bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SOFTWARE

Repliscan: a tool for classifying replication timing regions

Gregory J Zynda[1], Jawon Song[1], Lorenzo Concia[2], Emily E Wear[2], Linda Hanley-Bowdoin[2], William F Thompson[2] and Matthew W Vaughn[1*]

*Correspondence: vaughn@tacc.utexas.edu
[1]Texas Advanced Computing Center, University of Texas at Austin, 10100 Burnet Road, 78758-4497 Austin, TX, USA
Full list of author information is available at the end of the article

Abstract

Background: Replication timing experiments that use label incorporation and high throughput sequencing produce peaked data similar to ChIP-Seq experiments. However, the differences in experimental design, coverage density, and possible results make traditional ChIP-Seq analysis methods inappropriate for use with replicating timing.

Results: To accurately detect and classify regions of replication across the genome, we present Repliscan. Repliscan robustly normalizes, automatically removes outlying and uninformative data points, and classifies Repli-seq signals into discrete combinations of replication signatures. The quality control steps and self-fitting methods makes Repliscan generally applicable and superior to previous methods with thresholds inapplicable to different genomes.

Conclusions: Repliscan is simple and effective to use on organisms with different magnitude genome sizes and sequencing coverage as low as 2.4x.

Keywords: DNA replication; Repli-seq; Classification

Background

The most essential property of the cell is its ability to accurately duplicate its DNA and divide to produce two daughter cells [1]. The cell's replication cycle starts with G_1 phase, in which molecules essential for cell division are produced, then proceeds to replicating DNA in S phase. After all DNA in the genome is duplicated, the cell continues to grow in G_2 phase until it divides into two daughter cells at the end of Mitosis, or M phase, at which point it is ready to start the cell cycle again (Figure 1).

To ensure accuracy and efficiency, S phase is complex and highly regulated. Instead of duplicating in a single zipping motion, reminiscent of transcription, DNA is synthesized in regions at distinct times in eukaryotes, initiating at multiple origins of replication [2]. This synthesis process is taking place in a live cell, so replication mechanisms need to be coordinated with active transcription and chromatin configuration. For example, early replication correlates with chromatin accessibility in a cell [3].

To better understand the coordinated program of DNA replication, two types of protocols have been developed to examine genome-wide DNA replication during S phase: one based on the time of replication, TimEx [4, 5], and the other based on incorporation of a labeled precursor into newly replicated DNA, Repli-seq/Repli-chip [6, 7, 8, 9, 10]. Time of replication (TimEx) measures DNA coverage at sequential

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

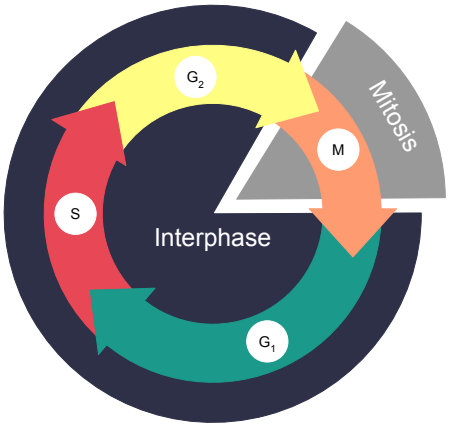

Figure 1 Overview of the cell cycle. Cell division takes place in two stages: interphase and mitosis. Interphase is when a cell copies its genome in preparation to physically divide during mitosis. Interphase starts with cell growth and preparation for DNA synthesis in Gap (G_1). After G_1, DNA is replicated in regions during the Synthesis (S) phase. The cell then transitions into a second growth phase - Gap 2 (G_2). When the cell has finished growing, the cell divides into two daughter cells in Mitosis (M).

times in S-phase. The normalized early S-phase signal should be mostly 1x coverage, additively transitioning to 2x coverage in late S-phase. In contrast to this method, Repli-seq/Repli-chip works by only sequencing newly replicated DNA. Theoretically, in a single cell, this means once a region is replicated, it should not appear in samples taken at later times, except in the case of allelic timing differences. Both methods have been shown to yield similar results [11, 12] for when and where genomic regions replicate, but each requires a distinct type of analysis. The methods described in this paper focuses on data produced by label incorporation (Repli-seq).

Data Description

In continuation to our analysis of *A. thaliana* chromosome 4 from 2010 [13], we updated our laboratory protocol to be more stringent and return higher-resolution data as described in Bass *et al.* 2014[14], Bass *et al.* 2015[15], and Wear *et al.* 2016[16]. We increased the sensitivity of the labelling process by using 5-Ethynyl-2'-deoxyuridine (EdU), which does not require harsh denaturation of DNA, unlike 5-Bromo-2'-deoxyuridine (BrdU) which was used in previous work. A flow cytometer is then used to separate labeled from unlabeled nuclei, and to separate labeled nuclei into different stages of S phase based on their DNA content. Next, DNA is extracted from sorted nuclei: the newly replicated DNA is immunoprecipitated and then sequenced using an Illumina sequencer. Previous protocols used microarrays for labeled DNA detection, which provided signal on probes at fixed intervals across a genome. Directly sequencing the immunoprecipitated DNA allows for single-base replication timing resolution for any organism with a reliably-assembled reference genome.

Following this Repli-seq protocol, we created an exemplar A. thaliana dataset for development, with sequenced nuclei from: G_1 (non-replicating control), early, middle, and late S phase. While the amplification, fragmentation, and sequencing

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of next generation sequencing (NGS) libraries should be unbiased and random, physical factors affect the sequenceability of each region. To correct for these effects, we use the raw non-replicating DNA from the G_1 control to normalize any sequenceability trends.

Introducing Repliscan

In addition to our updated laboratory protocol for generically measuring DNA replication, we needed to improve the sensitivity and robustness of our analytical method. In previous work, log-ratios and aggressive smoothing were used to classify genomic regions by their time of replication. While this yielded results with high true-positive rates, we found that this approach over-smoothed our high resolution sequencing data. We created the Repliscan method to analyze generic, DNA sequence-based replication timing data. Accepting any number of S-phase timepoints as input, Repliscan, removes uninformative or outlying data, smooths replication peaks, and then classifies regions of the genome by replication time.

