

A link between repetitive sequences and gene replication time

M. Regelson^a C.D. Eller^a S. Horvath^{a, b, 1} Y. Marahrens^{a, 1}UCLA Departments of ^aHuman Genetics and ^bBiostatistics, Gonda Center, David Geffen School of Medicine, Los Angeles, CA (USA)

Manuscript received 8 April 2005; accepted in revised form for publication by J. Greally, 8 August 2005.

Abstract. Genes display a wide range of replication times in S phase. In general, late replication is associated with transcriptionally repressive states and early replication with transcriptional competence. Rare examples of early-replicating repressive states have also been identified that are consistent with molecular evidence that repressive states are not all uniform in nature. Here we show that the replication times of over 4000 *Drosophila* genes correlate with the abundance of repetitive sequences in ~200-kb regions flanking the genes. In particular, Satellite-Related sequences (SRs) and the simple sequence repeats (SSRs) (CA)_n and (ACTG)_n were increasingly abundant in the regions flanking progressively lat-

er replicating genes, while (CATA)_n repeats were more abundant around earlier replicating genes. These four sequences comprise less than 0.5% of the 'euchromatic genome' in *Drosophila*, yet they account for 5% of the variation of gene replication timing. Although the effect is not strong, it is broad: 99% of the genome is within the region of correlation of at least one of the above repeats. The role of SSRs and non-centromeric SRs in the genome is not known. We propose that SSRs and SRs foster transcriptionally repressive states throughout the genome in order to minimize spurious transcription.

Copyright © 2006 S. Karger AG, Basel

The average replication times of genes vary greatly across the genome. Genes in late replicating regions tend to be repressed, while those in early replicating regions are more likely to be expressed (reviewed in Gilbert, 2002). Some transcriptionally inactive late replicating genes become early replicating when expressed, while other genes remain late replicating when expressed (Gilbert, 2002). Thus far, few factors are known to affect replication time. Histone acetylation

has been shown to determine the time in S phase that an origin of replication is activated in yeast (Vogelauer et al., 2002). The signaling molecules ATM and ATR have also been implicated in the temporal control of origin activation (Shechter et al., 2004) and both ATM (Schmidt and Schreiber, 1999) and ATR (Kim et al., 1999) have been reported to associate with histone deacetylases. In line with these findings, there is a correlation between replication time and epigenetic state with heterochromatin generally replicating later in S phase than euchromatin (Gilbert, 2002). Elements in the underlying DNA sequence that help to determine the epigenetic state may therefore indirectly influence times of replication. The only sequences that have previously been shown to be associated with both late replication timing and transcriptionally repressive states are sequence-specific protein binding sites in yeast (Rivier and Rine, 1992; Zappulla et al., 2002) and repetitive sequences.

For the majority of repetitive sequences no biological role is known and they are frequently considered 'junk' DNA.

Supported by USHHS Institutional National Research Service Award #T32 CA09056 (M.R.), and by NIH grant R01 HD041451 (Y.M.) and NSF DGE-9987641 (C.D.E.).

Request reprints from York Marahrens

UCLA Department of Human Genetics, Gonda Center
Room 4554B, 695 E. Charles E. Young Drive South
Los Angeles, CA 90095-7088 (USA)
telephone: 310-267-2466; fax: 310-794-5446
e-mail: YMarahrens@mednet.ucla.edu

Steve Horvath, UCLA Departments of Human Genetics and Biostatistics
Gonda Center, Room 4357A, 695 E. Charles E. Young Drive South
Los Angeles, CA 90095-7088 (USA)
telephone: 310-825-9299; fax: 310-794-5446
e-mail: SHorvath@mednet.ucla.edu

¹ Both authors co-directed this work.

However, repetitive sequences have been associated with replication time variation and the silencing of gene expression. For example, a tandem simple sequence repeat resides in an intron of the *FMRI* gene. Expansions in the length of this repeat past a certain threshold number result in the formation of a late replicating domain, transcriptional silencing of the *FMRI* gene, and Fragile-X syndrome (Webb, 1992; Hansen et al., 1993). In another example, high concentrations of the long interspersed nuclear element (LINE) transposon sequence have been proposed to be responsible for late replication and transcriptional silencing during X-inactivation (Lyon, 1998; Marahrens, 1999; Bailey et al., 2000; Hansen, 2003). Repetitive sequences have also been found to be important for centromere function (Lica et al., 1986; Wevrick and Willard, 1989; Haaf et al., 1992; Harrington et al., 1997; Platero et al., 1999; Bernard and Allshire, 2002) and telomere function (Chan and Blackburn, 2004).

Given the association between repetitive sequences and replication time in the above examples, we investigated to what extent replication time variation across a genome can be explained by repetitive sequence environment. The availability of a genome-wide replication timing profile makes *Drosophila melanogaster* Kc cells (Schubeler et al., 2002) ideal for studying this relationship. We compared published replication timing of more than 4000 genes with the repetitive sequence content in regions of various sizes about the genes. We present statistically significant correlations between several repetitive sequences and gene replication time. Furthermore, we present evidence that individual repetitive sequences correlate to replication time over remarkably large regions of the genome and may collectively influence the replication time of virtually all genes. Many biological principles underlying *Drosophila* genome function have been shown to be shared with other eukaryotes and our results may therefore extend to other eukaryotic organisms as well.

Materials and methods

Replication time profile

To perform our analyses, we integrated data from several different sources. Schubeler et al. (2002) constructed a genome-wide replication timing map using the *D. melanogaster* Kc cell line that was originally derived from embryo cultures (Cherbas et al., 1977). They isolated cells that incorporated BrdU into their DNA during early and late S-phase based on DNA content and then purified the BrdU-containing DNA from these fractions. They then simultaneously hybridized both the early and late fractions to a microarray containing 6500 cDNA sequences, 5543 of which represented expressed genes from the *Drosophila* Gene Collection (DGC) (Rubin et al., 2000) (<http://dgc.cgb.indiana.edu/vectors/store/dgc.html>), and determined the ratio of the early and late signals. These timing assays were replicated three times and the average results for 5221 clones are available at <http://parma.fhcr.org/DSchubeler/>. Gene expression data was obtained from the same Kc cell line using the same cDNA microarrays.

