifuge. Place on a magnet and wait until eluate is clear. Discard the supernatant.

3. Wash the beads twice with 250 µL Long Fragment Buffer (*see* **Note 15**). Spin down and put back on the magnet to remove the buffer. Repeat this step to remove all residual buffer.

4. Remove the tube from the magnetic rack and add 15 µL of Elution Buffer. Incubate 10 minutes at 37°C in a water bath. Put the tube back on the magnet and wait until the eluate is perfectly clear. Transfer the eluate in a new 1.5 mL DNA LoBind tube.

5. Quantify by Qubit the amount of DNA recovered using 1 µL of the eluate. The recovery aim is 400 ng of DNA. Libraries can be stored at 4°C for several days.

3.4.4 Priming the flow cell and loading the library

1. Equilibrate the MinION R9.4.1 flow cell at room temperature.

2. Open MinKNOW, insert the flow cell into the MinION device and perform the flow cell check. To do so, choose the flow cell type from the selector box. Then

mark the flow cell as "Selected". Click "Check flow cells" at the bottom of the screen and select "FLO-MIN106". Click "Start test". Be sure that the number of pores detected is above the warranty before proceeding any further.

3. Prepare the priming mix by adding 30 µL of mixed Flush Tether into one tube of Flush Buffer.

4. Open the priming port and draw back a small volume to remove any bubble (a few µL, *see* **Note 16**). Load 800 µL of the priming mix through the priming port and wait for 5 minutes. In the meantime, prepare the library.

5. In a new 1.5 mL DNA LoBind tube, mix 12 µL of DNA (ideally 200 ng, *see* **Note 17**) with 37.5 µL of Sequencing Buffer and 25.5 µL of Loading Beads previously intensively resuspended by vortexing for a final volume of 75 µL.

6. Open the SpotON port, and add 200 µL of the priming mix into the priming port (not into the SpotON port).

7. Using a wide bore tip, resuspend the prepared library and inject it into the SpotON port by droplets. Close the SpotON port and the priming port and proceed to data acquisition (*see* **Note 18**).

3.5 Nanopore sequencing and base calling

This part is highly dependent on the material available. Using the MinION Mk1C, the base calling and the demultiplexing can be done directly on the device during and after the run (*see* **Note 19**). Using the MinION Mk1B, it is also possible to base call during the run if the computer has enough computational power (please refer to ONT recommendations). If not, the raw FAST5 files can be processed after the run using Guppy (*see* **Note 20**). The analysis is quite resource-consuming and the use of HPC cluster or adapted workstation (*see* **Note 21**) is necessary.

Moreover, one MinION run will typically generate 100 to 300 Go of raw data. Therefore, appropriate storage capacity must be available.

3.5.1 Concomitant acquisition and base calling

1. For kit selection, select "SQK-LSL109" and if samples are multiplexed, select "EXP-NBD104" or "EXP-NBD114". Configure acquisition keeping default parameters.

2. For the base calling, choose "High-accuracy basecalling", and enable "barcoding" if you have multiplexed samples. This will split both FAST5 and FASTQ files according to their barcode. In "output", keep both FAST5 and FASTQ in the output directory. Keep the filtering value of the quality score (qscore) at 7 (*see* **Note 22**).

3. Start the sequencing and let it run for 72h.

During the run, MinKNOW displays the activity of the pores and plots a histogram of read lengths. N50 should typically be in the 30-60 kb range. If multiplexing, two barplots showing the number of bases and of reads attributed to each barcode are displayed in real time. FAST5 and FASTQ files are generated. In addition, a `sequencing_summary.txt` file containing useful details about every single read is made (in the FASTQ folder). Using the Mk1C, the basecalling will go on for several days after the end of the run.

3.5.2 Acquisition with no base calling

If your computer has limited resources do not forget to disable base calling as it could impact the proper acquisition of the raw signal. Once the sequencing is finish you will base call the data with Guppy.

1. Configure acquisition keeping default parameters. For kit selection, select "SQK-LSL109" and if samples are multiplexed, select "EXP-NBD104" or "EXP-NBD114".

