NEAR-SIGMOID MODELING TO SIMULTANEOUSLY PROFILE GENOME-WIDE DNA REPLICATION TIMING AND EFFICIENCY IN SINGLE DNA REPLICATION MICROARRAY STUDIES*

JUNTAO LI [1]
MAJID ESHAGHI[2]
JIANHUA LIU[3]
R. KRISHNA MURTHY KARUTURI[4†]

[1,4]*Computational & Mathematical Biology,* [2,3]*Systems Biology, Genome Institute of Singapore, #02-01, Genome, 60 Biopolis ST, Republic of Singapore 138672.*

DNA replication is a key process in cell division cycle. It is initiated in coordinated manner in several species. To understand the DNA replication in a species one needs to measure the half replication timing (or replication timing) and the efficiency of replication which vary across genome in higher eukaryotes. In the previous studies, no direct assessment of replication efficiency on a genomic scale was performed while the replication timing was indirectly assessed using average DNA. In this paper, we present a first-ever-method of directly measuring both half replication timing and efficiency simultaneously from a single DNA microarray time-course data. We achieve it by fitting the so called *near-sigmoid model* to each locus of the DNA. We use this model apply *S. pombe* DNA replication microarray data and show that it is effective for genome-scale replication timing and efficiency profiling studies.

1 Introduction

DNA replication is a very important process in cell cycle progression, takes place within a short cell cycle phase called S-phase. It is initiated at multiple sites or origins at varying times in eukaryotic genomes [1-3] within S-phase. It was shown that [4] some regions of the genome initiate replication early, some in the middle, and the others near the end, attributing to a strict timing and coordination of firing at origins with a few exceptions such as frog embryo [24]. The replication carried out by the fork initiated (also called *replication firing*) at a locus is called *active replication* and the site is called *origin of replication* or *origin*. Whereas the replication carried by the passing forks resulting from the nearby origins is called *passive replication*. The active and passive replications are well defined in S. cerevisiae and fuzzily defined in the other higher eukaryotes when efficiency of replication is relatively low.

The two genome-wide profiles of interest in DNA replication studies are *half replication timing* (or *replication timing*) and *replication efficiency*. Half replication timing of a locus is the time it takes to complete its replication in half of the cells or the probability that it is

† Corresponding author (karuturikm@gis.a-star.edu.sg)
* This work was supported by Genome Institute of Singapore and the Agency for Science, Technology and Research, Singapore.

replicated is 0.5. This is especially significant if the replication efficiency is less than 100% i.e. replication of the locus takes significant time of the S-phase. The importance of the half replication timing is its estimation stability to identify the origins of replication. Genome-wide DNA microarray analyses have been widely used to determine profiles of half replication timing at the genomic scale [4-7, 14]. It is measured indirectly by average DNA content at the loci and the peaks in the average DNA content profile show the origins or most likely active replication sites.

Flexible timing of firing or inefficient firing at origins was observed in several species such as human [7], *S. pombe* [8,9] and frog embryos [24] unlike in *S. cerevisiae* [4]. Replication efficiency is the measure of strictness of timing of replication of the locus under consideration i.e. 100% efficient locus is the one that is always replicated strictly at the same time in all cell cycles whereas an inefficient locus is replicated at different timings in different cell cycles within a given period within S-phase.

The efficiency of replication has been measured using different techniques such as Single strand DNA (ssDNA) [17] and DNA combing [9]. DNA combing technique analyzes DNA replication/firing only at single origins leaving it to be a tedious and low throughput technique to measure efficiency and half replication time. Hence it cannot be used for genome scale study. In case of ssDNA technique, the amount of nascent DNA accumulated at sites of replication initiation during HU treatment may not be proportional to firing efficiency on a genomic scale resulting in inaccurate assessment of replication efficiency. Moreover, the current approach requires two different technologies and datasets, one to measure half replication timing and the other to measure efficiency while not resulting in any advantage.

Though genome-wide microarray analyses have been widely used to determine profiles of half replication timing at the genomic scale, direct estimation of replication efficiency at various loci of the genome based on the genome-wide replication profiles has not been performed previously. In this paper, we demonstrate that replication efficiency, together with half replication timing, can be estimated using a novel approach - the *near-sigmoid modeling* for the increase in DNA copy number as a function of time at individual loci. The *near-sigmoid modeling* approach permits estimation of replication start timing and replication end timing at various loci of the genome. Based on these measurements, we attain the genome-wide profiles of half replication timing and replication efficiency. The rest of the paper describes near-sigmoid modeling and proceeds to show its efficacy on genome-wide profiling of DNA replication timing and efficiency of S. Pombe.

