



# Replicon: a software to accurately predict DNA replication timing in metazoan cells

Yevgeniy Gindin<sup>1,2</sup>, Paul S. Meltzer<sup>1</sup> and Sven Bilke<sup>1\*</sup>

<sup>1</sup> Genetics Branch, Center for Cancer Research, National Institutes of Health, Bethesda, MD, USA

<sup>2</sup> Graduate Program in Bioinformatics, Boston University, Boston, MA, USA

## Edited by:

Alexandre V. Morozov, Rutgers University, USA

## Reviewed by:

Jian Ren, Sun Yat-sen University, China

Dapeng Wang, University College London, UK

## \*Correspondence:

Sven Bilke, Genetics Branch, Center for Cancer Research, National Institutes of Health, 37 Convent Drive, Bethesda, MD 20892, USA  
e-mail: bilkes@mail.nih.gov

Eukaryotic DNA replication follows a strict temporal program where genomic loci are replicated at precise times during the S phase of the cell cycle. Yet, the mechanism in control of the timing program in metazoan cells is poorly understood. In a recent publication, the authors proposed an intuitive stochastic model of DNA replication and showed that it predicts replication timing with an accuracy approaching the level of experimental biological repeats. Here, we discuss an extended software implementation of the mechanistic model: Replicon. This package allows interested researchers to predict the global replication timing program in human cells from chromatin data.

**Keywords:** DNA replication, cell cycle, mathematical modeling

## 1. INTRODUCTION

The strict temporal order of DNA replication in metazoan cells has been observed for over half a century (Taylor, 1960). Yet, relatively little is known about what governs this process (for recent review, see Masai et al., 2010). Studies of fission yeast suggest that DNA replication is initiated in a time-stochastic manner, where the global temporal patterns of DNA replication timing arise from uncoordinated initiation events (Dai et al., 2005). While DNA replication timing has been modeled successfully in fission yeast (Hyrien and Goldar, 2010), no such model has existed for metazoan cells.

Among the roadblocks to modeling DNA replication timing in metazoan cells has been the lack of a comprehensive catalog of the locations and amplitudes of replication initiation (Martin et al., 2011; Besnard et al., 2012). Yet, even if such a catalog were to become available, the resulting model would require more than 100,000 parameters (Pope et al., 2013) (at least one per initiation site)—a process that would have to be repeated for new cell lineages and species.

Recently we have shown that DNA replication initiation sites are sufficiently localized by DNase hypersensitivity (HS) and that the amplitude of replication initiation at individual sites has negligible effect on the global DNA replication timing program (Gindin et al., 2014). The current work showcases a vastly expanded computational model of DNA replication timing in human cells—Replicon. To the best of our knowledge, Replicon is the only mechanistic model of replication timing in mammalian cells. Unlike any other model, it can predict, rather than reproduce, replication timing because our model has no free parameters (with the exception of one technical constant). We have shown earlier for human cells that based only on the knowledge of DNase hypersensitive sites, our model accurately predicts replication timing for arbitrary cell types (Gindin et al., 2014).

There are a number of mechanistic models for the simpler genomes of various yeast strains (See Hyrien and Goldar, 2010, for a recent review). They all rely on a detailed characterization of replicators, the sites of replication initiation. In addition to replicator locations, typically two more parameters are required per site, firing efficiency and average firing time. Unfortunately, metazoan genomes may contain more than 100,000 initiation sites, and metazoan replicators remain relatively poorly characterized. To use these models for metazoan cells therefore requires not only to fit several 100,000 parameters in order to reproduce timing data, but this procedure would have to be repeated again for each cell type. The same is true for another interesting approach aiming at identifying a histone code of replication timing in *Drosophila melanogaster* (Comoglio and Paro, 2014). Here, too, the authors were able to reproduce the timing behavior in a given cell type, but the authors found that model parameters had to be re-adjusted for every cell type.