Repeat environment

Gene location and transcription start site were extracted from the Berkeley *Drosophila* Genome Project (BDGP) *Drosophila* Annotation, Release 3.1, available through the UCSC genome browser (<http://genome.ucsc.edu/>). We converted between the DGC clone identifier and the BDGP identifier using the `bdgpGeneInfo.txt` file (<ftp://hgdownload.cse.ucsc.edu/goldenPath/dm1/database/>) that contains BDGP identifiers, FlyBase identifiers, and DGC clone identifiers. We eliminated clones from our analysis that had no associated BDGP identifier, with more than one associated BDGP identifier, or for which the BDGP identifier was associated with more than one clone. In addition, we omitted clones with ambiguous gene expression value, leaving us with data for a total of 4389 clones.

To determine the location of the repeats we studied, we used the `chr*_rmsk.txt` files downloaded from <ftp://hgdownload.cse.ucsc.edu/goldenPath/dm1/database/>. Having determined the gene associated with each clone and the locations of all identified repeats in the *Drosophila* genome, we then determined the repeat counts within windows about those genes using Perl scripts. These scripts and our processed data can be found at <http://www.genetics.ucla.edu/labs/horvath/Replication-AndRepeats/>. Note that for SSRs we combined occurrences of the reverse complement of a repeat pattern with the pattern itself. For example, $(CA)_n$ represents instances of both $(CA)_n$ and $(TG)_n$ in the *Drosophila* genome. When considering potential effects on replication time, the strand of the repeat was assumed not to matter. Replication time information was originally provided as \log_2 of a ratio, however we chose to use the raw ratio data for our analyses because it better satisfied the linear model assumptions (normality and homoscedasticity). We scaled the data to the interval (0,1) for ease of interpretation, with smaller values representing earlier replication and larger values representing later replication.

Statistical methods

All statistical analyses were conducted using the R software package (freely available at <http://cran.r-project.org/>). Pearson correlations and their *P* values were computed using the `cor.test` function. The `cor.test` statistic is based on Pearson's product moment correlation coefficient and follows a *t* distribution with $n - 2$ degrees of freedom if the samples follow independent normal distributions. This test is identical to the Wald test *P* value of the univariate linear regression model where replication time is regressed on the single covariate. The model-fitting diagnostics for the multivariable linear regression module (Supplementary Fig. 1, www.karger.com/doi/10.1159/000089869) provide evidence that the assumption of normality is satisfied at least approximately for our replication time data. We checked for constant error variance by plotting the residuals versus predicted values for each of our linear models (Supplementary Fig. 1, www.karger.com/doi/10.1159/000089869).

Linear regression modeling to determine the relationship between repeat environment and replication time was performed using the function `lm` (Chambers, 1992). The percent of variance explained by the linear model was determined using the (unadjusted) R^2 value of the model. We used the *F*-statistic to test whether the linear model with covariates (repeat counts) contains more predictive information for replication time than the null model that contains only an intercept term.

Since we considered 347 different repeat types, we carried out several approaches to adjust for multiple comparisons. We used permutation test approaches to estimate permutation *P* values.

For the multivariable linear model, the permutation test (Davison and Hinkley, 1997) was based on the *F*-statistic. To arrive at a permutation test *P* value we permuted the replication time, leaving the covariates (repeat counts) unchanged, and stored the *F*-statistic of each multivariable regression model. With $r = 2000$ repetitions of this procedure, the permutation test *P* value is $(s + 1)/(r + 1)$ (Davison and Hinkley, 1997), where s equals the number of permuted data sets for which the resulting test statistic was larger than or equal to that of the unpermuted, observed data set.

For the univariate linear regression model that tests whether a single repeat count is correlated with replication time, the permutation test applies to the null hypothesis of independence between covariate *X* and outcome *Y* when these are both random or when the conditional distribution of *Y* given *X* = x does not depend on x (Davison and Hinkley, 1997). To begin with, we chose a correlation threshold of 0.06, which on our data set of 4389 genes corresponds to an uncorrected *P* value of approximately 0.00001. We generated permutation samples by fixing the covariate values and randomly permuting the *y* values. As an additional check on significance, we computed the Bonferroni-corrected *P* values for our primary set of correlations. Since the *P* value for the correlation