2 Near-Sigmoid Modeling

In our approach, DNA replication process is described in three steps: *initiation*, *linear progression*, and *completion* of replication. The time period from the replication initiation

to the completion is defined as the duplication time *DT*. As one and only one copy of DNA at all loci would be synthesized during the S-phase, $\eta_r \cdot DT = 1$, where η_r is the (average) replication efficiency = *1/DT* i.e. rate of fraction of cells replicate the given locus in a given time upon initiation of replication, higher the *DT* lower the efficiency.

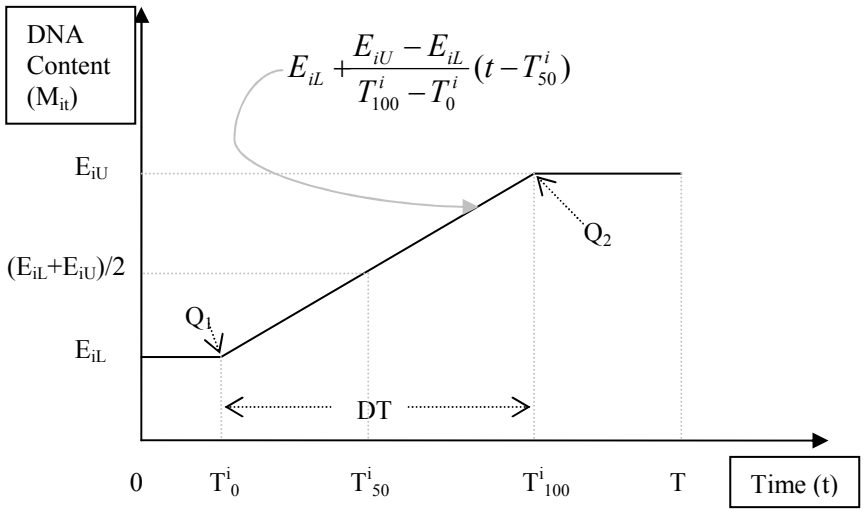

Figure 1. Graphical illustration of near-sigmoid model to measure half replication timing and efficiency. Two point of inflexion Q_1 and Q_2 define the model, a piecewise linear approximation of sigmoid, hence the name.

Near-sigmoid model represents the above replication model as a specialized 3-piecewise linear model as shown in figure 1. This model is called near-sigmoid model because it is a piecewise linear approximation of sigmoid model. The two points of inflexion $Q_1 = (T^i_0, E_{iL})$ and $Q_2 = (T^i_{100}, E_{iU})$ signify the quantitative state of replication of locus l_i just prior to the replication initiation and just after the replication completion respectively. T^i_0 and T^i_{100} indicate replication initiation and replication completion timings of l_i respectively; T^i_{50} is its half replication timing, the average of T^i_0 and T^i_{100} i.e. $T^i_{50} = (T^i_0+T^i_{100})/2$. T is the end of the experiment. Duplication time (DT^i), inverse of replication efficiency (η^i_r), is $T^i_{100} - T^i_0$. E_{iL} is the initial DNA content at l_i which remains constant till T^i_0 and starts increasing linearly at T^i_0 till it reaches E_{iU} at time T^i_{100}. T^i_0 and T^i_{100} obey the conditions $0 \le T^i_0 < T^i_{100} \le T$ and $1 \le E_{iL} \le E_{iU} \le 2$. E_{iL} and E_{iU} are ideally expected to be 1 and 2 respectively. We used them in this manner to signify the fact that some loci may have already replicated to varying extent before the release of the arrested cells to progress in which case $E_{iL} > 1$ and the experiment may have stopped before the end of S-phase i.e. $T <$ S-phase period resulting in $E_{iU} < 2$ for some late replicating or highly inefficient loci. The model is mathematically expressed as follows.

