Published online 19 June 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 2 **1**
doi: 10.1093/nargab/lqaa045

Efficient, quick and easy-to-use DNA replication timing analysis with START-R suite

Djihad Hadjadj[1,†], Thomas Denecker[2,†], Eva Guérin[1], Su-Jung Kim[1], Fabien Fauchereau[1], Giuseppe Baldacci[1], Chrystelle Maric[1,‡] and Jean-Charles Cadoret[1,*,‡]

[1]Pathologies de la Réplication de l'ADN, Université de Paris; Institut Jacques-Monod, UMR7592, CNRS, F-75006 Paris, France and [2]Institut de Biologie Intégrative de la Cellule, UMR9198, CNRS, Université Paris-Saclay, Université Paris-Sud, F-91405 Orsay, France

Received February 20, 2020; Revised May 19, 2020; Editorial Decision May 28, 2020; Accepted June 15, 2020

ABSTRACT

DNA replication must be faithful and follow a well-defined spatiotemporal program closely linked to transcriptional activity, epigenomic marks, intranuclear structures, mutation rate and cell fate determination. Among the readouts of the spatiotemporal program of DNA replication, replication timing analyses require not only complex and time-consuming experimental procedures, but also skills in bioinformatics. We developed a dedicated Shiny interactive web application, the START-R (Simple Tool for the Analysis of the Replication Timing based on R) suite, which analyzes DNA replication timing in a given organism with high-throughput data. It reduces the time required for generating and analyzing simultaneously data from several samples. It automatically detects different types of timing regions and identifies significant differences between two experimental conditions in ∼15 min. In conclusion, START-R suite allows quick, efficient and easier analyses of DNA replication timing for all organisms. This novel approach can be used by every biologist. It is now simpler to use this method in order to understand, for example, whether 'a favorite gene or protein' has an impact on replication process or, indirectly, on genomic organization (as Hi-C experiments), by comparing the replication timing profiles between wild-type and mutant cell lines.

INTRODUCTION

DNA replication is a highly regulated process involved in the maintenance of genome stability (1–3). Its accuracy relies partly on a spatio-temporal program that regulates timing and location of origin firing (4,5). Based on this program, replication is organized into large-scale domains that replicate at different times in S phase (6–8). Protocols developed to study the replication timing (RT) in specific cell lines have been established in different laboratories (9–13). A script dedicated to RT analysis was previously developed by David Gilbert's laboratory (10,12), but it required skills in bioinformatics and R language. In order to make the analysis of experimental results easier for biologists, we implemented an interactive suite, START-R (Simple Tool for the Analysis of the Replication Timing based on R) Analyzer and START-R Viewer, showing user-friendly interfaces. This START-R suite is dealing with RT experiments made with microarrays or with Repli-seq data (either Early/Late or S/G1 ratios) from different organisms. These web applications would make easier differential analyses of RT, by comparing conditions such as treated/untreated cells or mutated/normal cells. They are free and may be improved by developers, according to specific needs. In addition, RT profiles correlate with A/B compartment profiles predicted by chromosome conformation methods. Regions of the genome defined by Hi-C profiles as A compartments are also identified as Early replicated domains, whereas regions defined as B compartments are Late replicating domains (14). Furthermore, some replicating domains coincide with a subset of topologically associating domains and more closely with the ones located at compartment boundaries (15). Studies of RT programs with START-R suite take shorter time to perform and therefore open new research perspectives for many laboratories working in DNA replication, in chromosome conformation and in other closely related molecular processes.