2. Disable base calling.

3. Start the sequencing and let it run for 72h.

During the run, MinKNOW displays the activity of the pores and plots a histogram of read lengths. N50 should typically be in the 30-60 kb range.

3.5.3 Base calling with Guppy independently of data acquisition

1. Depending on your computational resources, download and install the CPU or GPU version of Guppy (full instructions at https://community.nanoporetech.com/downloads, *see* **Note 23**.)

2. Proceed to base calling with the command line below (if no multiplexing). `GUPPYDIR` refers to Guppy install directory and `WORKDIR` to the project working directory.

Base calling with Guppy, no multiplexing

```
$ GUPPYDIR/bin/guppy_basecaller -i WORKDIR/RawData/ -s
WORKDIR/Fastq/ -q 4000  \
--flowcell FLO-MIN106 --kit SQK-LSK109 --cpu_threads_per_caller 8
\
 --num_callers 1
```

With

- `-i`: location of the raw FAST5 files; all FAST5 files found in this directory will be processed.
- `-s`: location of the output FASTQ files
- `--cpu_threads_per_caller`: number of threads per caller
- `--num_caller`: number of callers

The last two parameters have to be adapted to your computational resource. Multiple FASTQ files as well as a `sequencing_summary.txt` file will be generated in the chosen directory.

3. If samples are multiplexed, use the same command line simply adding the `barcode_kits` option.

Base calling with Guppy, multiplexing

```
$ GUPPYDIR/bin/guppy_basecaller -i WORKDIR/RawData/ -s
WORKDIR/Fastq/   \
--flowcell FLO-MIN106 --kit SQK-LSK109 -q 4000 --barcode_kits EXP-
NBD104   \
--cpu_threads_per_caller 8 --num_callers 1
```

The FASTQ files corresponding to each barcode will be saved in the directory `Fastq/barcodeX/`, where `X` is the barcode number.

3.5.4 Demultiplexing of the FAST5 files

If samples are multiplexed you then have to split the FAST5 files according to the samples. This is performed using the tool ONT FAST5 API (*see* **Note 24**).

1. Make a subset list from the sequencing summary (generated during base calling) with the reads containing a specific barcode. This have to be done for each barcode (replacing `barecodeX` in the commands).

   Demultiplexing 1/2

   ```
   $ cd WORKDIR
   $ head -n 1 Fastq/sequencing_summary.txt > barcodeX_summary.txt
   $ grep barcodeX Fastq/sequencing_summary.txt >> barcodeX_summary.txt
   ```

2. Using this list, you can now split your FAST5 files using the following command (for each barcode). `APIDIR` refers to the installation directory of ONT FAST5 API.

   Demultiplexing 2/2

   ```
   $ APIDIR/bin/fast5_subset -i WORKDIR/RawData -s
   WORKDIR/RawData/barcodeX \
   -l WORKDIR/barcodeX_summary.txt --batch_size 4000
   ```

3.5.5 Concatenate FASTQ files

Basecalling during or after the run generates multiple FASTQ files for each sample. To allow subsequent RepNano analysis, it is necessary to generate one FASTQ per sample.

1. Without multiplexing: In the following command, `Sample.fastq is the name you choose to give to the FASTQ file containing all the reads.`

Concatenate FASTQ files, no multiplexing

```
$ cat WORKDIR/Fastq/*.fastq > WORKDIR/Fastq/Sample.fastq
$ [rm WORKDIR/Fastq/*fastq]
```

This last `rm` command removes the small FASTQ files to save disk space.

2. With multiplexing: `Sample_barcodeX.fastq` is the name you choose to give to the FASTQ file containing all the reads coming from the sample barcoded with barcodeX.

Concatenate FASTQ files, if multiplexing

```
$ cat WORKDIR/Fastq/barcodeX/*.fastq >
WORKDIR/Fastq/Sample_barcodeX.fastq
$ [rm WORKDIR/Fastq/barcodeX/*fastq]
```

This has to be done for each barcode (replacing `barecodeX` and `Sample` in the command).

Nanopore reads are now base called and ready for downstream analysis.