Let C_{it} be the relative DNA content in the synchronized S-phase cells with that of the reference genome at locus l_i at time t. Let C_{itk} be the k^{th} repeated measurement of C_{it}. Let

M_{it} be the estimation of $log(C_{it})$ as described by the near-sigmoid model in equation (3). In this model we assume that, like in typical microarray studies, $log(C_{itk}) \sim N(log(C_{it}), \sigma_i^2)$.

$$M_{it} = \begin{cases} E_{iL} & for \quad t \in [0, T_0^i] \\ E_{iU} & for \quad t \in [T_{100}^i, T] \\ E_{iL} + \dfrac{E_{iU} - E_{iL}}{T_{100}^i - T_0^i}(t - T_0^i) & for \quad t \in [T_0^i, T_{100}^i] \end{cases} \tag{3}$$

$$E_{iL} = \frac{1}{K_L^i} \sum_{t=0}^{T_0^i} \sum_{k=1}^{n_t^i} \log(C_{itk}) \qquad and \qquad E_{iU} = \frac{1}{K_U^i} \sum_{t=T_{100}^i}^{T} \sum_{k=1}^{n_t^i} \log(C_{itk})$$

$$K_L^i = \sum_{t=0}^{T_0^i} n_t^i \qquad and \qquad K_U^i = \sum_{t=T_{100}^i}^{T} n_t^i$$

Where n_t^i is the number of repeats for l_i at time t, the superscript 'i' signifies the fact that some observations may be missing and their numbers vary from locus to locus at each time point. K^i_L is the number of measurements up to T^i_0 and K^i_U is the number of measurements from time T^i_{100} to T, the end of the time-course.

2.1 Near-Sigmoid Model Parameter Estimation

As can be seen from the definitions in equation (3), we need to estimate optimal values for T^i_0 and T^i_{100} given $\{C_{itk}\}_i$ which automatically leads to the remaining parameters of the model. They are estimated by exhaustive search by minimizing the mean-squared error between M_{it} and $log(C_{itk})$ as depicted in equation (4).

$$O = \frac{1}{K^i} \sum_{t=0}^{T} \sum_{k=1}^{n_t^i} \left(M_{it} - \log(C_{itk})\right)^2$$
$$K^i = \sum_{t=0}^{T} n_t^i \tag{4}$$

Where K^i is the total number of measurements made for l_i at all time points put together.

2.2 Statistical Significance of Near-Sigmoid Fit

Upon estimation of the model parameters (Q_1 & Q_2), we have to examine whether the near-sigmoid fit is better than constant or flat fit which we evaluate using hypothesis testing and false discovery rate estimation. The statistical significance (or the p-value) of

the fit was calculated under the null hypothesis that $E_{iU} = E_{iL} = E_i$ and the alternative hypothesis is $E_{iU} > E_{iL}$. The following ANOVA table is used to derive the statistic F^i.

Table 1. ANOVA table for near sigmoid model fit. The F^i statistic follows central F distribution with degrees of freedom *3* and *(K^i-4) i.e.* $F_{3,K^i-4}(x)$. Higher the F^i better the fit.

	Sum of squares, *SS*	Degrees of freedom, *df*	F-statistic, F^i
Regression	$SSR^i = \sum_{t=0}^{T} (M_{it} - \overline{\log(C_i)})^2$ $\overline{\log(C_i)} = \frac{1}{K^i} \sum_{t=0,k=1}^{t=T,k=n_t^i} \log(C_{itk})$	3	$F^i = \frac{\frac{SSR^i}{3}}{\frac{SSE^i}{(K^i - 4)}}$
Error	$SSE^i = \sum_{t=0,k=1}^{t=T,k=n_t^i} (\log(C_{itk}) - M_{it})^2$	K^i-4	
Total	$TSS^i = \sum_{t=0,k=1}^{t=T,k=n_t^i} (\log(C_{itk}) - \overline{\log(C_i)})^2$	K^i-1	

P-value of the fit (p_i) is given by area under F_{3,K^i-4} from F^i to ∞. The p-values of all loci were used to obtain false discovery rate (FDR) with monotonicity correction [23]. The loci above an FDR cut-off are declared to be unfit or flat responsive loci.

2.3 Meaning of Flat Responsive or Unfit Loci

In principle, each and every locus of the DNA has to be replicated before the cell cycle progresses into its next stage i.e. G2 phase. The insignificant FDR (p-value) shows that the DNA content at the locus has not changed from the start of the study to the end which means either the probe is bad or the locus has replicated even before the release of the cells from synchronization block such as HU arrest to progress in the S-phase. We believe that almost all probes in an array were tested for their goodness, leaving only possibility that the loci have replicated even before the release of the cells to progress. Hence the flat responsive or unfit loci (probes) signify that the corresponding loci are early efficient replication regions. We show in the next section (3) that it is indeed the case in *S. pombe*.