Methods

The analysis of the replication time data starts like any other DNA sequencing analyses, with quality control, alignment to reference, and filtering. Quality control consisted of removing contaminating 3' universal sequencing adapters from the paired reads, and trimming the 5' ends with quality scores below 20 with the program Trim Galore![17] version 0.3.7, which is designed to maintain read pairs. While it is obvious that low-quality regions need to be removed or masked because those base calls are untrustworthy, any contaminating sequences from adapters hinder the alignment process even more because they are always high-quality and may comprise a large part of the read. Therefore, any reads in the output from Trim Galore! shorter than 40 base pairs were discarded, and any resulting singletons (unpaired reads) were not included for alignment.

We then used bwa-mem [18] version 0.7.12 with default parameters to align the quality-filtered reads to the TAIR10 *A. thaliana* reference genome [19]. After alignment, we filtered out any reads with multiple alignments using samtools[20] version 1.3. Removing these non-uniquely aligning reads is essential because they come from repetitive elements or other duplications in the genome that could replicate at different times, thereby confounding region classification into discrete replication times. After our stringent alignment requirements, fewer than 0.5% of our reads were identified as duplicates by samtools. We decided that removing the duplicates was unnecessary for our data due to the depth of our sequencing and localized nature of segmentation spikes.

Windowing

The DNA sequencing workflow leaves us with raw replication signals across a genome, which we must classify into distinct genomic regions and assign replication times. Our methods for this process build on methods from Lee *et al.*[13] and are illustrated in Figure 2.

At first glance, Repli-seq data appears similar to dense ChIP-seq [21], when viewed in a genome browser (Figure 3). However, instead of highlighting a limited number of coverage peaks as sites of molecular interactions, replication timing data

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

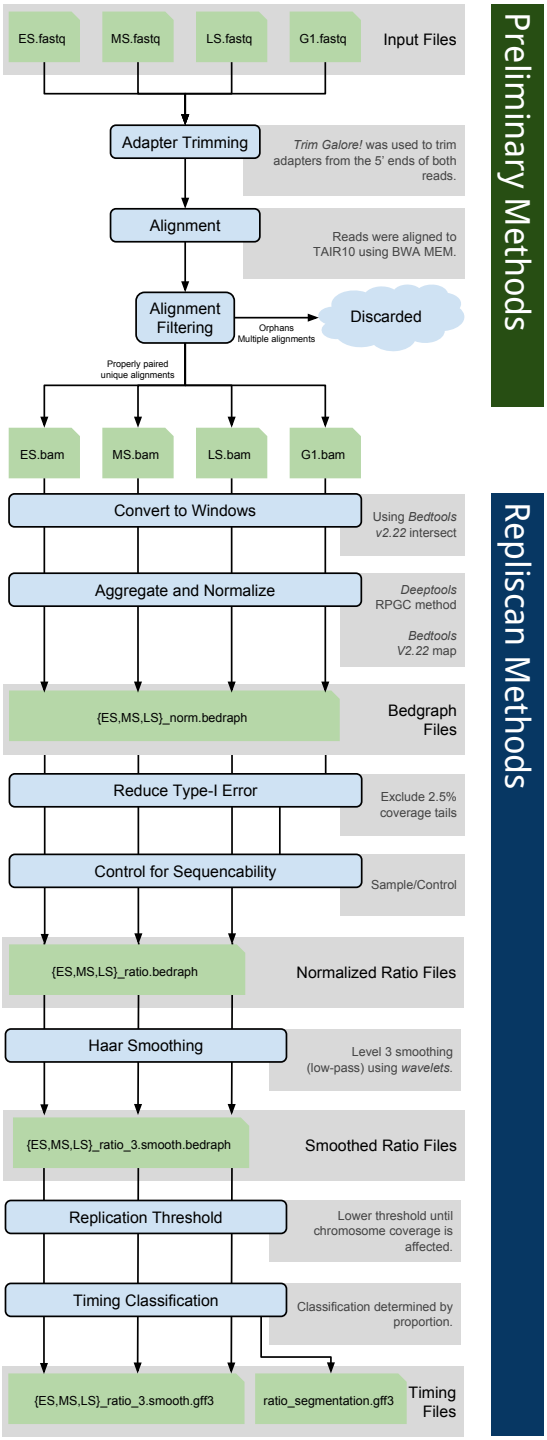

Figure 2 Repliscan workflow. Diagram of the preliminary alignment and quality control methods in green at the top, and the Repliscan methods in blue at the bottom.

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

consists of coverage across the entire genome accented with extremely wide peaks corresponding to regions of replication initiation and subsequent spreading. This background coverage with subtle, broad increases in depth makes deep coverage essential to reduce sampling error when detecting statistically-relevant differences. Even though the cost of sequencing has plummeted since 2007, deep-coverage DNA sequencing is still expensive for higher order eukaryotes.

Lee *et al.* defined putative replicons in *A. thaliana* and calculated the median length to be 107 kilobases [13]. To achieve greater signal depth in each replication timing sample, we transformed each BAM alignment file into 1 kilobase coverage windows using bedtools [22]. While this transformation slightly reduces the resolution of our analysis, Figure 3 shows that the proportion of sampling error to measured signal is greatly reduced with the increased coverage. The windows also put all changes in coverage on the same coordinate system, simplifying comparisons between samples and experiments.

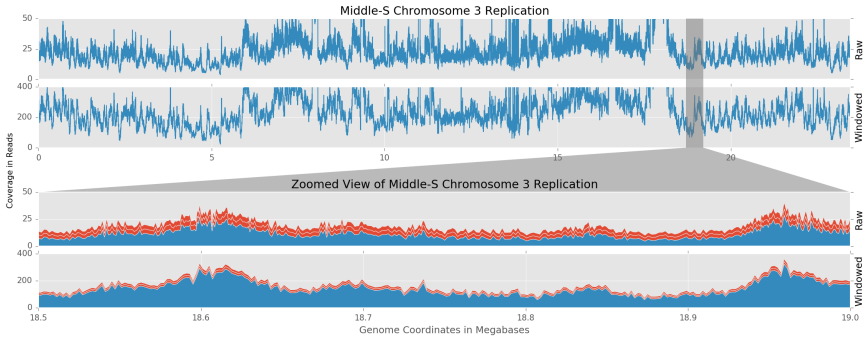

Figure 3 Replication signal and sampling uncertainty. The top two graphs show raw and windowed replication signal across *A. thaliana* chromosome 3. The bottom two graphs show raw and windowed replications signals at 18.5-19.0 megabases from the top view as represented by the gray selection area. The red bars represent sampling uncertainty ($\sqrt{\lambda}$ for Poisson distributions).