MATERIALS AND METHODS

START-R suite

START-R Analyzer and START-R Viewer (doi:10.5281/zenodo.3251905) were developed using the Shiny R package (W. Chang, J. Cheng, J. Allaire,

*To whom correspondence should be addressed. Tel: +33 1 57 27 80 74; Email: jean-charles.cadoret@ijm.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Y. Xie and J. McPherson, Shiny: web application framework for R, R package version 1.2.0, 2018, https://CRANR-project.org/package=shiny). The source code and the installation procedure are available on GitHub (https://github.com/thomasdenecker/START-R). All R packages used in the START-R suite are listed in the Readme file on GitHub. Users should install Docker and then follow installation procedures described for each operating system (Windows, Linux and Mac OS) in the Readme file (Figure 1). Although the installation of these web-based applications has been simplified as much as possible, it may even so require some computer knowledge (especially on Linux). Once installed, the START-R suite is an easy-to-use tool requiring no computer skills (see Supplementary Figure S1 and the Wiki section on GitHub for more details).

Validation of START-R suite using microarray data from other laboratories

We analyzed with the START-R suite the microarray data obtained by Hiratani *et al.* (16) of D3esc and D3npc9 cell lines during mouse cells' differentiation (GEO accession numbers: GSM450273 and GSM450285, respectively). As data extracted from the Nimblegen platform are in PAIR format, we used a script to convert data into a valid format for START-R Analyzer (convertPair.R, available on GitHub in 'supplement script' file, https://github.com/thomasdenecker/START-R).

Early/Late-seq data and conversion

In order to validate our software, we used data corresponding to Repli-seq 46C mouse cells—Early S fraction or Late S fraction, respectively (GSM2496038 and GSM2496039). Read mapping was obtained using Bowtie2 (2.3.4.2 version) with the very sensitive end-to-end option. Then, PCR duplicates were removed by RmDUP from SAMTools (2.0.1 version). BamCoverage parameters were fixed (3.1.2.0.0 version with default parameters) to a bin size of 10 kb (corresponding to the genomic distance between microarray probes) and reads per kilobase million (RPKM) normalization to generate a bedGraph file. A headline was added to the file to name the four columns (chr, start, end and gProcessedSignal for Early file or rProcessedSignal for Late file, respectively). Then, a script to convert and merge the bedGraph files from Early and Late samples to a format compatible with START-R Analyzer was developed. This script is available on GitHub in 'supplement script' file as convert_bamcoverage_file.R.

Validation of START-R suite using S/G1 data from multiple species

Different laboratories analyze variations of DNA copy number between G1 and S phase cells (S/G1 ratio) to study the RT program with Repli-seq. We used data obtained from different organisms such as *Drosophila*, zebrafish and humans (17–19), to validate the START-R suite (GSM3154888, GSM3154890, GSM2282090, SRX3413939-40). As previously, BAM coverage files from G1 and S fractions are converted to a format compatible with START-R Analyzer with the aforementioned 'convert_bamcoverage_file.R.' script.

RESULTS

RT analysis with START-R suite allows robust statistical analysis

We developed a software allowing an automatic detection of RT regions and a differential analysis between two conditions. The START-R suite is implemented into an HTML interface for more efficient use and sharing by biologists. START-R is built-in and packaged into a virtual environment with Docker (Figure 1). Thus, START-R can be easily deployed on a personal computer or on a server, and can run independently of any library updating. START-R provides as many parameters as possible for a comprehensive analysis of the RT program (Supplementary Figure S1A–K). Furthermore, we added new scaling, normalization and smoothing methods (Supplementary Figure S2) and also novel statistical approaches to detect differences between two samples. A classical differential analysis performed with START-R takes ∼15 min with a standard laptop computer. START-R Analyzer can run data from all organisms with different genome assemblies. This flexibility is one of the new aspects of the START-R suite that allows to analyze RT program in every organism (Supplementary Figure S1B). In addition, START-R generates RT profiles in PDF and all the files necessary for a better and customized visualization with START-R Viewer. START-R Analyzer also produces specific files that could be visualized with START-R Viewer, a specific genome browser tool dedicated to START-R suite.