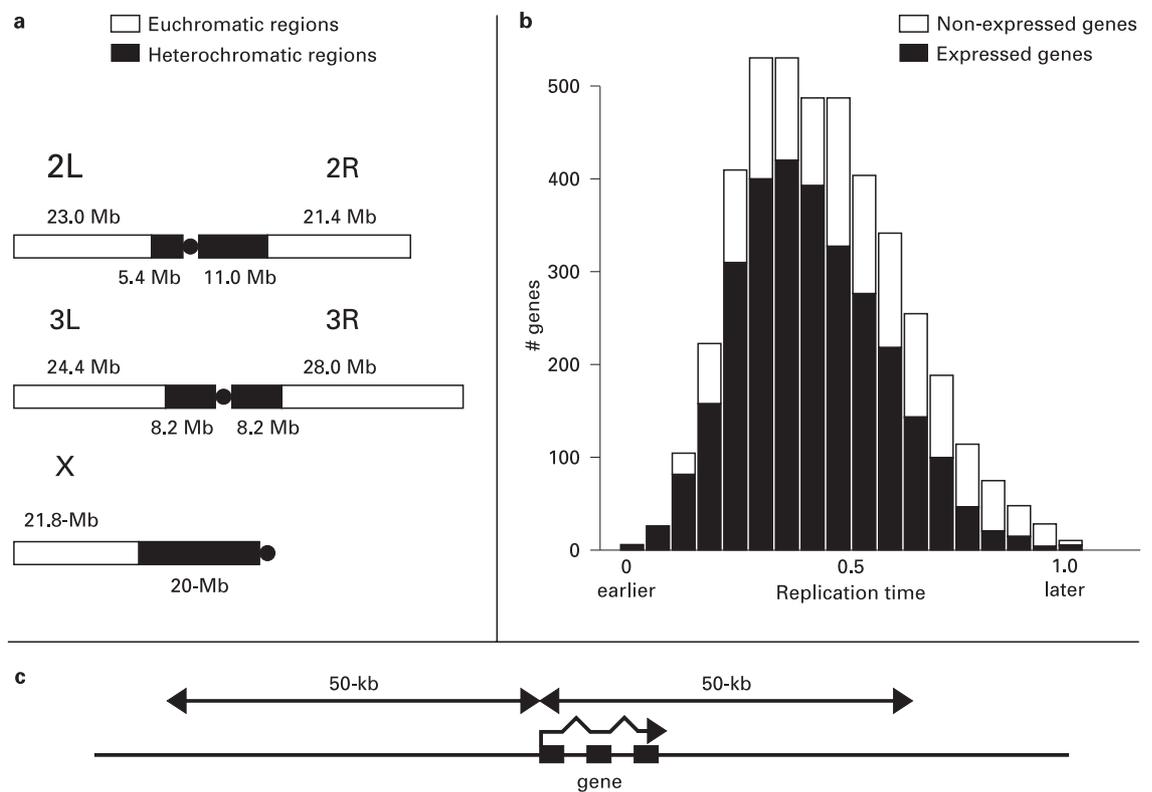


Fig. 1. 4389 *Drosophila* genes with known replication times were selected for analysis. **(a)** The *Drosophila* genome consists of alternating heterochromatic and (predominantly) euchromatic regions (Adams et al., 2000). All 4389 genes examined lie within the five euchromatic regions shown. **(b)** Distribution of the replication times in S-phase of the 4389 genes. Larger replication time values indicate later replication. **(c)** For each gene, repetitive sequence content was determined for a region extending from 50 kb upstream to 50 kb downstream of the transcriptional start site, for a net region size of 100 kb.

may be inaccurate for extreme values, our inference is based on the permutation test *P* value. We include the asymptotic *P* values of the correlation in our tables, but they should be considered as descriptive and not inferential measures.

We also considered the ‘False Discovery Rate’ (i.e., the number of observations above our threshold expected to be false positives). False Discovery Rate (FDR) (Benjamini and Hochberg, 1995; Storey, 2002) is an approach to the multiple comparisons problem which controls the expected proportion of false positives among observations exceeding a threshold value. The FDR threshold is determined from the observed *P* value distribution for a data set and thus adapts to the amount of signal present. We use the approach implemented in the ‘fdr.control’ function in the R package GeneTS (Benjamini and Hochberg, 1995; Storey, 2002) to estimate what *P* value threshold would lead to a false discovery rate of 1% as a check on the validity of our results.

Estimating the proportion of the genome whose replication time correlates with repetitive sequence

We computed a characteristic distance (the ‘region of correlation’) for the repetitive sequence types we studied in detail. In order to determine the proportion of the genome within this distance of one or more copies of the repeat, we considered each position in the genome. For each position, we counted the number of copies of the given repeat within a flanking region of the specified size. In this way we were able to determine the fraction of genomic positions within range of one, two, etc. copies of each repeat type. The perl scripts used to perform these computations are available at <http://www.genetics.ucla.edu/labs/horvath/Replication-AndRepeats/>.

Results

Determining the repeat environment and replication time

The *Drosophila* genome can be partitioned into alternating ‘heterochromatic regions’ that contain few genes, and ‘euchromatic regions’ that contain the vast majority of the genes (Adams et al., 2000). The largest euchromatic regions are the left and right arms of chromosome 2, the left and right arms of chromosome 3, and the long arm of the X chromosome (Fig. 1a). These five euchromatic regions will henceforth be referred to as the ‘euchromatic’ portion of the *Drosophila* genome.

Using published replication timing data (Schubeler et al., 2002), we selected genes with known expression status located in the aforementioned euchromatic portion of the *Drosophila* genome. These genes were found to exhibit a broad range of average replication times (Fig. 1b). Despite a clear correlation between replication time and expression (Schubeler et al., 2002), many genes with replication times in mid- or late-S phase were expressed (Fig. 1b).

The types and positions of repetitive sequences within the euchromatic portion of the *Drosophila* genome were identified using RepeatMasker (Smit et al., 1996–2004) output obtained through the UCSC genome browser. 7.2% of the sequence in the euchromatic portion of the *Drosophila* genome

was found to consist of dispersed repeats, primarily long transposon sequences. These repeats fall into the categories of LTR-transposons, DNA transposons, non-LTR retrotransposons (Long Interspersed Nuclear Elements (LINEs), Short Interspersed Nuclear Elements (SINEs), etc.), Satellite and Satellite-Related (SR) repeats, low-complexity repeats ('Low-complexity') and Simple Sequence Repeats (SSRs) (Table 1). The SSRs in our study consist of tandem repeated copies of a short sequence. We refer to segments of repetitive sequence, whether an SSR, a LINE or a SINE as 'repeats'. We found 347 different repeat types flanking the genes in our data set.

The repetitive sequence content in the regions flanking each gene was recorded. For each gene, we first determined chromosomal start and end locations from the BDGP annotation. We then considered the flanking regions of each gene extending up- and downstream from the transcription start site. Using the RepeatMasker output, we determined the types and locations of repetitive sequences within each of these regions.

Significant correlation between repeat environment and replication time

For the euchromatic portion of the *Drosophila* genome, we considered the relationship between gene replication time and the number of each of the individual repetitive sequence types in its environment. We performed our analysis with windows of 50 kb on either side of the transcription start site (Fig. 1c), but our results are robust with respect to the window size, as we will show below (Fig. 3). We will refer to the region defined by 50 kb to either side of a gene's transcription start site as the 100-kb flanking region of a gene.

We first regressed replication time on all of the 347 repeat types found flanking the genes in our dataset. The correspond-

ing multivariable linear model fitting index (R^2) equaled 0.238 and the model F-statistic was 3.74 with 339 and 4049 degrees of freedom. The corresponding F-test *P* value is smaller than $2.2e-16$, but we also carried out a permutation test to assess the significance of this finding. We permuted the replication times and regressed the permuted values on the repetitive sequence information. This calculation was performed 2000 times. In no case did we observe an F-statistic greater than 1.3, a maximum well below our observed value of 3.74. The permutation *P* value associated with this F-statistic is thus $(0 + 1)/(2000 + 1) = 0.0005$.