Our earlier publication (Gindin et al., 2014) included a core-version of our timing simulator. Since then we have substantially extended this simulator and added a new software suite, RepliconWrench. The core simulator, Replicon now allows to observe events in single cells as well as the entire cell ensemble. We have added simulations for a number of new measurement devices, including “DNA combing” experiments (Herrick et al., 2002), nascent strand bubble-chip and bubble-seq experiments to measure local initiation rates (Martin et al., 2011; Valenzuela et al., 2012), time of replication initiation measurements as well as global replication timing experiments that include a flow sorter (Ryba et al., 2011). To facilitate large-scale analyses, we implemented RepliconWrench—a software accessory tool that generates DNA replication initiation probability landscapes, a required Replicon input (see Methods), by querying relational databases of genome annotations or processing genome interval files.

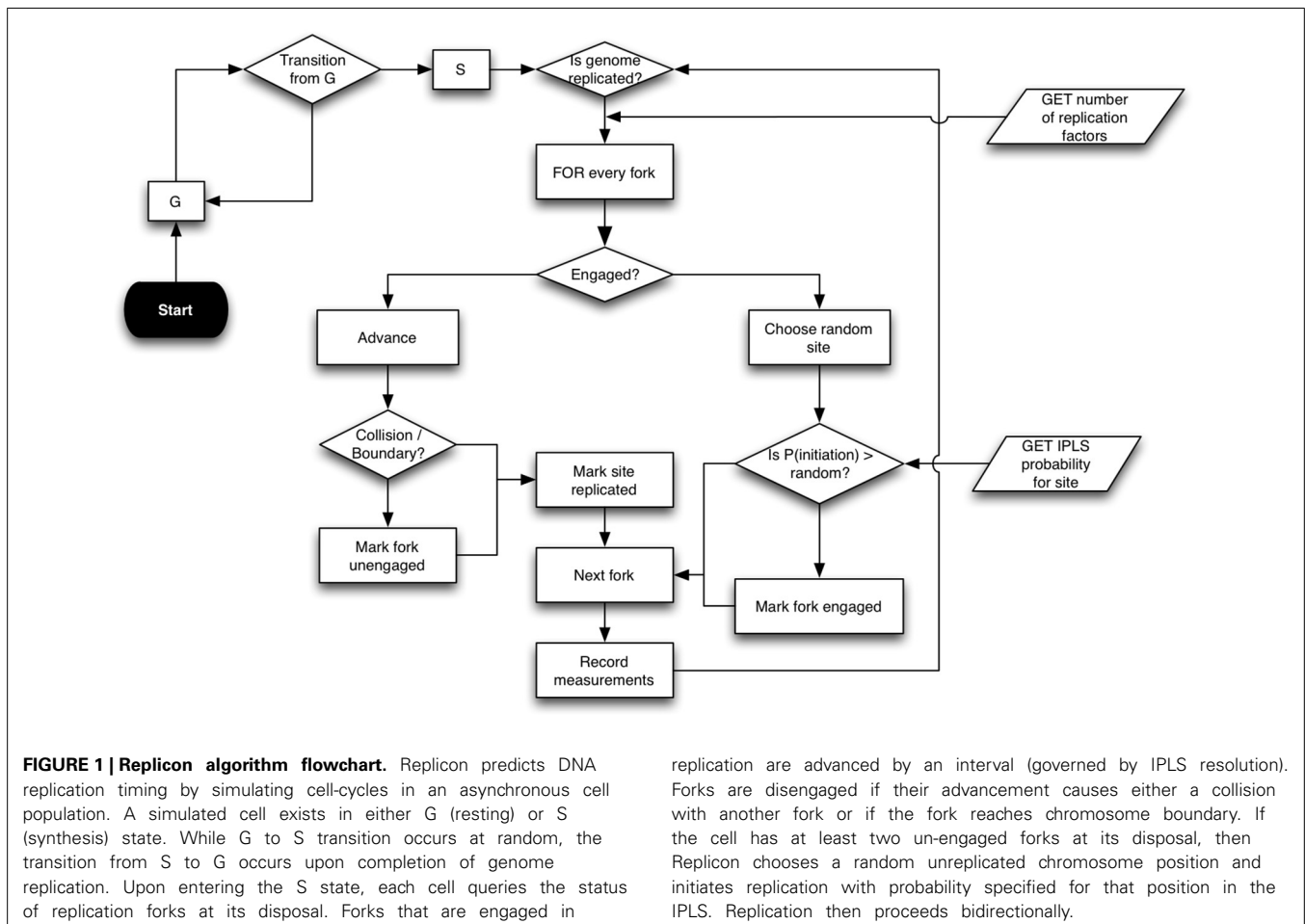
## 2. MATERIALS AND METHODS

DNA replication is simulated on a cell population consisting of independent cells (**Figure 1**). A Replicon-simulated cell exists in two states: S, when it is actively replicating its genome; and G, when it is at rest. The simulation starts with all cells in G state. Cells transition from G to S following an interval drawn from a normal distribution. There are two required inputs to Replicon. The inputs to each cell are: (1) sites of potential DNA replication initiation events; and (2) a number of replication factors (see **Table 1** for a complete list of command-line arguments).

The first input specifies the number of replication factors that bind to initiation sites. These factors replicate DNA bidirectionally at a constant velocity of 50 bp/s. Replication factors are recycled when they collide with one another or when they reach a chromosome's physical boundary. The second input, Initiation Probability Landscape (IPLS), specifies the probability of replication initiation for every genome locus. At each step in the cell-cycle simulation, a replication factor chooses a random, unreplicated position in the genome and begins replication with a probability specified for that locus in the IPLS. In our earlier study (Gindin et al., 2014) we had shown that the choice of exact number of replication factors parameter is not critical and that using one fork per 80 MB yielded accurate replication timing predictions for human cells.