We chose 1 kilobase windows because they not only reduce sampling error, but are also two orders of magnitude smaller than the expected *A. thaliana* replicons. This is important to detect fine-grained features, such as early-S replication spreading into middle and late-S. We suggest that, when adapting Repliscan to other species, the expected replicon size be factored into calculations that establish window size and sequencing depth.

Replicate Aggregation and Normalization

To further decrease sampling effects, and achieve consistent results between experiments, we used multiple biological replicates and adopted aggregation methods to either increase coverage or summarize replication signals using functions provided by "bedtools map" [22]. For experiments with low coverage, we pooled timing $t = 1..T$ replicates $r = 1..R$ together by summing coverage signal k across each window $i = 1..N$.

$$k_{it} = \sum_{r=1}^{R} k_{it_r}$$

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

When coverage was sufficient, we used the signal mean (or the more robust signal median) to clean up aberrant coverage. For these methods, replicates were first normalized for sequencing depth using sequence depth scaling [23]. This normalization step removed differences in sequencing depth between replicates by scaling each sample to an average depth of 1x.

$$k_{it} = \text{median}\left(\frac{N * k_{it_r}}{\sum_{i=1}^{N} k_{it_r}}\right)$$

After aggregation, the combined signals were normalized once more to scale any imbalances in replicates back to 1x, prior to making any comparisons between replication times.

$$k_{it} = \frac{N * k_{it}}{\sum_{i=1}^{N} k_{it}}$$

Our *A. thaliana* test data was relatively high coverage at 30x, so we used the median function to generate a robust signal, instead of using the default sum.

Reducing Type I Error

Repliscan aims to detect and highlight peaks of replication coverage, but some peaks may be too high and may in fact be false-positives caused by errors in the reference. For instance, if a repetitive element is present three times in the actual genome, but present only once in the reference sequence due to assembly error, all reads would align uniquely to the same location. If two of those elements replicate early and the third in middle S phase, the early peak would be twice as large and dominate the classification process. To reduce type I error arising from genomic repeats, we needed to detect and exclude these areas from the final classification because there is no way to resolve such duplications without improving the reference genome.

The distribution of sequencing coverage is bounded on the left at zero, with very long, positive tails. Before we can detect any outliers we first need to transform the data to actually fit a probability distribution. We found that both the square root and log transforms stabilized the spread and skew by shortening the positive tail and lengthening the [0,1) tail. Outliers were detected in the transformed data using four different methods:

fitting a gamma distribution to the log transformed data,

$$\log(K_t) \sim \Gamma(\alpha_t, \beta_t) \equiv \text{Gamma}(\alpha_t, \beta_t)$$

fitting a gamma distribution to the square root transformed data,

$$\sqrt{K_t} \sim \Gamma(\alpha_t, \beta_t) \equiv \text{Gamma}(\alpha_t, \beta_t)$$

fitting a normal distribution to the log transformed data,

$$\log(K_t) \sim \mathcal{N}(\mu, \sigma^2) \equiv \text{Normal}(\mu, \sigma^2)$$

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

or calculating the whisker bounds (WB) of a boxplot from the log transformed data

$$X_t = log(K_t)$$
$$\mathrm{IQR}(X_t) = P_{75}(X_t) - P_{25}(X_t)$$
$$\mathrm{WB}(X_t) = [P_{25}(X_t) - 1.5 * \mathrm{IQR}(X_t), \quad P_{75}(X_t) + 1.5 * \mathrm{IQR}(X_t)]\,,$$

where P is the percentile function.

We use scipy[24] version 0.15.0 to fit all probability distributions to the actual coverage windows. Any windows with coverage in the upper and lower 2.5% tails of the calculated probability distributions, or outliers when using whiskers, are considered unrepresentative and removed (Figure 4).

$$\log(k_{it}) = \begin{cases} 0 & P_{97.5}(\alpha_t, \beta_t) < X < P_{2.5}(\alpha_t, \beta_t) \\ k_{it} & \text{Otherwise} \end{cases}$$

The large spikes of coverage that this method removes may comprise a significant amount of coverage, so we perform another round of normalization to return the sample to 1x coverage. Each of the four methods has their own strengths and computation complexity. Most coverage data can be accurately modeled with the normal distribution. For cases when the transformed coverage distributions are still skewed, we suggest using the gamma distributions. If for some reason, the coverage data is multimodal, the whisker method will sufficiently summarize the data without overfitting.

Normalize for Sequenceability

Amplification, fragmentation, and shotgun sequencing DNA is a non-uniform random process. Coupled with imperfect alignment efficiency from repetitive regions and incomplete reference genomes, artificial peaks arising from differences in the efficiency with which specific genomic regions can be sequenced are easy to confuse with actual signal peaks. This does not have a significant impact on comparisons between samples, but makes it difficult to compare adjacent genomic regions. Our sequencing protocol included a sample of non-replicating G_1 DNA to correct for this phenomenon.