A large panel of new settings and tools for RT analysis

We based our method on four major steps: *normalization* (between Early and Late fractions, between two replicates and between two independent experiments; Supplementary Figure S2A), *smoothing* (different options are available; Supplementary Figure S2B), *identification* of transition timing regions (TTRs; Supplementary Figure S3) and *segmentation*. The originality of our approach is to first detect TTRs in order to better identify constant timing regions (CTRs; Supplementary Figure S3). The identification of TTRs is based on their intrinsic properties: regions that include more than three consecutive probes with significantly different Early/Late intensity log ratios are considered as TTRs (Supplementary Figure S3). The statistical significance of differences between intensity log ratios is calculated by the outlier box plot method (Supplementary Figure S4) (20). Following TTR detection, START-R Analyzer localizes CTRs: TTRs are subtracted from the genome (Supplementary Figure S3) and the remaining regions are considered as CTRs. At the end of these steps, START-R Analyzer automatically generates a BED file for CTRs and TTRs making easier further bioinformatic analyses and the display of the RT domains via a genome browser (Supplementary Figure S1F). A codebook is also generated to ensure the traceability of options chosen for each analysis.

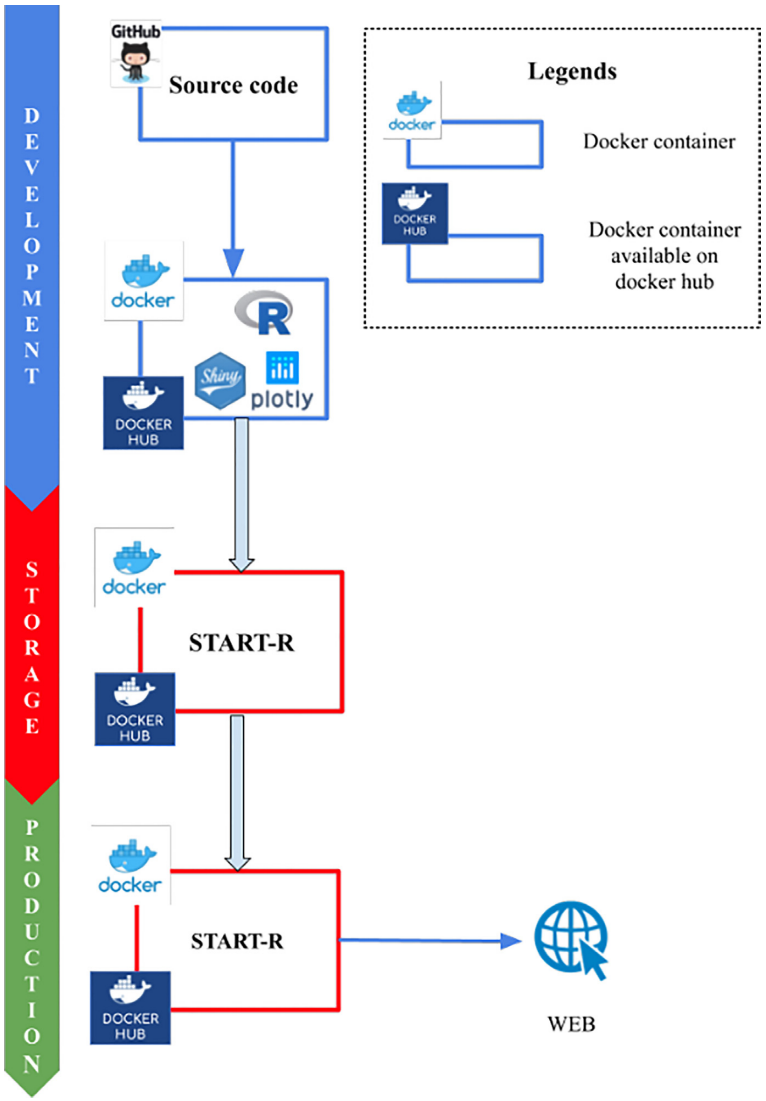

Figure 1. Stack overview of the START-R suite. The START-R suite is able to analyze all types of genome-wide RT data formats like microarray data, Repli-seq data with Early/Late and S/G1 ratios, and RT data from multiple organisms. The START-R suite was developed using the Shiny R package (W. Chang, J. Cheng, J. Allaire, Y. Xie and J. McPherson, Shiny: web application framework for R, R package version 1.2.0, 2018, https://CRANR-project.org/package=shiny) and Plotly visualization tools (C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec and P. Despouy, plotly: create interactive web graphics via 'plotly.js', R package version 4.7.1, 