3.6 Data analysis with RepNano

Please refer to RepNano repository for manual and updates (https://github.com/organic-chemistry/RepNano).

3.6.1 Installation

The installation requires Git (*see* **Note 25**). We recommend using a dedicated Conda environment to assure all software dependencies are met prior to running the analysis (*see* **Note 26**). In particular, the preprocessing step uses Tombo (ONT, *see* **Note 27**) package.

Installation of the computational environment and of Tombo

```
$ conda create --name RepNanoEnv --override-channels -c bioconda -c
defaults \
 python=3.6 keras pandas numba tqdm joblib ont-tombo matplotlib
```

All further steps assume the `RepNanoEnv` conda environment is activated using:

```
$ conda activate RepNanoEnv
```

RepNano installation `SOFTDIR` refers to the directory where you wish to install RepNano.

```
$ cd SOFTDIR
$ conda activate RepNanoEnv
$ git clone https://github.com/organic-chemistry/RepNano.git
$ cd RepNano
$ python setup.py develop
```

These operations result in a working RepNano installation in directory `SOFTDIR/RepNano`, which will be referred to as `REPNANODIR`.

Installation of *simplification* python module

```
$ [conda activate RepNanoEnv]
$ pip install simplification
```

3.6.2 BrdU detection

The pipeline consists in the alignment of the basecalled Oxford Nanopore reads on a reference genome followed by prediction of BrdU content by the neural network

(CNN) and the transition matrix (TM) approaches ***(4)***. Sequencing and basecalling results in FAST5 and FASTQ files (*see* **Section 3.5**). For each sample, you have:

- several FAST5 files (raw current values) containing 4000 reads each (*see* **Note 28**),
- one FASTQ file with all the sequences.

A test dataset is available at https://www.opendata.bio.ens.psl.eu/FORK-seq. The analysis procedure is described using this dataset as an example. In order to use it you have to download *TestDataSet.tgz* (1,1G) and decompress it. `TESTDIR` refers to the directory where the procedure will be run.

Download test dataset

```
$ cd TESTDIR
$ wget https://www.opendata.bio.ens.psl.eu/FORK-seq/TestDataSet.tgz
$ tar -zxvf TestDataSet.tgz
```

1. **Genome indexing [Facultative]**

   The preprocessing starts with the alignment of FASTQ sequences on a reference genome using Minimap2 ***(8)***. In the `--ref` option, you can either put the FASTA file of your genome or the corresponding Minimap2 index (recommended as it is faster). To generate Minimap2 index you have to run the following command (*see* **Note 29**).

   Minimap2 index

   ```
   $ [conda activate RepNanoEnv]
   $ cd TESTDIR
   $ python REPNANODIR/src/repnano/data/index.py TestDataSet/chr4.fa
   TestDataSet/chr4.mmi
   ```

This step has to be done only once.

2. **Preprocessing**

   The preprocessing includes the mapping on the reference genome and a realignment of the current values to fit the reference sequence (Tombo resquiggle function). Every FAST5 is processed separately (parallelizable). This will create a `preprocessed_file.fast5` file that will contain the current values aligned on the reference genome (*see* **Note 30**).

Preprocessing

```
$ [conda activate RepNanoEnv]
$ [cd TESTDIR]
$ python REPNANODIR/src/repnano/data/preprocess.py \
--hdf5 TestDataSet/FORKSeq_combined.fast5 --fastq
TestDataSet/FORKSeq_combined.fastq \
--ref TestDataSet/chr4.mmi --output_name
TestDataSet/preprocessed_file.fast5 --njobs 6
```

With

- `--hdf5`: input FAST5 file
- `--fastq`: input FASTQ file
- `--ref`: reference genome (.fa or .mmi)
- `--output_name`: name of the output FAST5 file (*see* **Note 31**)
- `--njobs`: number of jobs run in parallel; to be adapted to the hardware

The script outputs a description of the reads that failed to be processed. It usually corresponds to 10 to 25 % of the reads. The errors are explained below.