In G_1, the cell is growing in physical size but no DNA replication is taking place, so the copy number of each sequence in the genome is at the 2C level. Variations in sequenceability can thus be separated from variations in signal attributable to DNA replication. Dividing each of the S-phase samples by the G_1 sample normalizes each of the windows by giving the ratio of treatment coverage over expected coverage.

$$r_t = \frac{k_t}{k_1}, \text{ where } k_1 \text{ is the control.}$$

To better illustrate this process, consider two replication coverage windows next to each other: the first one is easy to sequence, and therefore produces more fragments per unit input DNA than the second window, which is hard to sequence. The normalization step would lower the signal from the first window, dividing it by a big

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

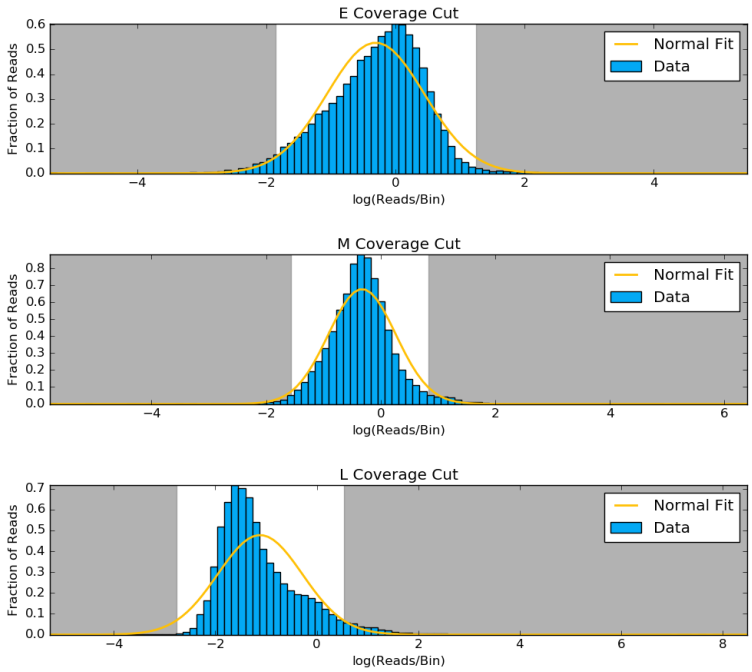

Figure 4 Outlying coverage in chromosome 3. Based on the normal distribution fit (yellow) to the log transformed coverage distribution of early (E), middle (M), and late (L) S-phase data, windows that fall in the tails shaded in gray are removed from the analysis.

coverage number from G_1. It would also raise the signal from the second window, which would be divided by a smaller G_1 number, making the two windows more comparable and reducing background noise. We recommend that such a control be implemented in all DNA sequencing based experiments to detect replication timing.

Haar Wavelet Smoothing

Data sampling is always affected by noise. Noise can either be averaged out with more sampling through replicates and by application of robust statistical methods, or by summarizing ranges of data and losing some resolution. Adding replicates for additional statistical power is cost-prohibitive, especially for larger genomes. Instead, we adopted the Haar wavelet transform to summarize replication data as an orthonormal series generated by the Haar wavelet. Using wavelets[25] version 1.0, we performed a maximum overlap discrete wavelet transform with the Haar wavelet using reflected boundaries and level 3 smoothing on a per-chromosome basis for each sample and replicate.

We specifically chose the Haar wavelet over other smoothing methods because it is a square function with discrete boundaries and thus resembles the signals we aim to detect. General smoothing methods like LOESS, moving average, or gaussian kernel produce stabilized trends from data, but they also change signal boundaries by spreading out the peaks. As demonstrated in Figure 5A with simulated data, the Haar wavelet accurately removes low-amplitude noise to reconstruct the original signal without artificially expanding the peaks of replication signal. Applying the moving average, LOESS, and Haar wavelet to actual *A. thaliana* data in Figure 5B

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

shows that both the moving average and LOESS can capture large trends, but the Haar wavelet excels at highlighting subtle peaks in the data without under smoothing.

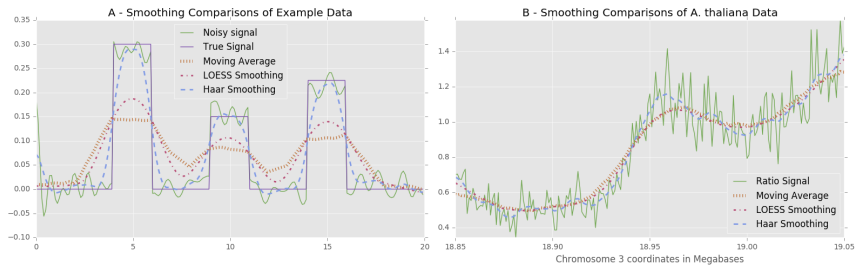

Figure 5 Smoothing comparisons. A - Noise (green) is added to an original signal (purple), and then smoothed with a moving average (orange), LOESS (red), and a Haar wavelet (blue). Both the moving average and LOESS spread out the peaks and artificially lower signal levels, while the Haar wavelet keeps bounds and peak heights close to the original. **B** - The *A. thaliana* middle S-phase normalized signal (green), is smoothed with a moving average (orange), LOESS (red), and the level 3 Haar wavelet (blue) for comparison.

Defining Replication

The analysis to this point yields a smoothed ratio of normalized replication ratio signals $r_{c_w t}$ in windows $w = 1..Y$ per chromosome $c = 1..X$, with a range of $[0, \infty)$ that can be compared to each other, and leads to the question of which signals can be considered confidently as resulting from DNA replication. Lee *et al.*[13] originally considered array-based replication signals greater than the control as actively replicating in their investigation of *A. thaliana* as follows.

$$\text{replicating}_{ct}(w) = \begin{cases} 1, & \text{if } r_{c_w t} > 1 \\ 0, & \text{otherwise} \end{cases}, \text{ where } c_w = i$$

The Repliscan software allows users to adopt this threshold method, but we also include more robust methods to define replication. The simple threshold approach above is appropriate when considering replication as a ratio, but, because all signals from the early, middle, and late S-phase samples represent labeled - and therefore, replicating - DNA, even signals that are less than the control must be considered as reflecting some level of replication activity. In other words, even though there may be noise in the data, all replication signals should be genuine because EdU is only incorporated into newly replicated DNA. Instead of simply choosing a smaller ratio threshold, we implemented a percentile cutoff based on the distribution of the ratios. By default, this method removes the lowest 2% of the values for a chromosome in a given sample.

$$\text{replicating}_{ct}(w) = \begin{cases} 1, & \text{if } r_{c_w t} > P_{0.02}(r_{ct}) \\ 0, & \text{otherwise} \end{cases}$$

While this method is a data-dependent means for establishing a cutoff, it was not considered ideal for an automatic analysis for two reasons. First, a cutoff is still being

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

dictated, even if it is more robustly supported than in previous analyses. Second, this cutoff will always remove a flat percentage of the values, even if there is no evidence they are not high-quality data points. To improve on these deficiencies, we implemented a threshold for replication that depends on the information provided in addition to the data.