2017, https://CRANR-project.org/package=plotly). The START-R software (START-R Analyzer and START-R Viewer) are open-source web-based applications (doi:10.5281/zenodo.3251905). For the storage and production steps, the START-R suite was concatenated using Docker in order to install, use and share it easily. These software can be used with different operating systems: Windows, Mac OS and Linux. The source code and the installation procedure are available on GitHub (https://github.com/thomasdenecker/START-R). To install the START-R suite, users should install Docker and follow the Readme file containing the installation procedure (https://github.com/thomasdenecker/START-R/blob/master/README.md). Finally, in order to run the START-R suite, users should double-click on the START-R file (Windows) or launch the command line (Mac OS X, Linux), followed by opening an internet browser at the following URLs: http://localhost:3838/ for START-R Analyzer and http://localhost:3839/ for START-R Viewer.

We added a step allowing the differential analysis of RT programs from two experiments. Thus, we can now compare RT profiles obtained in different conditions and/or with different cell lines to identify genomic elements that can modify the RT program. Our differential analysis includes three different methods of comparison: the Mean method, the Euclidean method and the Segment comparison method. When the goal is to identify most regions with strong RT changes, we recommend using the most stringent Mean method. The less stringent Segment and Euclidean methods allow the detection of a larger set of RT changes, while increasing the risk of obtaining false positives.

The last major implementation is START-R Viewer (Supplementary Figure S1K). This web-based interface allows the visualization of the RT profile generated by START-R Analyzer in dynamic charts obtained with the Plotly library (C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec and P. Despouy, plotly: create interactive web graphics via 'plotly.js', R package version 4.7.1, 2017, https://CRANR-project.org/package=plotly).