- `seq_not_found`: The read ID from the FAST5 file is not found in the FASTQ file (*see* **Note 32**)

- `Alignment not produced`: The read couldn't be mapped on the reference.
- `Read event to sequence alignment extends beyond bandwidth`: The read could be mapped but the alignment of the current values on the reference sequence failed.
- `Found non canonical bases (['N'])`: The read mapped on a region of the reference that contains non canonical bases (*see* **Note 33**).

3. **BrdU detection**

   RepNano BrdU detection is then called on the `preprocessed_file.fast5` file.

BrdU detection

```
$ [conda activate RepNanoEnv]
$ [cd TESTDIR]
$ python REPNANODIR/src/repnano/models/predict_simple.py \
TestDataSet/preprocessed_file.fast5 --output
TestDataSet/BrdU_calls/output_file.fa \
--overlap 10 --bigf
```

With

- `--output`: output file name
- `--overlap`: [INT] RepNano BrdU detection with the trained neural network model calculates BrdU ratio on 96 bp windows. This parameter defines the number of overlapping windows used to calculate the final average ratio. This will be done by shifting the signal times with a step.

- `--bigf`: Before 2019, the output of ONT basecallers was different and one FAST5 was generated for each read. Remove this option if you work with this old format.

RepNano prediction step generates two files:

- `output_file.fa`: a FASTA file of the read sequences where T have been replaced by T, X or B according to the transition matrix (TM) approach ***(4)***,
- `output_file.fa_ratio_B`: a FASTA-like file with the BrdU ratio for each base of the sequence as computed by the neural network (CNN; see `overlap` argument above).

Mapping information is stored in the definition line (starting with '>') of each read in *output_file.fa* (*see* **Note 34**). Below is the description of the different fields.

- *mapped_chrom*: chromosome
- *mapped_start*: start coordinate
- *mapped_end*: end coordinate
- *mapped_strand*: strand on the reference (+/-)
- *clipped_bases_start*: number of bases clipped at the beginning of the read
- *clipped_bases_end*: number of bases clipped at the end of the read
- *num_matches*: number of exact matches between the read and the reference sequence
- *num_deletions*: number of deletions in the read compared to the reference
- *num_insertions*: number of insertions in the read compared to the reference
- *num_mismatches*: number of mismatches between the read and the reference sequence

3.6.3 Fork detection

To detect replication forks and their orientations, as well as replication initiation and termination events, you have to run the following command, where `BrdU_calls` is the folder containing the outputs of RepNano BrdU detection (from previous step, *see* **Note 35**), `DetectionFOLDER` is the location for the output files and `prefix` is a prefix you choose for the output file names (it can be a sample ID for instance) (*see* **Note 36**).

Fork detection

```
$ [conda activate RepNanoEnv]
$ python REPNANODIR/src/repnano/detection/ForkPrediction-CNN-TM.py
BrdU_calls DetectionFOLDER prefix
```

This will generate 6 files (forks, initiation and termination events for CNN and TM methods) with the events extracted from all the files contained in the 'BrdU_calls' folder. Below is the description of the fields.

- forks: `prefix_CNN.forks` and `prefix_TM.forks`
  - *chrom* : chromosome
  - *start* : first coordinate on the chromosome (independent of fork direction)
  - *end* : second coordinate on the chromosome
  - *direction* : fork direction (R, L for right, left, respectively) relative to the reference sequence
  - *FASTA* : file containing the read
  - *read* : read identifier
  - *strand* : strand on the reference

  - *JumpScore* : score reflecting the height of the signal, the higher the more confident one can be that this is a true fork and not noise.
  - *AsymScore* : score reflecting the asymetry of the signal, the higher the more confident one can be about the direction of the fork.

- initiation events: `prefix_CNN.init` and `prefix_TM.init`
  - *chrom* : chromosome
  - *middle* : estimated position of the initiation event. This is the midpoint between the starts of the two diverging forks.
  - *end1* : first coordinate of the first fork, for an initiation event this is the end of the leftward-going fork
  - *start1* : second coordinate of the first fork, for an initiation event this is the start of the leftward-going fork
  - *start2* : first coordinate of the second fork, for an initiation event this is the start of the rightward-going fork
  - *end2* : second coordinate of the second fork, for an initiation event this is the end of the rightward-going fork
  - *FASTA* : file containing the read
  - *read* : read identifier
  - *strand* : strand on the reference
- termination events: `prefix_CNN.term` and `prefix_TM.term`
  - *chrom* : chromosome
  - *middle* : estimated position of the termination event. This is the midpoint between the starts of the two converging forks.