To maximize the fraction of a chromosome with valid replication signal (or information), we designed an optimization method that incorporates as much of each chromosome as possible by analyzing the rate of coverage. Using data from all time points, coverage is defined as the fraction of windows with a signal greater than the threshold in at least one replication time.

$$\text{coverage}(T_c) = \frac{r_{c_w t} > T_c}{Y}, \text{ where } T_c \text{ is the threshold for chromosome } c$$

Starting from the point of the largest absolute change in coverage (slope), the replication threshold is lowered until the absolute chromosome fraction per sample/control coverage differential goes below 0.1, effectively leveling out.

$$mT_c = \arg\max_{T_c}(|\text{coverage`}(T_c)|)$$
$$\hat{T}_c = \arg\max_{T_c < mT_c}(|\text{coverage`}(T_c)| < 0.1)$$
$$R_{c_w t} = \begin{cases} r_{c_w t}, & \text{if } r_{c_w t} > \hat{T}_c \\ 0, & \text{otherwise} \end{cases}$$

This coverage differential method starts from the maximum amount of information (max absolute differential) and then lowers the threshold, incorporating smaller signals, until the rate of genome coverage slows down (Figure 6). Such a search pattern circumvents any local optima in the coverage signal that may have stalled a gradient descent. That being said, we implemented the threshold to run on a per-chromosome basis to minimize the effect of any structural differences between chromosomes.

The end result is a method that includes as much of the genome and coverage information as possible, but only returns windows with acceptable levels of coverage. Our method is generically applicable to experiments using the same Repli-seq protocol because the threshold is calculated from the data. A critical benefit is that users are not required to be masters of their data or this tool, and can instead focus on interpretation.

Classification/Segmentation

Given a signal that can confidently be considered as arising from DNA replication, we are able to classify segments of the genome according to when in the cell cycle they are replicated. Suppose that in one of the windows in Chromosome 3, we have the following levels of replication in Table 1.

We already know that any values below 0.92 in Chromosome 3 are not considered replicating, so the middle S-phase value would become 0 and we would say this window replicates in both early and late S-phase. However, the late replication level

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

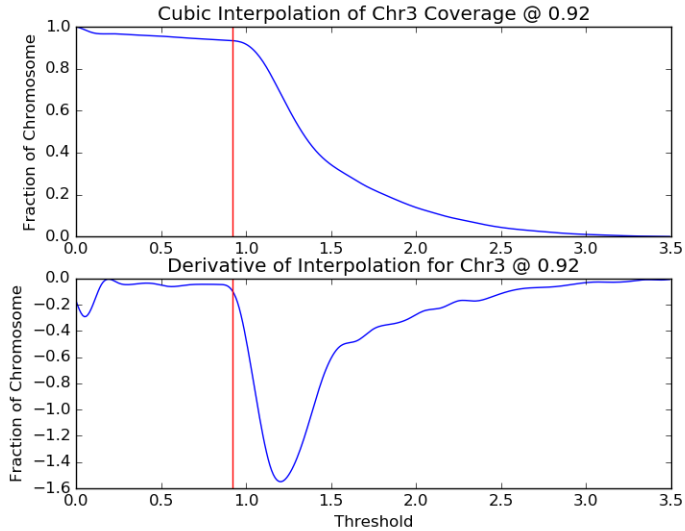

Figure 6 Replication threshold from coverage. The upper plot shows how much of *A. thaliana* chromosome 3 will be kept for downstream analysis as a function of the signal threshold. The lower plot shows the chromosome coverage differential as a function of the threshold. The vertical red line in each plot marks the optimal threshold of 0.92.

Table 1 Example coverage values to demonstrate replication timing classification.

Time	Early	Middle	Late
Coverage	0.93	0.8	3.0
Replicating	0.93	0	3.0

is 3 times higher than that of early, which is just past the threshold for replication. Instead of making another replication threshold, we implemented a general solution to compare values against each other using a proportion.

First, on a window-by-window basis, we take the infinity norm of all values, which means we divide all values by the maximum for that window position.

$$S_{ct}(w) = \frac{R_{c_w t}}{\|R_{c_w}\|_\infty}$$

This operation scales the largest value to 1 and the others to the range [0,1]. A time signal is then classified as predominantly replicating $C_c t(w)$ if the normalized value is greater than 0.5, or at least half the size of the largest signal for that window.

$$C_{ct}(w) = \begin{cases} 1, & \text{if } S_{ct}(w) > 0.5, \\ 0, & \text{otherwise} \end{cases}$$

The infinity-norm ensures that the largest value will always be classified as replicating, and this classification method allows for a window to be called strongly replicating at more than one time in S-phase (e.g. both early and late) when other signals are within 50% of the maximum value. Besides 0.5 being easy to test for, this creates an equally partitioned solution space in the form of an n-dimensional hypercube. In the case of the *A. thaliana* data, the space is a 3-dimensional cube with

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

each dimension being one of the time points: early, middle, and late S-phase. The 0.5 partition then creates 8 equal-sized sub-cubes corresponding to each possible combination of times:

{Non-replicating, Early, Middle, and Late}

along with

{Early-Middle, Middle-Late, Early-Late, and Early-Middle-Late}

S-phase replication combinations.

Results and Discussion

Data

To demonstrate the ability of our methods to adapt to different datasets, we ran our pipeline on the A. thaliana Col-0 cell culture data (PRJNA330547) that was used to develop these methods, and a separate similarly prepared Zea mays B73 replication timing dataset (PRJNA327875) also from our lab.

A.thaliana

The A. thaliana experiment was comprised of 3 early S bioreplicates, 3 middle S bioreplicates, 3 late S bioreplicates, and 1 G_1 sample. Each bioreplicate was paired-end sequenced to 36x coverage. The unique and properly-paired alignment rate for each sample was approximately 85%, yielding a total of 30x viable replication data from each sample. Due to the high coverage, we decided to use 1 kilobase windows and merge bioreplicates with the median function for our analysis.