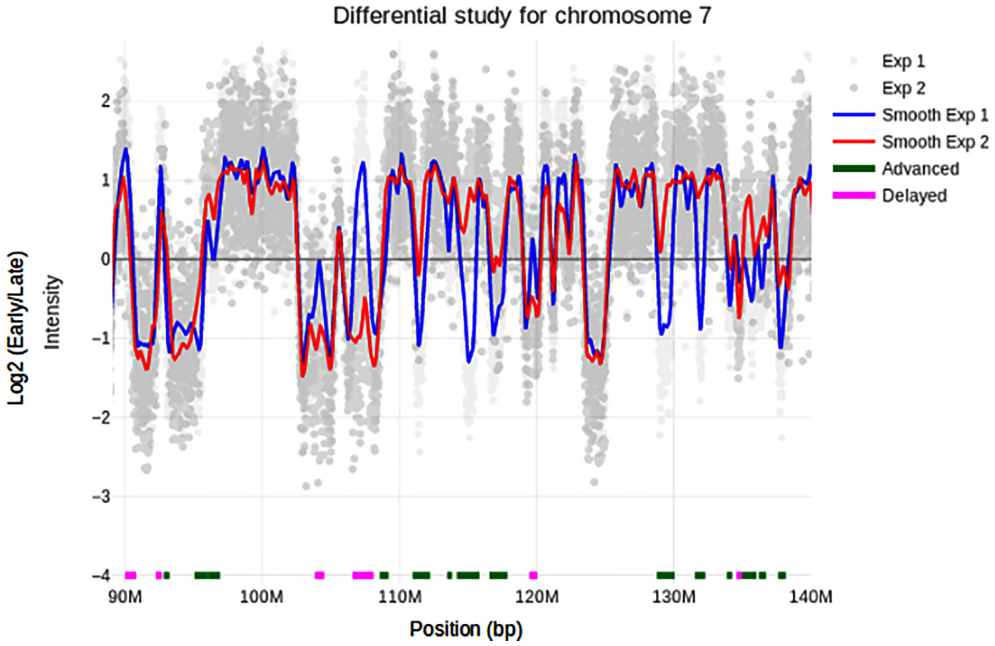

Figure 2. Genomic characteristics of regions harboring different RT programs. START-R Viewer allows the visualization of RT data with many features. The START-R differential analysis of RT profiles is shown here for a portion of chromosome 7 in mouse D3esc (blue) and D3npc9 (red) cells. We used START-R Analyzer with the standard options: Loess Early/Late normalization, scale inter-replica normalization, inter-experiment standardization, Loess method for smoothing (span = 300 kb), 2.5 for SD difference between two segments and mean comparison analysis with a Holm's method *P*-value of 0.05 for the differential analysis. The advanced (green) and delayed (pink) regions identified with START-R Analyzer are indicated underneath the RT profiles. Light gray and gray spots indicate data from both RT experiments. With these parameters, 2066 CTRs were detected in the genome and 910 regions showed different RT between D3esc and D3npc9 cells. Box plots illustrating genomic characteristics (GC content, LINE-1 content and gene coverage) of regions harboring different RT programs are shown in Supplementary Figure S6.

One can easily identify CTRs, TTRs (Supplementary Figure S5A) and significantly advanced or delayed regions (Supplementary Figure S5B). We therefore developed a genome browser to optimally display the maximum of data generated by START-R Analyzer. The START-R suite also automatically generates files with different output formats essential for further molecular characterizations and compatible for classical bioinformatic tools and/or for GALAXY genomic tools (21).

START-R analysis of RT programs during differentiation in mouse: a new analysis of previous data

To validate our START-R based-approach without *a priori* consideration, we decided to re-analyze the data generated by the Gilbert's group concerning the changes of RT program during cell differentiation in mouse D3esc and D3npc9 cell lines (16). We converted these raw data with the convertPair.R script available in our GitHub project into the correct format for START-R Analyzer. RT profiles generated with START-R Analyzer (Figure 2) and molecular signatures (Supplementary Figure S6) are identical to the ones previously described by Gilbert's group. Each modified timing region had a particular molecular signature: Late-to-Early or advanced regions show a GC/LINE-1 density and gene coverage similar to constant Early regions, while Early-to-Late or delayed regions showed GC/LINE-1 density and gene coverage similar to constant Late regions.

Validation of START-R with Early–Late Repli-seq data from mouse

Nowadays, many data of RT program are generated with Repli-seq experiments, but their analysis is time-consuming and often requires bioinformatics skills. We analyzed the Early/Late Repli-seq data from Marchal *et al.* (12). We specifically developed a supplemental script to convert the BAM coverage file to a log Early/Late file (convert_bamcoverage_file.R) to be sure that the integration into the START-R pipeline was correct. Then, we compared the RT smooth profiles generated from Early/Late Repli-seq data with those generated by the same group using microarrays (Figure 3A). Profiles are almost identical to the ones described by Marchal *et al.* (12). Thus, START-R Analyzer and START-R Viewer can be easily used to analyze Early/LateIDEX Repli-seq data, showing their versatility and their simplicity of use.