The IPLS file format is a simple two column file where the first column specifies a genome location and a second column specifies a probability of initiating replication at that location. Chromosome designation need not be specified in the IPLS file as Replicon simulates replication on one chromosome at-a-time. The choice of genome intervals in the IPLS files, which determines the resolution of the resulting replication time predictions, is user-driven. In our work, we found that a 500 bp IPLS resolution is well-suited for comparison with existing experimental data (see the accompanying Replicon web-site for a through description). IPLS files may be generated from “bedgraph” files either with a custom script or with a provided Replicon accessory tool—RepliconWrench. The RepliconWrench suite handles conversion of “bedgraph” or similarly formatted files into chromosome-specific IPLSs. Additionally, RepliconWrench could be used to query relational databases and generate IPLSs directly from genome annotations, such as DNase HS sites.

Replicon is written in C++ and supports multi-threading. In our simulations, Replicon was executed on a moderately sized SGE-managed computing cluster. A whole-genome simulation of replication timing for human cells required 15 min when running on 22, 4-core 2.93 GHz Linux nodes (one node per chromosome). RepliconWrench is written in Java and is compatible with Java 1.5 or higher.



**Table 1 | Replicon command-line parameters.**

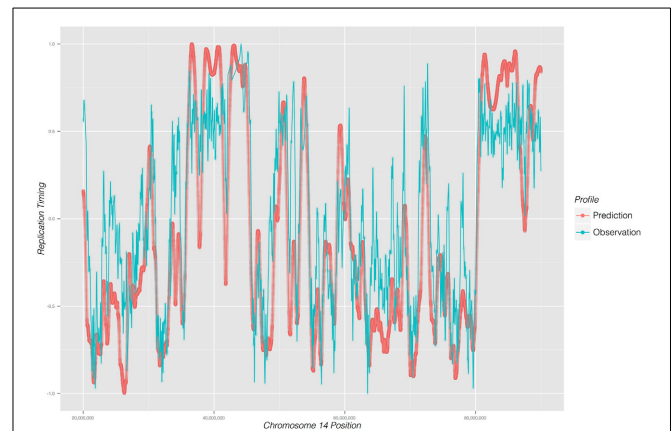
Parameter name	Parameter action
Flow-sorter	Comma-separated list of flow-sort boundaries
Indices	Measure S-phase and replicator engagement fraction
Initiate	Measure the global initiation rate as a function of S-phase fraction
Nascent	Simulate nascent strand measurements
Ncells	Number of cells (default: 1000)
Nfork	Number of replication forks (default: 50)
Nmeas	Number of measurements (default: 2000)
Ntherm	Number of thermalization sweeps (default: 80000)
Overwrite	Overwrite result file
Pprogress	Probability to move replication fork (ignored if 2nd column is present in landscape file) (default: 1)
Prelease	Probability to release replication fork (ignored if 3rd column is present in landscape file) (default: 0)
Singlemolecule	Simulate single molecule experiments
Stepsize	Sweeps per measurement (default: 20000)
Threads	Number of threads (default: 1)
Timing	Perform direct timing measurements