- *start1* : first coordinate of the first fork, for a termination event this is the start of the rightward-going fork
- *end1* : second coordinate of the first fork, for a termination event this is the end of the rightward-going fork
- *end2* : first coordinate of the second fork, for a termination event this is the end of the leftward-going fork
- *start2* : second coordinate of the second fork, for a termination event this is the start of the leftward-going fork
- *FASTA* : file containing the read
- *read* : read identifier
- *strand* : strand on the reference

4. Notes

1. FORK-seq relies on the incorporation of BrdU from the external medium into yeast genomic DNA. Wild-type *Saccharomyces cerevisiae* cannot incorporate external nucleosides and therefore cannot be used for FORK-seq experiments. The minimal changes are the addition on a nucleoside transporter to allow the BrdU to enter the cells, and a thymidine kinase (TK) to phosphorylate it and make it a suitable substrate for DNA polymerases. We use the MCM869 yeast strain, which in addition to expressing human ENT1 and *Herpes simplex* TK ***(9;10)***, is deficient for the thymidylate synthase CDC21 ***(10;6)***. This makes it unable to produce its own thymidine and therefore fully dependent on external supply. In principle, any other strain able to incorporate BrdU at a high rate should be suitable to FORK-seq experiments.

2. The pH may vary depending of the YPD batch. If necessary, adjust it to 6.5-7.

3. MCM869 doubling time is about 150 minutes.

4. We have noticed that the 30 min thymidine starvation induces stress, as shown by Rad53 phosphorylation, which might be problematic depending on the biological question. Experiments without starvation nevertheless gave very similar BrdU profiles, albeit with very little labeling in early S-phase. Improved strains are under development.

5. In the original paper, we used two steps with different BrdU/thymidine ratios in order to orient the forks, but found out that this was dispensable as the changes in incorporation are not abrupt enough to resolve the steps; in fact, the signal shape is almost identical with a single 4 min BrdU pulse; the orientation of the forks instead relies on the asymmetrical rates of increase and decrease of BrdU incorporation during the pulse and the chase, respectively.

6. DNA extraction using QIAGEN Genomic-tip 20/G and QIAGEN Genomic DNA Buffer Set that contains buffers TE, Y1, G2, QBT, QC, and QF. Resuspend the frozen pellet into 1 mL Y1 buffer (Please refer to QIAGEN Genomic DNA Handbook for more details; do not use more than the equivalent of 100 mL of OD_{600}=1 cell suspension per column). Transfer to 2 mL-tubes. Add 100 µL of zymolyase (100U) and incubate at 37°C for 1h to 3h, inverting the tube from time to time. A white jelly fish should form. Centrifuge at 5000 g for 10 min at 4°C. Remove and discard the supernatant. Resuspend the pellet by slow pipeting in 2 mL G2 buffer with 10 µL RNAse A. Incubate for 15 min at 37°C. Add 20 µL Proteinase K solution (20 mg/mL stock) and incubate at 50°C for 1h. Invert the tube ~10 times to get a clear solution. Centrifuge at 5000 g for 10 min at 4°C and transfer the supernatant to a new tube. Equilibrate the QIAGEN Genomic-tip 20/G with 2 mL of QBT buffer. Vortex the sample 10s full speed and load on the column. Wash the column 3 times with 1 mL QC buffer. Put the column on a clean 5 mL tube and elute gDNA with 1 mL QF prewarmed to 50°C. Repeat this elution step. Precipitate the DNA adding 1,4 mL isopropanol and inverting the tube several times. Centrifuge at 5000 g for 30 min at 10°C. Remove the supernatant and wash the DNA pellet with 1 mL 70% EtOH. Let the pellet dry and resuspend it in 100 µL TE.