3 S. pombe DNA replication Data Analysis using Near-Sigmoid Modeling

To investigate the efficacy of our approach we present its application on genome-wide DNA replication timing and efficiency. Appropriately normalized DNA microarray data on DNA replication was obtained from [15]. It is based on *S. pombe* genome-wide ORF-specific microarray and the increase in DNA copy numbers at individual loci (from all

three chromosomes) was studied in cells released after HU block. The microarray has an average resolution of one locus per ~2.4 Kb. Each locus (or ORF) was represented by two different 50-mer oligonucleotide probes whose average ratio was used for profiling. The length and resolution of the time-course are 60min and 5 min respectively, two repeats are available at each time point

We applied the near-sigmoid model to fit DNA copy number increase as a function of time at individual loci for estimation of replication initiation timing T_0 and completion timing T_{100}. To this end, the T_0 and T_{100} at the majority of loci (> 96% loci of the genome) were attained based on the criterion of FDR less than 0.01%.

Our methodological approach is validated by the following genome-wide observations: (1) predicted replication origins are close to A+T islands and the previously predicted origins; (2) the unfit or flat responsive loci are close to A+T islands and other previously predicted origins; (3) chrIII has early half replication timing and lower efficiency relative to the remaining two chromosomes; (4) telomeres on chrI and chrII are late replicating and more efficient. The importance of A+T islands in our observations stems from the fact that the origins of replication in *S. pombe* were shown to be close to A+T islands [21]. The detailed results are described in the following subsections 3.1 through 3.5.

3.1 Replication origins are close to A+T islands and other predicted origins

The origins of replication were predicted using Peak finder software [20] on T_{50} profile. Of the 516 origins predicted 285 overlap with A+T islands [21 (A+T)] and 360, 48, 305, 295, 239 and 318 match with the peaks predicted by [17, 21(Validated), 22 (ORC1), 22 (MCM6), 16 (Wt), 16 (ΔCDS1)] respectively.

3.2 Flat responsive loci coincide with A+T islands and predicted origins

193 loci were flat responsive which were analyzed to check whether they match closely to A+T islands and the predicted origins of replication. The islands are expected to be close to the origins of replication, especially the early ones. Of the 193 flat responsive loci, 146, 173, 36, 155, 154, 129 and 147 loci match with A+T islands [21 (A+T)] and the other predicted origins [17, 21(Validated), 22 (ORC1), 22 (MCM6), 16 (Wt), 16 (ΔCDS1)] respectively. This shows that these loci are indeed close to the A+T islands and predicted origins of replication which proves the efficacy of our approach and interpretation.

3.3 Early replication timing of chrIII relative to chrI & chrII

Further evaluation of the effectiveness of our approach to measure half replication timing was further carried out by comparing the half replication timing of chrIII with that of chrI & chrII. ChrIII was observed to be early half replicating as compared to that of chrI & chrII [17]. This was reinforced in our analysis with P-value $< 2.2 \times 10^{-16}$. The box plots of the half replication timings of chrI & chrII put together and chrIII are shown in figure 2.

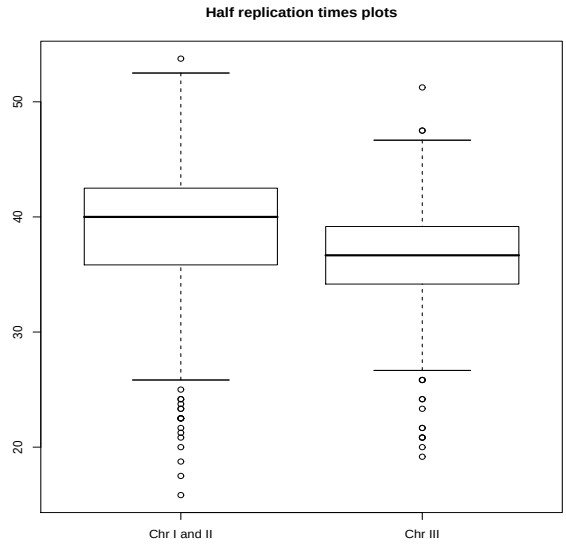

Figure 2. Boxplots of half replication timing (vertical axis) of loci on chrI & chrII and chrIII (horizontal axis). chrIII is systematically has early half replication timing compared to the remaining chromosomes, p-value < $2.2x10^{-16}$ by Wilcoxon rank-sum test.