Validation of START-R with S–G1 Repli-seq data from *Drosophila*, zebrafish and humans

Other laboratories use the ratio of DNA content between G1 and S phases to analyze the RT program. We wanted to know whether START-R suite can run the correct analyses with this type of data and also with other organisms than mice and humans. We performed exactly the same pipeline used for Early/Late Repli-seq data described above with *Drosophila*, zebrafish and human S/G1 data (17–19). Then and as expected, START-R can be run with S/G1 log ratio

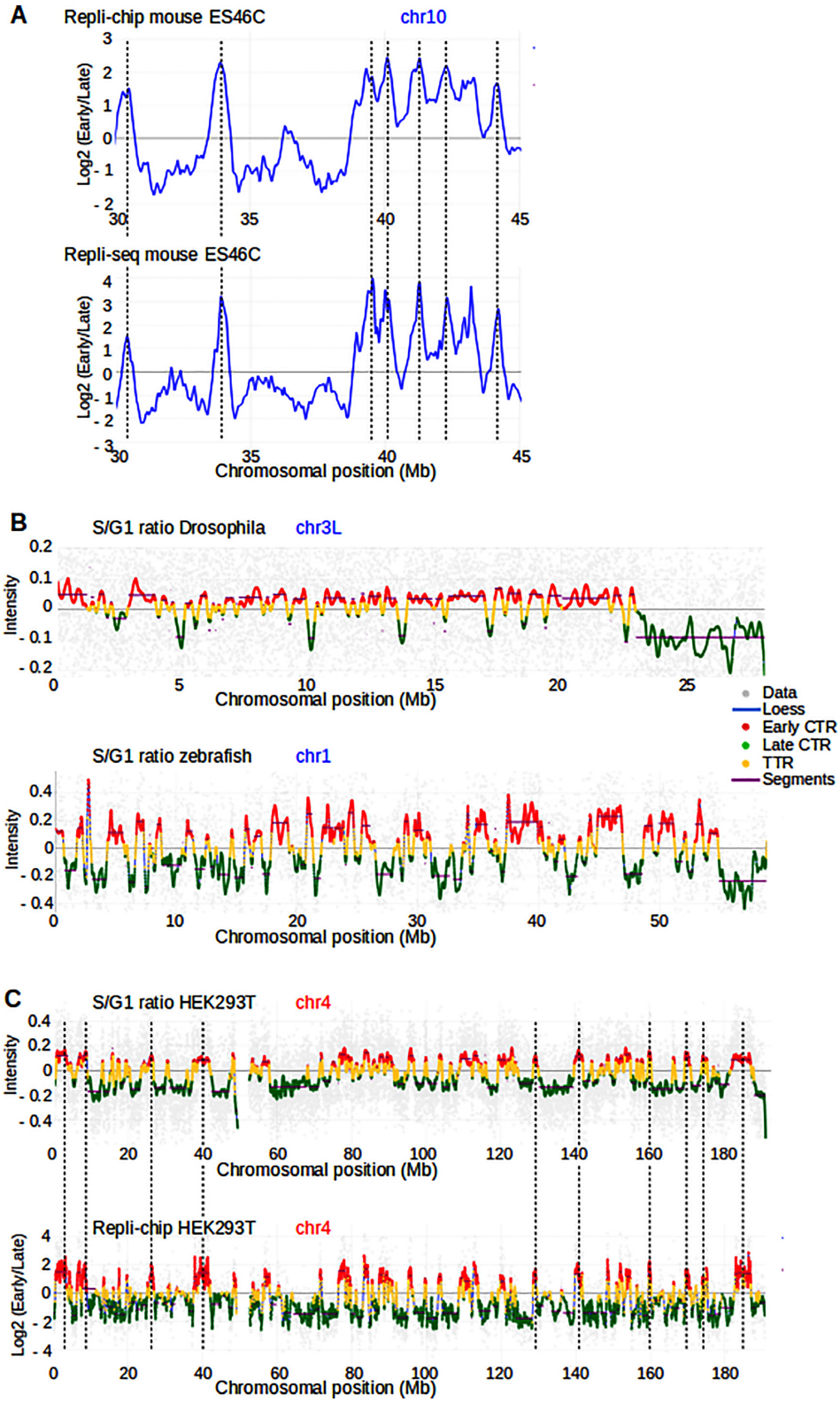

Figure 3. The START-R suite allows analysis and visualization of both Repli-chip and Repli-seq data from different model systems. (**A**) RT profiles of a portion of mouse chromosome 10 from ES46C cell line are generated using Repli-chip (top panel) and Repli-seq data (bottom panel) with START-R web applications. Dashed vertical lines show common RT regions between both profiles. (**B**) RT profiles obtained by S/G1 ratios are shown for the left part of *Drosophila* chromosome 3 (3L) and for zebrafish chromosome 1 (blue lines). The profiles display distribution of Early and Late CTRs, in red and green, respectively, and of TTRs, in yellow. Segments corresponding to regions of constant timing are shown in purple. Gray spots indicate data from RT experiments. (**C**) RT profiles of human HEK293T chromosome 4 are generated using S/G1 ratio and Repli-chip data. The empty space inside the RT profiles represents the centromere region.

data (Figure 3B and C). We observed similar profiles as the ones already observed for these different organisms.

DISCUSSION

In this study, we show a new automated protocol for analyzing RT profiles (Figure 1), obtained with different methods, in all organisms. As a proof of concept, we succeed in generating RT analyses for human, mouse, *Drosophila* and zebrafish genomes (Figures 2 and 3). START-R suite's user-friendly interface allows choices between different parameters at all steps used to generate RT profiles (Supplementary Figure S1). Compared to the previous methods (10,12), START-R Analyzer first detects TTRs and thus better refines and improves the CTR detection (Supplementary Figures S3 and S5A). In addition, START-R Analyzer contains new calculation methods for differential analyses between two conditions or cell lines (Supplementary Figure S5B).