Z. mays

In the Z. mays experiment, there were 3 early S bioreplicates, 3 middle S bioreplicates, 2 late S bioreplicates, and 2 G_1 technical replicates. Each bioreplicate was paired-end sequenced to about 5x coverage. While there were more reads than the A. thaliana experiment, the Z. mays genome is much larger, so coverage was lower. Using the B73 AGPv3 genome assembly, the unique and properly-paired alignment rate for each sample was approximately 99%, yielding a total of 5x viable replication data from each bioreplicate. We decided to use the same 1 kilobase windows for this data, and deemed the summation of bioreplicates was necessary to achieve enough coverage to highlight peaks in the data.

Segmentation Overview

Using 1 kilobase windows, median aggregation for *A. thaliana*, and sum aggregation for *Z. mays*, we used our default pipeline to classify the replication timing of our data. We generated Figure 7 to show the replication segmentation classification of Chromosome 3 in *A. thaliana* and Chromosome 10 in *Z. mays*.

In both instances, early replication is concentrated toward the ends of the chromosome arms, with middle and late replication becoming more prominent closer to the centromere and the highest concentration of late replicating sequences in the

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

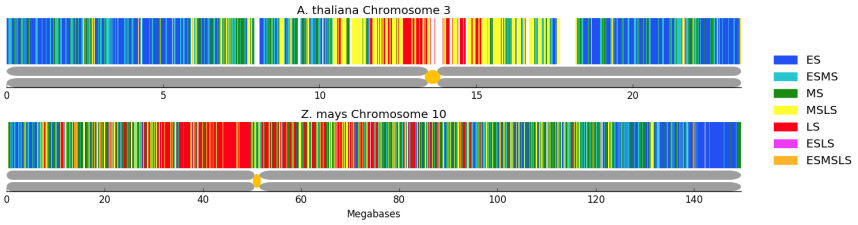

Figure 7 Comparison of A. thaliana and Z. mays segmentation. Following the segmentation legend on the right, *A. thaliana* chromosome 3 (top) and *Z. mays* chromosome 10 (bottom) have been classified into segmentation regions by repliscan. The large white regions in the *A. thaliana* figure are unclassified regions due to outlying coverage. Below each replication segmentation is a depiction of the chromosome, with the centromere location marked in yellow [26, 27].

heterochromatin surrounding the centromere. These timing maps demonstrate that the method developed using the *A. thaliana* data was successfully applied to the lower coverage *Z. mays* data after only choosing to aggregate replicates using the sum instead of the median.

Segment Composition and Size

Instead of viewing the chromosomes as a whole, we can also get an idea of predominant replication times by looking at the proportional composition. Figure 8 shows that Early, Early-Middle, and Middle-Late S-phase replication makes up most of the segmentation profiles for *A. thaliana* Chromosome 3. About 6% of the chromosome is missing around the centromere and heterochromatic knob, which probably would have been classified in the Middle to Late times based on what we do see. In *Z. mays*, we see a more uniform distribution of Chromosome 10, which is 5-fold larger, across the replication segmentation classes. Lee *et al.*[13] previously hypothesised a two-stage replication program, but our results, which were generated at higher resolution, show a more even spread (Figure 8).

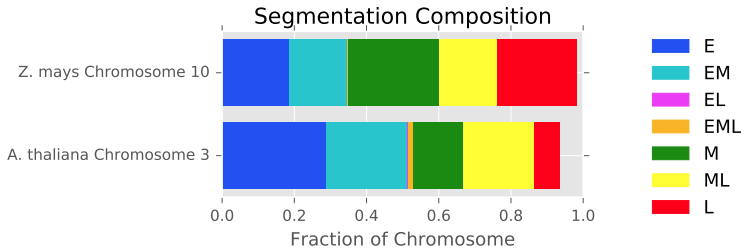

Figure 8 Composition of replication segmentation. The segment composition shows that replication in *A. thaliana* is skewed towards early S replication, while *Z. mays* has an even distribution across early, middle, and late S. We can also see that the non-sequential early-late (EL) and early-middle-late (EML) classifications comprise a very small proportion of the classified segments in both cases.

The Early-Late and Early-Middle-Late comprise a small portion of the chromosomes in both organisms and could arise naturally in the data through allelic and cell population differences. Figure 9 shows a different summary of the segmentation breakdown, highlighting the segment size distribution with boxplots. Once again,

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Early-Late and Early-Middle-Late segments are distinct in that their lengths are small relative to the other timing categories.

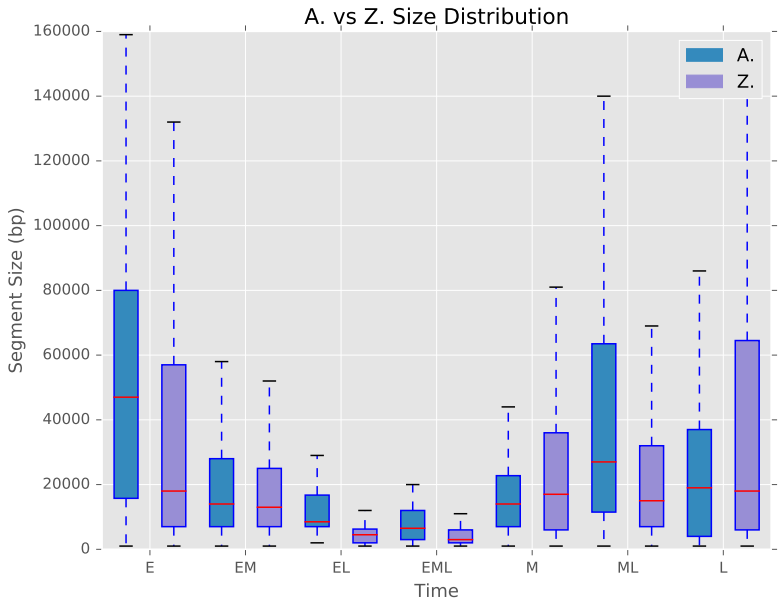

Figure 9 Segment size distribution. Boxplots for every combination of replication time, illustrating the distribution of segment sizes. Early (E) and mid-late (ML) S were largest in *A. thaliana*, while early and late (L) were largest in *Z. mays*.