3.4 Low efficiency of replication of chrIII relative to chrI & chrII

The evaluation of the effectiveness of our model to measure efficiency of replication is carried out by comparing the overall efficiency of chrI & chrII with that of chrIII. ChrIII was observed to be less efficient compared to that of chrI &chrII, with P-value < $2.2x10^{-16}$ and the box plots of efficiency are shown in figure 3. This observation was supported by the fact that chrIII was shown to have lot more origins [22] with earlier half replication timing compared to the remaining two chromosomes. This is due to the fact that inefficient origins are expected to have systematically early half replication timing in order to complete replication and there should be many such origins owing to their inefficiency.

3.5 Higher efficiency of telomeres compared to the other regions

The evaluation of the effectiveness of our model to measure efficiency of replication is carried out by comparing telomere regions with that of the others in chrI & chrII. Telomeres are defined as the regions in chrI for < 0.2Mb and >5.4Mb, for chrII < 0.2Mb and > 4.4Mb. We observed that the telomeres are more efficient (p-value is 0.06) as telomeres are known to be late replicating which is possible only if they are more efficient than the other regions. The box plots of efficiency of all loci in telomeres and the other regions are shown in figure 4.

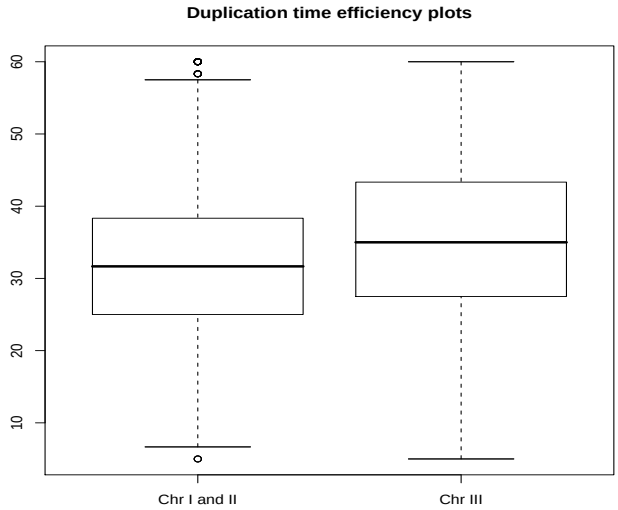

Figure 3. Boxplots of duplication times (vertical axis) of loci on chrI & chrII and chrIII (horizontal axis). chrIII is systematically has higher duplication time or lower efficiency compared to the remaining chromosomes, p-value < 2.2×10^{-16} by Wilcoxon rank-sum test.

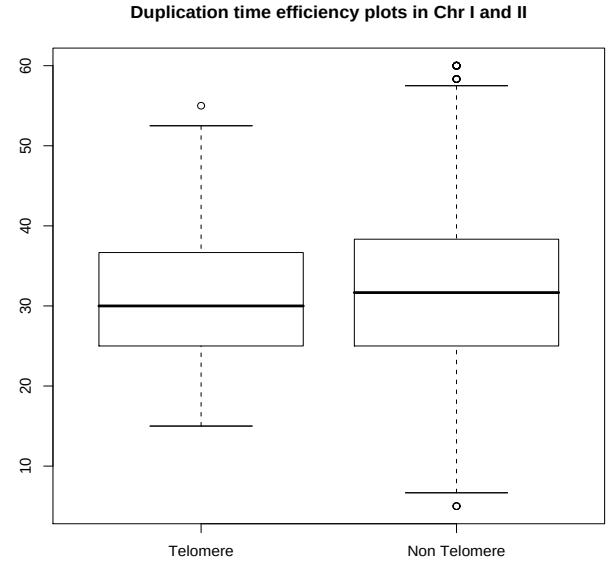

Figure 4. Boxplots of duplication times (vertical axis) of loci on telomeres and non-telomeres (horizontal axis) of chrI & chrII. Telomeres, systematically have lower duplication time or higher efficiency compared to the non-telomere regions, p-value < 0.06 by Wilcoxon rank-sum test.

4 Discussion

We have presented the first-of-its-kind approach that permits direct assessment of genome-wide replication efficiency at individual loci from the same dataset used to

estimate the half replication timing. This is a significant step in DNA replication studies since the direct and accurate genomic scale estimation of replication efficiency at various loci of the genome based on the genome-wide replication profiles which has not been performed previously though the genome-wide microarray analyses have been widely used to determine profiles of average DNA replication timing at the genomic scale. In this paper, we demonstrated that replication efficiency, together with average replication timing, can be estimated using a novel approach - the near-sigmoid fitting for the increase in DNA copy number as a function of time at individual loci. Based on their estimates at each locus, we attained the genome-wide profiles of half replication timing and replication efficiency. We have shown the efficacy of our approach by various observations and their concordance with the literature on the analysis of *S. pombe* DNA replication microarray data.