7. Ensure that the RNase used is completely free of DNase activity.

8. Using NanoDrop quantification is not recommended as it often overestimates the concentration.

9. Low A260/280 ratio indicates protein contamination and low A260/230 ratio indicates RNA contamination.

10. You may also use a low concentration agarose gel electrophoresis but the resolution of large DNA fragments is not as good. It still provides information about accidental fragmentation and RNAse treatment efficiency.

11. Possible causes for small DNA fragments: Every step has to be performed keeping in mind that high molecular weight DNA is fragile. Be gentle, do not vortex and avoid repeated pipetting. DNase contamination might also occur, be sure to use DNAse-free reagents and water.

12. It is recommended to start from 2 µg of DNA. When multiplexing the samples, it is possible to start from less DNA (down to 500 ng) but the number of sequencing reads might be reduced.

13. If no thermal cycler is available, a thermoblock can be used.

14. A smaller volume of beads can be used in order to remove more small DNA fragments, yet we noticed a significant loss of long DNA fragments when doing so.

15. Do not wash with 70% ethanol as it can damage molecular motors and result in low pore occupancy.

16. Set a P1000 pipette to 200 µL, insert the tip into the priming port and turn the wheel until the dial shows 220-230 µL or until you can see a small volume of buffer entering the pipette tip. Care must be taken when drawing back buffer from the flow cell. The array of pores must be covered by buffer at all times. Removing more than 20-30 µL risks a loss of sequencing channels. Be careful not to introduce any air bubble.

17. ONT recommends loading 5-50 fmol of a prepared library. Yet, this quantification depends on the size of the DNA molecules. Using 200 ng, we typically obtain 10-15 Gbp for non-barcoded libraries and 5-7 Gbp for barcoded ones. Barcoding reduces significantly the amount of resulting data and we recommend a maximum of 4 barcodes per run.

18. Pore activity is monitored in real time by MinKNOW. If pore activity becomes too low during the run (i.e. below 30%), it is possible to reload the flowcell with the leftover library. To do so, repeat steps 3 to 7.

19. It takes typically 6 days to base call a 72h run on the Mk1C (with Guppy 4.3, please be aware that ONT softwares evolve very quickly). For high-frequency usage of the sequencer, an independent workstation dedicated to base calling is recommended.

20. If a sequencing core facility is running your samples, inform them that you will need the raw FAST5 files, as they may do immediate archiving (or even deletion) of those big files.

21. Guppy base caller is available for Windows, Mac OS and Linux and both CPU and GPU versions are possible to get. We have tested only the Linux version using a dedicated workstation with 2x Intel Xeon Silver 4214 (12 cores), a NVIDIA GeForce RTX2080 TI graphic card (providing 4352 CUDA cores), 128 G of RAM and two 1 To SSD disks in RAID 0 for the data. Note that Guppy only runs on a selected set of GPUs (see Guppy documentation). With this configuration, the GPU version of Guppy base call a 4000-reads FAST5 file in a few minutes, while it takes more than 10h using the CPU version and 20 threads on an HPC cluster (depending on read length and CPU speed, typically 1 to 5 min per read on a single CPU). RepNano BrdU detection is routinely done using an HPC cluster, but we also have successfully tested the workflow on personal computers with Linux and Mac OS. The example provided correspond to using these Unix family operating systems.

22. We haven't tested the filtering based on read length although it seems quite reasonable to discard reads below 5 kb for replication fork detection with the present protocol.

23. You need a full member ONT account to access some of their softwares, including the Guppy base caller. You will get such account if you buy a sequencer. You can also join an existing account of a collaborator, see https://community.nanoporetech.com/.

24. ONT FAST5 API can be downloaded at https://github.com/nanoporetech/ont_fast5_api

25. To install Git please refer to https://git-scm.com/downloads.

26. To install Conda please refer to https://docs.conda.io/en/latest/miniconda.html.

27. Complete description of Tombo at https://nanoporetech.github.io/tombo/tombo.html.

28. Before 2019, the output of ONT base callers was different and one FAST5 was generated for each read. A procedure to work with the old files is available on our GitHub repository https://github.com/organic-chemistry/RepNano.