### 3. RESULTS

Replicon's accuracy can be ascertained by calculating a correlation coefficient between its prediction and experimentally observed data. For that purpose we use the mouse ENCODE data (ENCODE Project Consortium and others, 2011) to obtain empirical DNA replication timing data. By way of demonstrating Replicon, we predict a DNA replication timing profile for mouse stem cells using DNase DGF data as input (Sabo et al., 2004) and compare the prediction with experimental observations (Hiratani et al., 2008).

As illustrated in **Figure 2**, Replicon's predicted DNA replication timing profile closely reproduces experimentally observed DNA replication data. The correlation between Replicon's prediction and empirical data for mouse stem cells, averaged across 19 autosomal chromosomes, is 0.88 (root-mean-square deviation = 3.18). This is a remarkable level of prediction accuracy considering differing measurement approaches between simulation and experiment. Whereas Replicon calculates an expectation value for each genomic bin, empirical data are reported as a wavelet-smoothed ratios of early vs. late replication timing (Hiratani et al., 2008). Moreover, the quality of the Replicon's predictions are affected by experimental noise both from DNase DGF and replication timing experiments. These results signify the robustness of both the Replicon's algorithm and the DNA replication timing program.

DNA combing technique has been used extensively to measure DNA replication kinetics on a single molecule scale. Herrick and colleagues, for instance, used this technique in an *in vitro* system to derive some of the quantitative parameters that govern DNA replication (Herrick et al., 2002). Here, we wanted to know if relationships between cell-cycle time and growth of newly synthesized DNA, observed and formulated by Herrick *et al.* in *Xenopus laevis*, could be recovered with the simulated DNA replication cycle performed by Replicon using *Mus musculus* DNase



**FIGURE 2 | DNA replication timing profile for a 65 mega-base region on chromosome 14 of mouse embryonic cells.** Replication time (early to late; y-axis) is plotted as a function of genome position (x-axis). Predicted profile (red) was generated from DNase I DGF data; experimentally observed timing profile is depicted in blue.

I hypersensitivity sites. To that end, we captured (**Figure 3**) the changing lengths of eyes (newly replicated), holes (yet-to-be replicated) DNA regions, and the distances between Replicon bubbles (eye-to-eyes) during the S phase. Our results match closely those observed by Herrick and colleagues. This is remarkable given that our simulation is based on mammalian cells, where DNA replication is initiated at well-defined loci, while, in *X. laevis*, DNA replication initiation sites are selected at random. This results further demonstrates that Replicon can be used to model single-cell observations of DNA replication such as those often performed using DNA combing.