To gain a more detailed understanding of the relationship between replication time and repetitive sequence environment, we next considered the correlation between gene replication time and each repeat type individually. The correlations with the greatest statistical significance (as measured by *P* value) for 100-kb total flanking regions were determined (Table 2). Nearly all of the repeats that correlated most significantly with late or early replication time were SSRs with

Table 1. Abundance of the various interspersed repeats in the euchromatic portion of the *Drosophila* genome

Class of repeat	Count	Euch. %
DNA transposons	3978	0.8
LINEs, SINEs, etc.	1314	1.0
Low-complexity	30398	1.0
LTR transposons	3633	3.0
RNA genes	100	0.0
Satellite and SR	493	0.2
SSRs	30181	1.2
Unknown	488	0.1

Table 2. List of repetitive sequences with significant correlation to replication time ($P < 1.0e-5$). *P* values were obtained from the correlation between the number of repeats in the gene's 100-kb flanking region and replication time. For each repeat type we show the correlation, the associated *P* value, the Bonferroni-corrected *P* value, the number of genes in the computation, and the number of copies of the repeat in the euchromatic portion of the genome. Repeats with fewer than 50 copies in the genome are omitted ((CCCG)_n, (CACCG)_n, (CAGTC)_n), as are repeats showing significant correlation on only one chromosome arm ((ATG)_n, DMR.DM).

Name	Correlation	<i>P</i> value	Bonferroni <i>P</i> value	Gene count	Total count
CA/TG	0.15	0.0E+00	0.0E+00	3718	5774
ACTG/CAGT	0.09	2.9E-10	1.0E-07	920	401
SR	0.09	5.4E-09	1.9E-06	664	421
CATA/TATG	-0.08	3.8E-08	1.3E-05	2089	939
AT-rich	0.07	2.3E-06	8.1E-04	4389	26656
TCCA/TGGA	0.07	3.0E-06	1.0E-03	735	286
CAGAGA/TCTCTG	0.07	3.1E-06	1.1E-03	418	156
CAGG/CCTG	0.07	8.4E-06	2.9E-03	111	51
CGGAA/TTCCG	0.07	9.8E-06	3.4E-03	134	53
CAAAA/TTTTG	-0.07	1.1E-05	3.9E-03	475	115
LINEJ1_DM	0.07	1.4E-05	5.0E-03	215	106
A/T	0.06	2.6E-05	9.1E-03	2668	1932
CATCC/GGATG	0.06	5.5E-05	1.9E-02	467	173

tandem reiterations of one to five bases. The correlations between replication time and repetitive sequence count were fairly small in magnitude, but some displayed a high level of significance. The correlations and *P* values for all repeats occurring in the regions flanking the genes in our set are reported in Supplementary Table 1, www.karger.com/doi/10.1159/000089869.

To determine the extent of potential random effects, we performed a permutation test. Initially we observed that a correlation coefficient of 0.06 corresponded to a nominal *P* value of approximately 0.00001. We permuted the timing data 2000 times and computed the correlation coefficients against the counts of each repeat type. For each permutation, we counted the number of times that one or more of the correlation coefficients exceeded 0.06: in this case no values exceeded the threshold in 1952 trials, one value exceeded it in 42 trials, two values exceeded it in four trials and three values exceeded it in two trials. The permutation *P* value associated with this threshold is thus $(42 + 4 \times 2 + 2 \times 3 + 1)/(2000 \times 347 + 1) = 0.00008$. In addition, we computed the Bonferroni correction, which would interpret the nominal threshold of *P* = 0.00001 as a corrected *P* value of $0.00001 \times 347 = 0.00347$.

Clearly there is a large difference between the approximate (asymptotic) *P* value at our correlation coefficient threshold and the permutation *P* value. We take the conservative approach of basing our inference on the permutation test *P* value. However, for the sake of completeness, we report the asymptotic *P* values of the correlation as a descriptive measure.

As a further check on the suitability of a nominal *P* value threshold of 0.00001, we considered the 'False Discovery Rate' (FDR) expected with this data set and distribution of *P* values. Using the R function `fdr.control`, we determined that the *P* value threshold associated with an FDR of 1% ($Q = 0.01$) is 0.0007. Since our *P* value threshold is well below that value, we expect an FDR of less than 1%.

116 out of 347 repeat types considered (33%) were found to have a *P* value < 0.01. However we only report findings that were significant at a *P* value level of 0.00001 to avoid false positives due to multiple comparisons, as discussed above. Of the 347 repeat types considered, 18 (5%) showed *P* values below our threshold of 0.00001 (Table 2). Of these, three were found to have fewer than 50 copies in the genome ((CCCG)_n, (CACCG)_n and (CAGTC)_n) and were omitted from the analysis. An additional two repeat types ((ATG)_n and DMR.DM) were omitted because the significance of their correlation was limited to one chromosome arm (Supplementary Table 2, www.karger.com/doi/10.1159/000089869).

The repetitive sequence that correlated most significantly with late or early replication time in the euchromatic portion of the *Drosophila* genome was the tandemly reiterated CA sequence (CA)_n. (CA)_n was more abundant in the 100-kb flanking regions of late replicating genes than the flanking regions of early replicating genes. Several additional SSRs were significantly more abundant near late replicating genes, while tandemly iterated CATA ((CATA)_n) was significantly more abundant in the regions flanking early replicating genes. The only repeats that correlated significantly with replication time but were not SSRs were the SR repeat and LINEJ1_DM (Ta-

ble 2). We not only observed the relationship between replication time and repetitive sequences on the whole data sets but typically also on each chromosome individually, as can be seen from Supplementary Table 2, www.karger.com/doi/10.1159/000089869. This provides an internal validation of our findings.