Timing of firing at origins is proximated by the average replication timing of the peak loci. This approach has limitations on identifying (inefficient) late-firing origins closely located to the other (efficient) early-firing origins [2,4]. As firing at origins is relatively inefficient in *S. pombe* [8,9,16], it would not only under-estimate the number of efficient late firing origins, but also would fail to identify most, if not all, inefficient late-firing origins. This is because inefficient late-firing origins are unlikely to be self-sufficient in replication of the origin DNA. Nevertheless, this is still the most effective way to predict origins of replication at the genomic scale [14].

Acknowledgments

We thank Edison T. Liu and Neil D. Clarke for their support during this work.

References

1. M. Barranco and J. R. Buchler. Thermodynamic properties of hot nucleonic matter. *Phys. Rev.,* C22:1729—1737, 1980.
2. H. Müller and B. D. Serot. Phase transition in warm, asymmetric nuclear matter. *Phys. Rev.,* C52:2072—2091, 1995.
3. V. Baran, M. Colonna, M. Di Toro and A. B. Larionov. Spinodal decomposition of low-density asymmetric nuclesr matter. *Nucl. Phys.,* A632:287—303, 1998.
4. V. Baran, M. Colonna, M. Di Toro and V. Greco. Nuclear fragmentation: Sampling the instability of binary systems. *Phys. Rev. Lett.,* 86:4492—4495, 2001.
5. Gilbert DM (2001) Nuclear position leaves its mark on replication timing. J Cell Biol 152: F11-15.
6. MacAlpine DM, Bell SP (2005) A genomic view of eukaryotic DNA replication. Chromosome Res 13: 309-326.
7. Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. Annu Rev Biochem 71: 333-374.
8. Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, et al. (2001) Replication dynamics of the yeast genome. Science 294: 115-121.

9. Schubeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, et al. (2002) Genomewide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. Nat Genet 32: 438-442.
10. Yabuki N, Terashima H, Kitada K (2002) Mapping of early firing origins on a replication profile of budding yeast. Genes Cells 7: 781-789.
11. Jeon Y, Bekiranov S, Karnani N, Kapranov P, Ghosh S, et al. (2005) Temporal profile of replication of human chromosomes. Proc Natl Acad Sci U S A 102: 6419-6424.
12. Kim SM, Huberman JA (2001) Regulation of replication timing in fission yeast. Embo J 20: 6115-6126.
13. Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N (2006) DNA replication origins fire stochastically in fission yeast. Mol Biol Cell 17: 308-316.
14. Eklund H, Uhlin U, Farnegardh M, Logan DT, Nordlund P (2001) Structure and function of the radical enzyme ribonucleotide reductase. Prog Biophys Mol Biol 77: 177-268.
15. Majid Eshaghi, R. Krishna M. Karuturi, Juntao Li, Zhaoqing Chu, Edison T. Liu and Jianhua Liu. Global Profiling of DNA Replication Timing and Efficiency Reveals that Efficient Replication/ Firing Occurs Late During S-phase in Cells Released after HU-block in S. pombe, PLoS One, 2(8): e722. doi:10.1371/journal.pone.0000722.
16. Feng W, Collingwood D, Boeck ME, Fox LA, Alvino GM, et al. (2006) Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. Nat Cell Biol 8: 148-155.
17. Heichinger C, Penkett CJ, Bahler J, Nurse P (2006) Genome-wide characterization of fission yeast DNA replication origins. Embo J 25: 5171-5179.
18. MacNeill SA, Fantes PA (1997) Genetic and physiological analysis of DNA replication in fission yeast. Methods Enzymol 283: 440-459.
19. Rhind N (2006) DNA replication timing: random thoughts about origin firing. Nat Cell Biol 8: 1313-1316.
20. Glynn EF, Megee PC, Yu HG, Mistrot C, Unal E, et al. (2004) Genome-wide mapping of the cohesin complex in the yeast S. cerevisiae. PLoS Biol 2: E259.
21. Segurado M, de Luis A, Antequera F (2003) Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. EMBO Rep 4:1048-1053.
22. Hayashi M, Katou Y, Itoh T, Tazumi M, Yamada Y, et al. (2007) Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. Embo J 26: 1327-1339.
23. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc.B *57*, 289–300.
24. Herrick J, Jun S, Bechhoefer J, Bensimon A (2002). Kinetic model of DNA replication in eukaryotic organisms. J Mol Biol. 320(4):741-50.