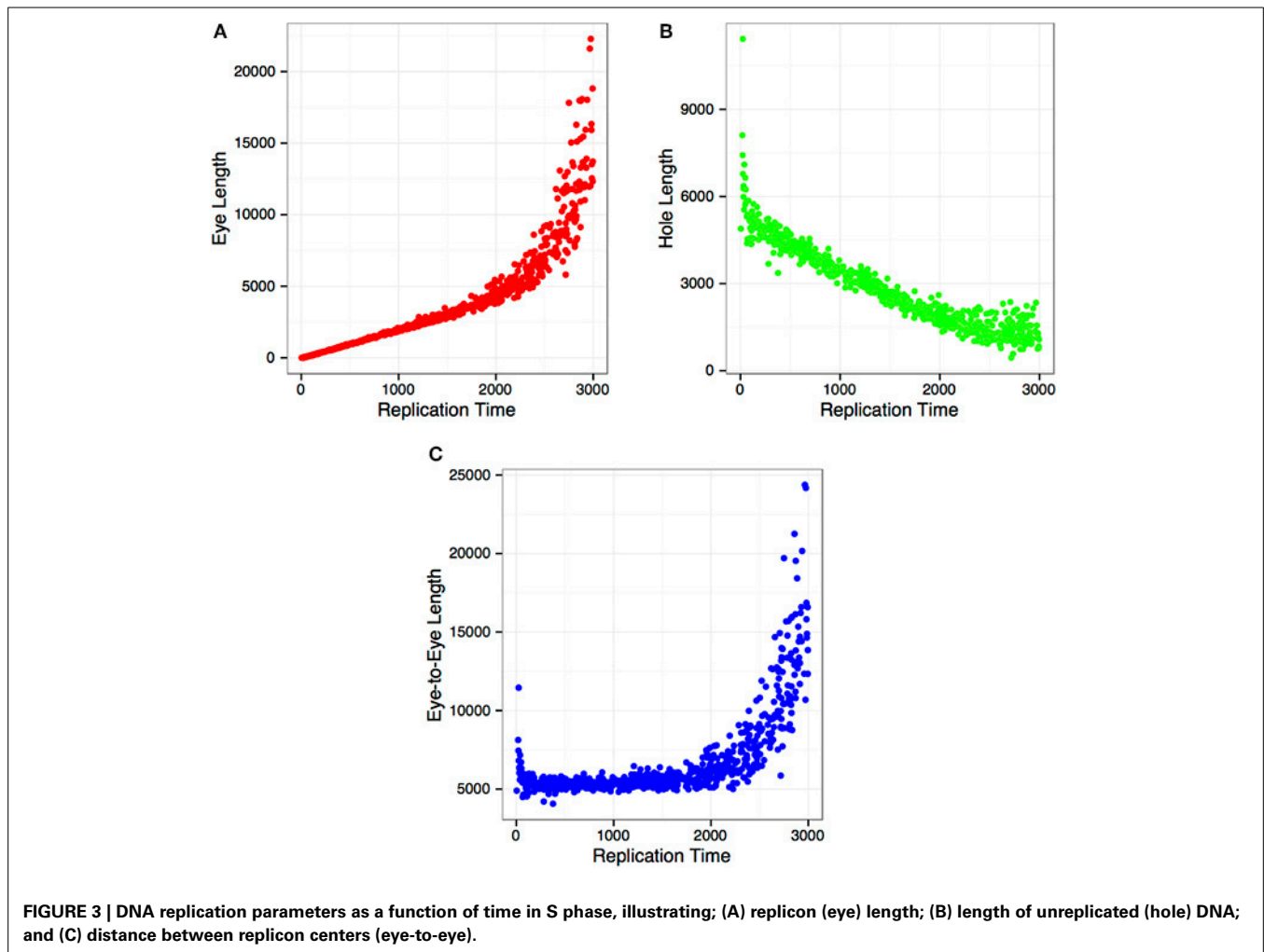
### 4. DISCUSSION

Replicon provides a way by which to accurately model DNA replication timing in metazoan cells. Here, we based our predictions on an IPLS derived from DNase I digestion data. Other IPLSs could just as easily be used to investigate, for instance, combinatorial effects of histone marks and DNA sequence motifs on DNA replication timing.

The association of replication timing with central cellular processes, such as differentiation or cytogenetic aberrations, has been revealed in recent years. But empirical replication timing data is not always available. In some instances it may not even be accessible if, for example, only formalin-fixed tissues is available. Given the close level of agreement between predicted and empirical DNA replication timing profiles, Replicon can be utilized to generate DNA replication timing predictions, provided appropriate genomic annotations are available.

### AVAILABILITY AND REQUIREMENTS

Project name: Replicon;  
 Project home page: <https://github.com/RepliconBioinfo/>;  
 Operating system: Platform independent;  
 Programming language: C++ and Java;  
 License: GNU;  
 Restrictions: None.