29. We use only one chromosome (chr4) for the reference in the test dataset, use a FASTA file containing the sequences of all the chromosomes for real data.

30. On Mac OS, if the command predict_simple.py fails to run and returns the error: *Initializing libiomp5.dylib, but found libomp.dylib already initialized.*, please add `nomkl` to the software environment using `$ conda install nomkl` (in RepNanoEnv).

31. If the directory where you want to save the output FAST5 doesn't exist, it will be automatically created.

32. This should not happen, please check that you give the correct path to the FASTQ file corresponding to the FAST5 file.

33. We recommend to replace 'N' bases in the reference FASTA by any canonical base, for instance using `$ sed 's/N/A/g' ref_with_N.fa > ref_corrected.fa`. This will replace all 'N' by 'A' bases.

34. Both outputs follow the orientation of the read and not the one of the reference.

35. All the files contained in the directory will be processed and a single list will be output.

36. A number of parameters are set up at the beginning of ForkPrediction-CNN-TM.py and can be modified to make the detection more or less stringent.

Acknowledgement

This work was supported by the Ligue Nationale Contre le Cancer, the Association pour la Recherche sur le Cancer, the Agence Nationale de la Recherche [ANR-15-CE12-0011-01, ANR-18-CE45-0002 and ANR-19-CE12-0028], the Fondation pour la Recherche Médicale [FRM DEI201512344404] and the Cancéropole Ile-de-France [PLBIO16-302].

The authors thank the sequencing facility (especially Corinne Cruaud and Stéfan Engelen) of François Jacob Institute of Biology (Genoscope, CEA, Evry) for their great help in setting up ONT nanopore sequencing in OH lab. We also thank IBENS IT support and Genomic Paris Centre platform for technical advices, as well as the computing cluster BioClust (Labex Memolife). Finally, we thank all members of the OH lab, especially Benoît Le Tallec (BT PhD supervisor), Laurent Lacroix and Etienne Jean for lively and supportive discussions.

References

1. Hyrien O (2015) Peaks cloaked in the mist: the landscape of mammalian replication origins. J Cell Biol 208, 2:147-60

2. Técher H, Koundrioukoff S, Azar D et al (2013) Replication Dynamics: Biases and Robustness of DNA Fiber Analysis. J Mol Biol 425, 23:4845-55

3. Guilbaud G, Rappailles A, Baker A et al (2011) Evidence for Sequential and Increasing Activation of Replication Origins Along Replication Timing Gradients in the Human Genome. PLoS Comput Biol 7, 12:e1002322

4. Hennion M, Arbona JM, Lacroix L et al (2020) FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. Genome Biol 21, 1:125

5. Muller CA, Boemo MA, Spingardi P et al (2019) Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. Nat Methods 16, 5:429–36

6. Ma E, Hyrien O and Goldar A (2012) Do replication forks control late origin firing in Saccharomyces cerevisiae? Nucleic Acids Res 40 5: 2010–9

7. Denis E, Sanchez S, Mairey B et al (2018) Extracting high molecular weight genomic DNA from Saccharomyces cerevisiae. Research Square Preprint, DOI:10.1038/Protex.2018.076

8. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 18:3094–3100

9. Lengronne A, Pasero P, Bensimon A et al (2001) Monitoring S phase progression globally and locally using BrdU incorporation in TK(+) yeast strains. Nucleic Acids Res 29, 7:1433–1442

10. Vernis L, Piskur J and Diffley JF (2003) Reconstitution of an efficient thymidine salvage pathway in Saccharomyces cerevisiae. Nucleic Acids Res 31, 19:e120

11. Siow CC, Nieduszynska SR, Muller CA et al (2012) OriDB, the DNA replication origin database updated and extended. Nucleic Acids Res 40:682–686

Figure caption

Figure 1 : Illustration of FORK-seq. In purple are shown the BrdU incorporation signals from reads containing an initiation event mapped within a 100 kb window on yeast chromosome IX. Green bars represent known origins from OriDB **(11)**. Purple arrows show fork orientation and green triangles highlight termination events.