The repetitive sequences (ACTG)_n, (CA)_n, (CATA)_n, and SR displayed the most significant association with replication time and were selected for further study. For a subset of our analyses, we also considered a set of repeats with somewhat less significant associations ((A)_n, AT-rich, (CAAAA)_n, (CAGAGA)_n, (CAGG)_n, (CATCC)_n, (CGGAA)_n, LINEJ1_DM, and (TCCA)_n). As controls, we also considered some abundant repeats that showed little or no association with replication time ((CAA)_n, (CAG)_n, (CATATA)_n, DNAREP1_DM, (GA)_n, ROO_I and (TA)_n). In order to estimate the total impact of these repeats on replication time, we regressed replication time on the repeat counts using a linear model (Chambers, 1992). Using this model, we found that counts of the repeats (CA)_n, (CATA)_n, (ACTG)_n, and SR in the 100-kb region flanking a gene's transcription start site were collectively highly significant covariates of replication time (asymptotic F-test *P* value < 2.2e-16). Using the model fitting index (*R*²) we found that 4.7% of the variation in replication time can be explained by considering these counts. Including the repeats (A)_n, AT-rich, (CAAAA)_n, (CAGAGA)_n, (CAGG)_n, (CATCC)_n, (CGGAA)_n, LINEJ1_DM, and (TCCA)_n explains 7.6% of replication timing variation. The coefficients of the repeat counts in the multivariable regression model can be found in Supplementary Table 3, www.karger.com/doi/10.1159/000089869. Our model indicates that repeat environment explains a modest portion of the variance in replication time with high significance. Other factors are therefore clearly involved in fully determining a gene's replication time. Although we find evidence for moderate correlations between the repeat counts, these correlations are not large enough to cause problems with multi-collinearity in the linear regression model (Supplementary Table 4, www.karger.com/doi/10.1159/000089869).

Although the correlations between individual repetitive sequence types and replication time are small, the linear model shows that the combination of a set of repeat types has a stronger association. To illustrate this, we pooled repeats with a significant negative correlation ((CATA)_n and (CAAAA)_n), those with a significant positive correlation ((A)_n, (ACTG)_n, AT-rich, (CA)_n, (CAGAGA)_n, (CAGG)_n, (CATCC)_n, (CGGAA)_n, LINEJ1_DM, SR and (TCCA)_n), and the abundant repeats with little correlation to replication time mentioned above ((CAA)_n, (CAG)_n, (CATATA)_n, DNAREP1_DM, (GA)_n, ROO_I and (TA)_n). We combined these repeats in an optimal fashion by using the linear regression coefficients to form a weighted average quantity for each set: *N*₁, *N*₂, and *N*₃ respectively. The correlation between the quantity and replication time is 0.10 for *N*₁, 0.23 for *N*₂ and 0.05 for *N*₃. While these correlations are still small, they reflect the increased predictive power gained by combining repeat types. The correlation for *N*₃ remains below our correlation threshold for significance.

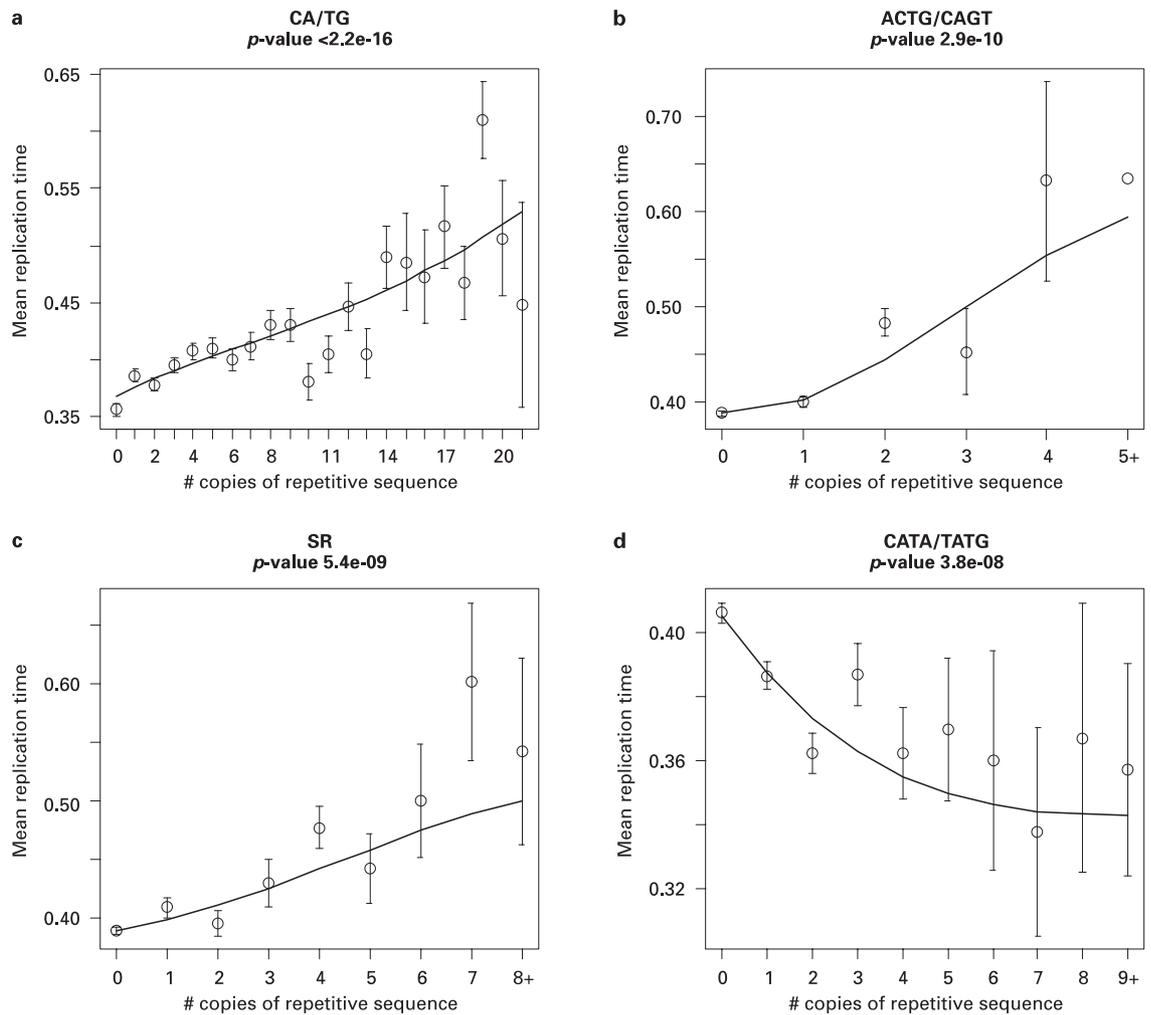


Fig. 2. Plot of average gene replication time versus the number of copies (tracts) of a repetitive sequence in the 100-kb region flanking the gene's transcription start site. For each count, the error bars represent one standard error from the mean replication time. The solid line represents a local regression curve ('loess' function in R) relating mean replication time to number of repeats. The P value is determined from the Pearson correlation between replication time and repeat count. **(a)** $(CA)_n$; **(b)** $(ACTG)_n$; **(c)** SR; and **(d)** $(CATA)_n$.