## AUTHOR CONTRIBUTIONS

Yevgeniy Gindin wrote the manuscript. Yevgeniy Gindin and Sven Bilke developed software and performed statistical analysis. Sven Bilke and Paul S. Meltzer conceived the study, supervised the work and edited the manuscript.

## ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

## REFERENCES

- Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., et al. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.* 19, 837–844. doi: 10.1038/nsmb.2339
- Comoglio, F., and Paro, R. (2014). Combinatorial modeling of chromatin features quantitatively predicts DNA replication timing in *Drosophila*. *PLoS Comput. Biol.* 10:e1003419. doi: 10.1371/journal.pcbi.1003419
- Dai, J., Chuang, R.-Y., and Kelly, T. J. (2005). DNA replication origins in the *Schizosaccharomyces pombe* genome. *Proc. Natl. Acad. Sci. U.S.A.* 102, 337–342. doi: 10.1073/pnas.0408811102
- ENCODE Project Consortium and others. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9:e1001046. doi: 10.1371/journal.pbio.1001046
- Gindin, Y., Valenzuela, M. S., Aladjem, M. I., Meltzer, P. S., and Bilke, S. (2014). A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Mol. Syst. Biol.* 10:722. doi: 10.1002/msb.134859
- Herrick, J., Jun, S., Bechhoefer, J., and Bensimon, A. (2002). Kinetic model of DNA replication in eukaryotic organisms. *J. Mol. Biol.* 320, 741–750. doi: 10.1016/S0022-2836(02)00522-3
- Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.-W., et al. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* 6:e245. doi: 10.1371/journal.pbio.0060245
- Hyrien, O., and Goldar, A. (2010). Mathematical modelling of eukaryotic DNA replication. *Chromosome Res.* 18, 147–161. doi: 10.1007/s10577-009-9092-4
- Martin, M. M., Ryan, M., Kim, R., Zakas, A. L., Fu, H., Lin, C. M., et al. (2011). Genome-wide depletion of replication initiation events in highly transcribed regions. *Genome Res.* 21, 1822–1832. doi: 10.1101/gr.124644.111
- Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., and Oda, M. (2010). Eukaryotic chromosome DNA replication: where, when, and how? *Ann. Rev. Biochem.* 79, 89–130. doi: 10.1146/annurev.biochem.052308.103205
- Pope, B. D., Aparicio, O. M., and Gilbert, D. M. (2013). SnapShot: replication timing. *Cell* 152, 1390–1390.e1. doi: 10.1016/j.cell.2013.02.038
- 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. doi: 10.1038/nprot.2011.328
- Sabo, P. J., Hawrylycz, M., Wallace, J. C., Humbert, R., Yu, M., Shafer, A., et al. (2004). Discovery of functional noncoding elements by digital analysis

- of chromatin structure. *Proc. Natl. Acad. Sci. U.S.A.* 101, 16837–16842. doi: 10.1073/pnas.0407387101
- Taylor, J. H. (1960). Asynchronous duplication of chromosomes in cultured cells of Chinese hamster. *J. Biophys. Biochem. Cytol.* 7, 455–464. doi: 10.1083/jcb.7.3.455
- Valenzuela, M. S., Hu, L., Lueders, J., Walker, R., and Meltzer, P. S. (2012). Broader utilization of origins of DNA replication in cancer cell lines along a 78 kb region of human chromosome 2q34. *J. Cell. Biochem.* 113, 132–140. doi: 10.1002/jcb.23336

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 06 August 2014; accepted: 13 October 2014; published online: 03 November 2014.

Citation: Gindin Y, Meltzer PS and Bilke S (2014) Replicon: a software to accurately predict DNA replication timing in metazoan cells. *Front. Genet.* 5:378. doi: 10.3389/fgene.2014.00378

This article was submitted to *Bioinformatics and Computational Biology*, a section of the journal *Frontiers in Genetics*.

Copyright © 2014 Gindin, Meltzer and Bilke. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.