This flexibility gives the users the opportunity to choose the differential analysis method and different parameters according to their questions. It also automatically generates files with different output formats essential for further molecular characterizations and compatible for classical bioinformatic tools and/or for GALAXY genomic tools (21). Then, START-R Viewer produces a nice interface to visualize all the data generated by START-R Analyzer. Furthermore, START-R suite freeware are available on GitHub and their source codes are open to anyone who wants to improve them, as, for example, for studies of allelic changes of RT.

In conclusion, it is now possible for any biologist or laboratory to readily explore new or previous RT data simply and quickly. Thus, a large number of laboratories can today use our software to find out whether their experimental conditions are affecting the RT process or are correlated with other molecular mechanisms. START-R also allows to determine what parts of the genome are impacted and in which proportion and to characterize further those loci. Thanks to the accessibility of our approaches and software, their speed and efficiency, new research perspectives can be efficiently envisaged.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We thank Gaëlle Lelandais for helpful discussions. We also acknowledge the ImagoSeine core facility of the Institut Jacques-Monod, member of the France BioImaging (ANR-10-INBS-04) supported by the Region Île-de-France (E539). This project was supported by the generous legacy from Ms Suzanne Larzat to our group.

FUNDING

La Ligue Nationale Contre le Cancer RS16/75-108, RS17/75-135; GEFLUC; Institut National du Cancer INCa-10493; IDEX Université de Paris ANR-18-IDEX-0001.

Conflict of interest statement. None declared.