The finding that SSRs and SRs correlate with replication time raised the issue of whether multiple copies of a given repetitive sequence were necessary to establish this correlation or whether single copies were sufficient. The frequency of a given repeat in the 100-kb flanking region of a gene was plotted against replication time of that gene. These plots indicated a consistent trend of later replication time in the presence of a larger number of $(CA)_n$ repeats (Fig. 2a). This trend was also evident for $(ACTG)_n$ (Fig. 2b), and SR (Fig. 2c) repeats. By contrast, the presence of a larger number of $(CATA)_n$ repeats was associated with earlier replication time (Fig. 2d).

We also considered the role of length in the relationship between repeat count and replication time (Supplementary Table 5, www.karger.com/doi/10.1159/000089869). For the repeat types $(ACTG)_n$, SR, and $(CATA)_n$, we found that counting the number of large copies of the repeat (those whose

length is in the upper quartile of the length distribution) led to higher correlations with replication time than counting the number of smaller tracts (Supplementary Table 5B–D, www.karger.com/doi/10.1159/000089869). This suggests a threshold length beyond which the correlation increases. For $(CA)_n$ repeats, length did not appear to affect the relationship between count and replication time (Supplementary Table 5A, www.karger.com/doi/10.1159/000089869).

Evidence that the relationship of repetitive sequences to replication timing extends throughout the majority of the genome

To investigate how the significance of correlation (as measured by minus \log_{10} of the Pearson correlation P value) varied over distance, we considered increasingly large regions flanking the transcription start site of genes (Fig. 3). For all four repeats, the significance of the correlation and/or corre-

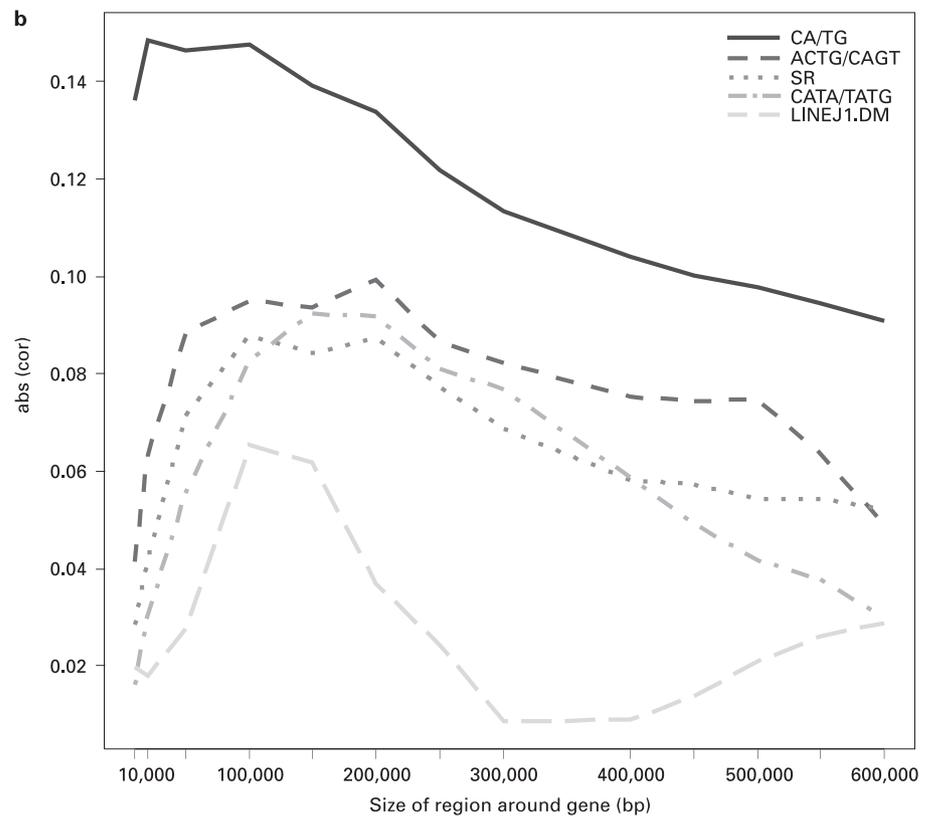
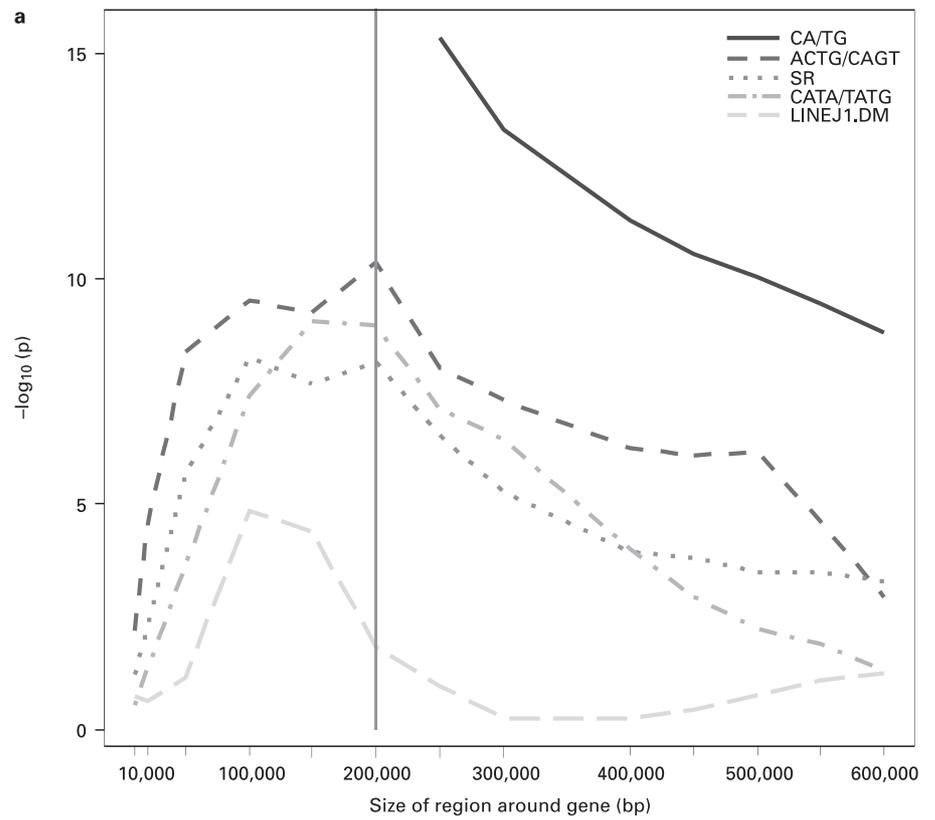