Downsampling and Stability of Results

The relatively small genome size of *A. thaliana* allowed us to obtain extremely deep sequencing coverage, which is currently cost-prohibitive for larger genomes. To estimate a minimum coverage requirement for our methods, we simulated experiments with lower coverage via downsampling. We first generated 3 technical replicates by randomly sorting the original alignment files. We removed reads from each of the replicates in 1% increments without replacement. Each of the 300 (100 x 3) simulated experiments were analyzed using both median and sum aggregation, and no (none), log gamma, square root gamma, normal, and whisker outlier removal. To account for differences arising from the sorting order, the final classification for each window was determined by majority across the 3 replicates. Classification ties were broken by treating the early, middle, and late time classification combination as a 2-bit binary number, and taking the median.

After confirming that the segmentation profiles from all three 100% replicates were identical to our original segmentation, differences for each run type were calculated as percent differences from the 100% version. All differences were compounded and plotted as a fraction of the whole chromosome in Figure 10. The most obvious results are the spikes of differences in both the median and sum log transformed gamma runs when the iterative fitting function failed to converge (Figure 11).

Shifting attention to the square root gamma experiments in Figure 10, we see that while the fit function never fails to converge, but there is increased variability

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

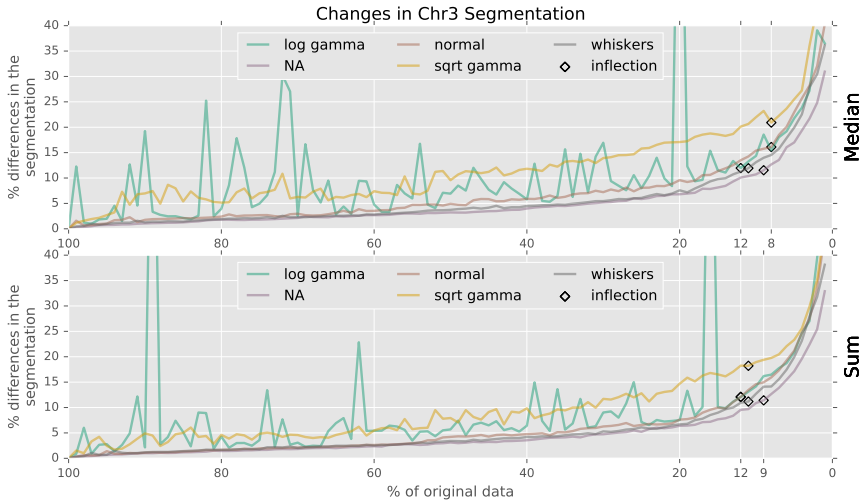

Figure 10 Segmentation differences in downsampled data. After downsampling the *A. thaliana* data, the accuracy of median (top) and sum (bottom) aggregation, and outlier detection using log gamma, none (NA), normal, square root gamma, and whiskers. Inflection points in the differences are labeled with black diamonds.

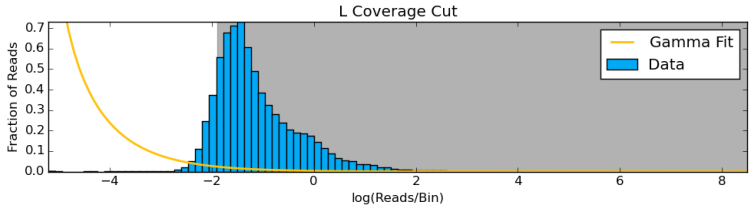

Figure 11 Unconverged log gamma fit. Most of the data is removed when the iterative fitting function fails to converge with the log transformed gamma distribution. Instances like this produce the spikes of differences in Figure 10.

of results among each level of downsampling. All other probability functions are very stable between downsampling runs. We even see that summing the coverage to 90x provides no improvement over the median - even at low coverage levels. The inflection point of differences show that the most stable method was aggregating replicates with the median operation and removing coverage by fitting a normal distribution to the log transformed data. Results from this method began to drastically diverge when downsampled to 8%, or 2.4x coverage. This indicates 5x coverage for the commonly studied species *Z. mays* (2.3 gigabase genome[28]) is sufficient to calculate a replication profile, which is quite tractable for a laboratory of modest financial means.

Conclusions

Based on our results from running Repliscan on both *A. thaliana* and *Z. mays* data, we have demonstrated that our methods offer a general solution to analyze replication timing experiments that use label incorporation and a G_1 control. We have significantly improved on previous methods by incorporating non-destructive Haar

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

smoothing, defining replication through optimization, and classification through proportion. When run using the same parameters with data from different organisms, the methods automatically tuned their thresholds to adjust for differences in coverage. Downsampling our data even showed our methods provided stable results at as little as 2.4x coverage.

In-depth explorations of results with these two species will be published separately. We hope these methods provide a path for greater understanding of the DNA replication program in plants, humans, and other higher organisms.

List of Abbreviations

TimEx: Time of replication
Repli-seq: Replication label incorporation sequencing
Edu: 5-Ethynyl-2'-deoxyuridine
BrdU: 5-Bromo-2'-deoxyuridine
NGS: Next generation sequencing
SRA: Sequence read archive
G_1: Gap 1 of cell division
G_2: Gap 2 of cell division
S: Synthesis phase of cell division
E: Early S-phase replication
M: Middle S-phase replication
L: Late S-phase replication
WB: Whisker bounds

Availability and requirements

Project name: Repliscan
Project home page: https://github.com/zyndagj/repliscan
Operating systems: Linux, OS X
Programming languages: Python v2.7
Other requirements: scipy v0.15.0+, samtools, bedtools v2.24.0+, wavelets v1.0, numpy, matplotlib.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Acknowledgements
Special thanks to everyone in the Hanley-Bowdoin and Thompson labs.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the NCBI Sequence Read Archive (SRA) BioProjects PRJNA330547 and PRJNA327875.

Funding
This work is supported by NSF IOS-1025830 "Epigenome Dynamics During DNA Replication".

Competing interests
The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Author's contributions

GJZ developed and implemented the algorithm, and wrote the manuscript. EEW, LC, WFT, and LHB provided biologial expertise and produced the data. JS provided feedback on the methods and helped revise the manuscript. All authors read, helped revise, and approved the final manuscript.