REFERENCES

1. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
2. Macheret,M. and Halazonetis,T.D. (2015) DNA replication stress as a hallmark of cancer. *Annu. Rev. Pathol. Mech. Dis.*, **10**, 425–448.
3. Técher,H., Koundrioukoff,S., Nicolas,A. and Debatisse,M. (2017) The impact of replication stress on replication dynamics and DNA damage in vertebrate cells. *Nat. Rev. Genet.*, **18**, 535–550.
4. Dileep,V., Rivera-Mulia,J.C., Sima,J. and Gilbert,D.M. (2015) Large-scale chromatin structure–function relationships during the cell cycle and development: insights from replication timing. *Cold Spring Harb. Symp. Quant. Biol.*, **80**, 53–63.
5. Rivera-Mulia,J.C. and Gilbert,D.M. (2016) Replicating large genomes: divide and conquer. *Mol. Cell*, **62**, 756–765.
6. Ryba,T., Hiratani,I., Lu,J., Itoh,M., Kulik,M., Zhang,J., Schulz,T.C., Robins,A.J., Dalton,S. and Gilbert,D.M. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.*, **20**, 761–770.
7. Cornacchia,D., Dileep,V., Quivy,J.P., Foti,R., Tili,F., Santarella-Mellwig,R., Antony,C., Almouzni,G., Gilbert,D.M. and Buonomo,S.B.C. (2012) Mouse Rif1 is a key regulator of the replication-timing programme in mammalian cells. *EMBO J.*, **31**, 3678–3690.
8. Desprat,R., Thierry-Mieg,D., Lailler,N., Lajugie,J., Schildkraut,C., Thierry-Mieg,J. and Bouhassira,E.E. (2009) Predictable dynamic program of timing of DNA replication in human cells. *Genome Res.*, **19**, 2288–2299.
9. Hansen,R.S., Thomas,S., Sandstrom,R., Canfield,T.K., Thurman,R.E., Weaver,M., Dorschner,M.O., Gartler,S.M. and Stamatoyannopoulos,J.A. (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 139–144.
10. Ryba,T., Battaglia,D., Pope,B.D., Hiratani,I. and Gilbert,D.M. (2011) Genome-scale analysis of replication timing: from bench to bioinformatics. *Nat. Protoc.*, **6**, 870–895.
11. Dileep,V., Didier,R. and Gilbert,D.M. (2012) Genome-wide analysis of replication timing in mammalian cells: troubleshooting problems encountered when comparing different cell types. *Methods*, **57**, 165–169.
12. Marchal,C., Sasaki,T., Vera,D., Wilson,K., Sima,J., Rivera-Mulia,J.C., Trevilla-García,C., Nogues,C., Nafie,E. and Gilbert,D.M. (2018) Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat. Protoc.*, **13**, 819–839.
13. Petryk,N., Kahli,M., d'Aubenton-Carafa,Y., Jaszczyszyn,Y., Shen,Y., Silvain,M., Thermes,C., Chen,C.L. and Hyrien,O. (2016) Replication landscape of the human genome. *Nat. Commun.*, **7**, 10208–10220.
14. Miura,H., Takahashi,S., Poonperm,R., Tanigawa,A., Takebayashi,S.I. and Hiratani,I. (2019) Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nat. Genet.*, **51**, 1356–1368.
15. Marchal,C., Sima,J. and Gilbert,D.M. (2019) Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.*, **20**, 721–737.
16. Hiratani,I., Ryba,T., Itoh,M., Yokochi,T., Schwaiger,M., Chang,C.W., Lyou,Y., Townes,T.M., Schübeler,D. and Gilbert,D.M. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.*, **6**, 2220–2236.
17. Armstrong,R.L., Penke,T.J.R., Strahl,B.D., Matera,A.G., McKay,D.J., MacAlpine,D.M. and Duronio,R.J. (2018) Chromatin conformation and transcriptional activity are permissive regulators of DNA replication initiation in *Drosophila*. *Genome Res.*, **11**, 1688–1700.
18. Siefert,J.C., Georgescu,C., Wren,J.D., Koren,A. and Sansam,C.L. (2017) DNA replication timing during development anticipates

transcriptional programs and parallels enhancer activation. *Genome Res.*, **8**, 1406–1416.
19. Massey,D.J., Kim,D., Brooks,K.E., Smolka,M.B. and Koren,A. (2019) Next-generation sequencing enables spatiotemporal resolution of human centromere replication timing. *Genes (Basel)*, **10**, E269.
20. Krzywinski,M. and Altman,N. (2013) Error bars. *Nat Methods*, **10**, 921–922.
21. Afgan,E., Baker,D., Batut,B., Van Den Beek,M., Bouvier,D., Ech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.