Fig. 3. (a) Change in the significance of correlation (as measured by minus \log_{10} of the Pearson correlation P value) when increasingly large regions flanking the transcription start site of genes are considered. By inspection, it is evident that the significance of the correlation drops off sharply for $(ACTG)_n$, $(CATA)_n$, and SR when the flanking region size exceeds 200 kb. **(b)** Change in correlation.

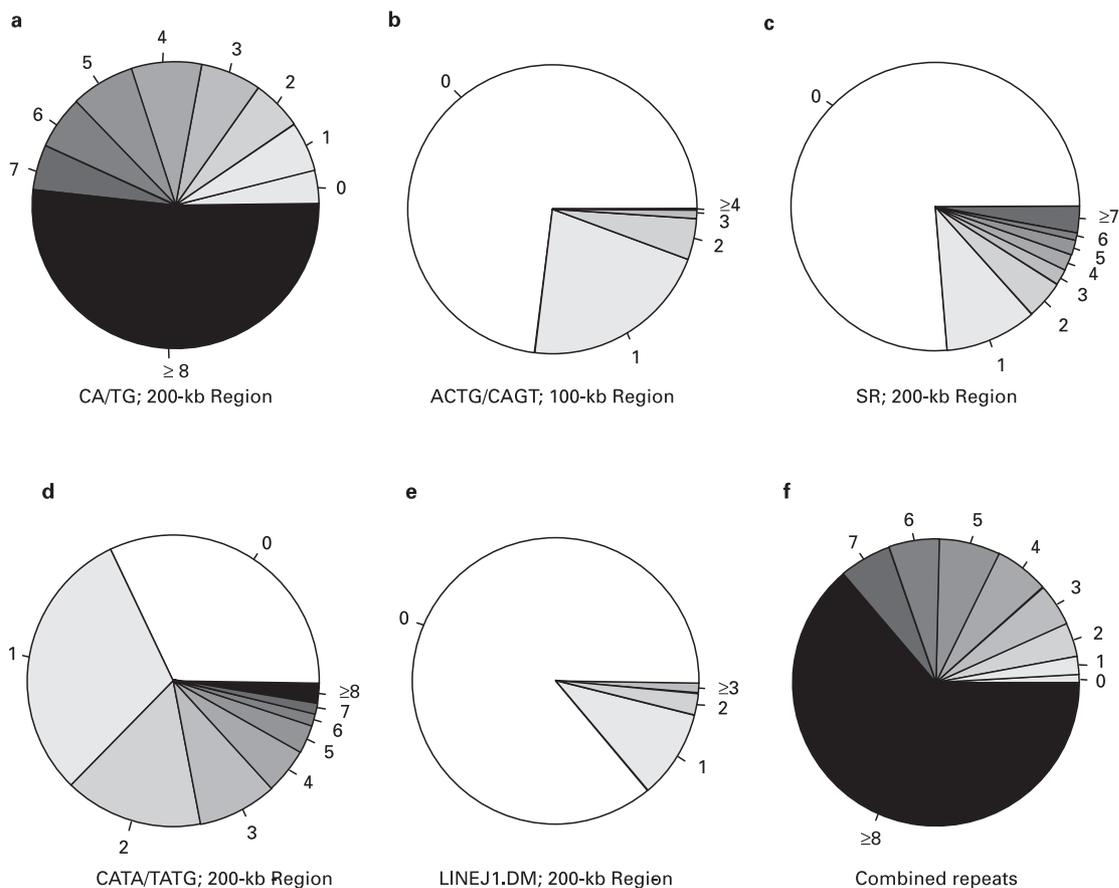


Fig. 4. Proportion of the euchromatic portion of *Drosophila* chromosome(s) within the range of correlation of $(CA)_n$, $(ACTG)_n$, $(CATA)_n$, SR, and LINEJ1.DM. For each repetitive sequence type the proportions of the genome in range of one, two, three, etc. copies of the repeat or repeat are shown.

lation coefficient showed a marked decline beyond a certain flanking region size: after 100 kb for $(CA)_n$ and LINEJ1_DM and after 200 kb for $(ACTG)_n$, $(CATA)_n$ and SR (Fig. 3a). The *P* value associated with the correlation to $(CA)_n$ was below the precision of the test for flanking regions up to size 250 kb, however a drop in the correlation coefficient was observed beyond 100 kb (Fig. 3b). We define a 'region of correlation' for a repeat based on these distances and propose that genes within the region of correlation may have their replication time influenced by the given repeat.

Given the strikingly large regions of correlation for the five repeats under consideration, we sought to determine the proportion of the *Drosophila* genome that fell within the regions of correlation of each of these repetitive sequences.

Our calculations indicated a 100-kb region of correlation for $(CA)_n$ and LINEJ1_DM repeats and a 200-kb region for $(ACTG)_n$, $(CATA)_n$, and SR. We used these regions to analyze the euchromatic portion of the *Drosophila* genome. We found that the proportion of the *Drosophila* genome within the region of correlation of one or more repeats was 85% for $(CA)_n$ (Fig. 4a), 46% for $(ACTG)_n$ (Fig. 4b), 67% for $(CATA)_n$ (Fig. 4d), 24% for SR (Fig. 4c), and 13% for LINEJ1_DM (Fig.