Author details

[1]Texas Advanced Computing Center, University of Texas at Austin, 10100 Burnet Road, 78758-4497 Austin, TX, USA. [2]Department of Plant and Microbial Biology, North Carolina State University, 27695-7612 Raleigh, NC, USA.

References

1. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Cell junctions, cell adhesion, and the extracellular matrix (2002)
2. Hand, R.: Eucaryotic dna: organization of the genome for replication. Cell **15**(2), 317–325 (1978)
3. Hansen, R.S., Canfield, T.K., Lamb, M.M., Gartler, S.M., Laird, C.D.: Association of fragile x syndrome with delayed replication of the fmr1 gene. Cell **73**(7), 1403–1409 (1993)
4. Woodfine, K., Fiegler, H., Beare, D.M., Collins, J.E., McCann, O.T., Young, B.D., Debernardi, S., Mott, R., Dunham, I., Carter, N.P.: Replication timing of the human genome. Human molecular genetics **13**(2), 191–202 (2004)
5. Woodfine, K., Beare, D.M., Ichimura, K., Debernardi, S., Mungall, A.J., Fiegler, H., Collins, V.P., Carter, N.P., Dunham, I.: Replication timing of human chromosome 6. Cell Cycle **4**(1), 172–176 (2005)
6. Schübeler, D., Scalzo, D., Kooperberg, C., van Steensel, B., Delrow, J., Groudine, M.: Genome-wide dna replication profile for drosophila melanogaster: a link between transcription and replication timing. Nature genetics **32**(3), 438–442 (2002)
7. Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., Ikemura, T.: Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. Human Molecular Genetics **11**(1), 13–21 (2002)
8. Hayashi, M., Katou, Y., Itoh, T., Tazumi, M., Yamada, Y., Takahashi, T., Nakagawa, T., Shirahige, K., Masukata, H.: Genome-wide localization of pre-rc sites and identification of replication origins in fission yeast. The EMBO journal **26**(5), 1327–1339 (2007)
9. White, E.J., Emanuelsson, O., Scalzo, D., Royce, T., Kosak, S., Oakeley, E.J., Weissman, S., Gerstein, M., Groudine, M., Snyder, M., *et al.*: Dna replication-timing analysis of human chromosome 22 at high resolution and different developmental states. Proceedings of the National Academy of Sciences of the United States of America **101**(51), 17771–17776 (2004)
10. MacAlpine, D.M., Rodríguez, H.K., Bell, S.P.: Coordination of replication and transcription along a drosophila chromosome. Genes & development **18**(24), 3094–3105 (2004)
11. Farkash-Amar, S., Lipson, D., Polten, A., Goren, A., Helmstetter, C., Yakhini, Z., Simon, I.: Global organization of replication time zones of the mouse genome. Genome research **18**(10), 1562–1570 (2008)
12. Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.-W., Lyou, Y., Townes, T.M., Schübeler, D., Gilbert, D.M.: Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol **6**(10), 245 (2008)
13. Lee, T.-J., Pascuzzi, P.E., Settlage, S.B., Shultz, R.W., Tanurdzic, M., Rabinowicz, P.D., Menges, M., Zheng, P., Main, D., Murray, J.A., *et al.*: Arabidopsis thaliana chromosome 4 replicates in two phases that correlate with chromatin state. PLoS Genet **6**(6), 1000982 (2010)
14. Bass, H.W., Wear, E.E., Lee, T.-J., Hoffman, G.G., Gumber, H.K., Allen, G.C., Thompson, W.F., Hanley-Bowdoin, L.: A maize root tip system to study dna replication programmes in somatic and endocycling nuclei during plant development. Journal of experimental botany, 470 (2014)
15. Bass, H.W., Hoffman, G.G., Lee, T.-J., Wear, E.E., Joseph, S.R., Allen, G.C., Hanley-Bowdoin, L., Thompson, W.F.: Defining multiple, distinct, and shared spatiotemporal patterns of dna replication and endoreduplication from 3d image analysis of developing maize (zea mays l.) root tip nuclei. Plant molecular biology **89**(4-5), 339–351 (2015)
16. Wear, E.E., Concia, L., Brooks, A.M., Markham, E.A., Lee, T.-J., Allen, G.C., Thompson, W.F., Hanley-Bowdoin, L.: Isolation of plant nuclei at defined cell cycle stages using edu labeling and flow cytometry. Plant Cell Division: Methods and Protocols, 69–86 (2016)
17. Krueger, F.: Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files (2015)
18. Li, H.: Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. arXiv preprint arXiv:1303.3997 (2013)
19. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., *et al.*: The arabidopsis information resource (tair): improved gene annotation and new tools. Nucleic acids research **40**(D1), 1202–1210 (2012)
20. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., *et al.*: The sequence alignment/map format and samtools. Bioinformatics **25**(16), 2078–2079 (2009)
21. Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B.: Genome-wide mapping of in vivo protein-dna interactions. Science **316**(5830), 1497–1502 (2007)
22. Quinlan, A.R., Hall, I.M.: Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)
23. Diaz, A., Park, K., Lim, D.A., Song, J.S., *et al.*: Normalization, bias correction, and peak calling for chip-seq. Stat Appl Genet Mol Biol **11**(3), 9 (2012)
24. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python. [Online; accessed 2016-11-28] (2001–). http://www.scipy.org/
25. Percival, D.B., Walden, A.T.: Wavelet Methods for Time Series Analysis vol. 4. Cambridge university press, The Edinburgh Building, Cambridge CB2 2RU, UK (2006)

bioRxiv preprint doi: https://doi.org/10.1101/094177; this version posted January 29, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

26. Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S., Jiang, J.: Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional dna element of arabidopsis thaliana centromeres. Genetics **163**(3), 1221–1225 (2003)
27. Zhao, H., Zhu, X., Wang, K., Gent, J.I., Zhang, W., Dawe, R.K., Jiang, J.: Gene expression and chromatin modifications associated with maize centromeres. G3: Genes— Genomes— Genetics **6**(1), 183–192 (2016)
28. Schnepf, E.: From prey via endosymbiont to plastids: comparative studies in dinoflagellates. In: Lewin, R.A. (ed.) Origins of Plastids vol. 2, 2nd edn., pp. 53–76. Chapman and Hall, New York (1993)