4e). More than 99% of the genome was within the region of correlation of at least one of these repetitive sequences (Fig. 4f).

Discussion

We find that the replication times of *Drosophila* genes are correlated to the repetitive sequence content in their flanking regions. Our evidence indicates that the presence of several simple sequence repeats, as well as LINEJ1_DM elements and the Satellite-Related sequences, correlates with gene replication time over remarkably large (>100 kb) distances extending from each repeat. The four repetitive sequences with the greatest correlation to replication timing, $(CA)_n$, $(CATA)_n$, $(ACTG)_n$, and SR, comprise less than 0.5% of the 'euchromatic' portion of the *Drosophila* genome sequence, yet their regions of correlation collectively cover more than 99% of this 'euchromatic genome'. Our data indicates that repeats account for 7% of replication time variation in the genome. Other factors that have been shown to correlate with replication time are transcription factor binding sites (Zappulla et al., 2002), differ-

ences among replication origins (Fox et al., 1993), gene expression (Schubeler et al., 2002; Alter and Golub, 2004), and chromatin structure (Ferguson and Fangman, 1992; Kitsberg et al., 1993; Vogelauer et al., 2002; Cohen et al., 2003).

SSRs, LINEJ1_DM elements, and SRs may influence replication time through the spread of heterochromatin from these elements. Heterochromatin has been shown to form at repetitive sequences (Dorer and Henikoff, 1994; Pal-Bhadra et al., 1997; Garrick et al., 1998; Selker, 1999) and to spread into adjacent regions (Muller, 1930; Hansen et al., 1997; Wakimoto, 1998). In a recent well-characterized example, heterochromatin was shown to spread ~10 kb from a single 1360 transposon in *Drosophila* and silence genes within this region (Sun et al., 2004). Replication time appears to reflect chromatin structure because the type of chromatin that encompasses an origin of replication determines the time in S phase at which that origin is activated (Ferguson and Fangman, 1992; Kitsberg et al., 1993; Vogelauer et al., 2002; Cohen et al., 2003), with most heterochromatin structures replicating later in S phase than euchromatin (Lima-de-Faria and Jaworska, 1968; Gilbert, 2002). A minority of heterochromatin structures are known to replicate early in S-phase: the centromeres of *S. pombe* (Kim et al., 2003) and the inactive X chromosome in the extraembryonic tissue of preimplantation mouse embryos (Sugawara et al., 1983). There is evidence that different heterochromatin structures are associated with different replication times (Chadwick and Willard, 2004) and the early-replicating heterochromatin are presumably distinct structures.

Our findings suggest that different repeat types may be associated with different types of heterochromatin. We propose three categories of repetitive sequences. Repeats from the first category, (CA)_n, (ACTG)_n and SR, were more abundant around later replicating genes and may be associated with chromatin structure(s) that confer later replication timing. A second category, represented by (CATA)_n, was more abundant around earlier replicating genes, suggesting that this class is associated with a heterochromatin structure that confers early replication timing. A third category of repeats that includes (TA)_n, (CAG)_n, (CAA)_n, ROO_I, DNAREP1_DM, (GA)_n, and (CATATA)_n has little or no effect on replication timing in *Drosophila* Kc cells, even though some of these repeats were as abundant or more abundant as class one repeats. There are at least three possible explanations for an absence of correlation with replication time. First, the heterochromatin structures associated with some types of repeats may be neutral with respect to replication timing. Second, heterochromatin may not spread far enough from some types of repeats to encompass origins and influence replication time. Third, category three repeats may not be associated with heterochromatin. We point out that category three repeats may be a heterogeneous group with different repeats not correlating with replication time for different reasons.

Why might different repeats be associated with different chromatin structures? One possibility is that some or all repeats foster the formation of heterochromatin, but different repeats foster the formation of different heterochromatin structures. Another possibility is that repressed chromatin may simply be more tolerant of certain repetitive sequences

compared to euchromatic regions. In this case, different types of repressed chromatin might display tolerance for different types of repeats. Note that the aforementioned two possibilities are not mutually exclusive.

There is evidence that different copies of the same repeat type can display different chromatin properties due to differences in chromosomal context. We would expect this variation to weaken the already indirect correlation between repeat type and replication time. This will likely cause us to underestimate the importance of repeats in influencing replication time. For example, the spread of heterochromatin from an SSR in *Drosophila* has been shown to be influenced by the proximity of this SSR to the centromere on the same chromosome (Henikoff et al., 1995). Another example in mammals indicates that context also influences the LINE transposon. An analysis of LINE-1 elements in DNA methyltransferase 3b-deficient cells revealed that the LINE-1 elements on the inactive X chromosome display chromatin properties that are distinct from LINE-1 elements located elsewhere (Hansen, 2003). These differences extend to replication time where the inactive X chromosome replicates later than the other chromosomes. The relationship between LINE-1 elements and replication time is therefore complicated by the LINE-1 elements on the inactive X chromosome replicating later than the LINES on other chromosomes.

If late replicating genes are encased in heterochromatin that has spread beyond the repeats, how do some late replicating genes still manage to be expressed? An explanation presents itself from reports that a subset of expressed genes actually alternate between 'on' and 'off' transcription states (Ross et al., 1994; Wijgerde et al., 1995; Milot et al., 1996; Kimura et al., 2002; Wilson et al., 2002; Osborne et al., 2004). These genes may be alternating between a euchromatic state in which they are expressed, and a heterochromatic state in which they are silenced. This would require heterochromatin to intermittently spread outward and encase the gene. There are several examples where the leading edge of heterochromatin has been shown to advance and retreat (Muller, 1930; Gottschling et al., 1990; Henikoff, 1996a, b; Wakimoto, 1998). According to this model, genes within reach of the intermittent spread of heterochromatin from a larger number of sources spend a greater proportion of their time being silenced by heterochromatin.

Whether the SSRs and SRs not associated with telomeres or centromeres serve a function is not known. The observation that SSRs display conservation among plant species points to a biological role (Yu et al., 2004). No roles for Satellite and Satellite-Related sequences are known, other than to direct the formation of centromeres. We propose that one purpose of SSRs and SRs is to impart repressive heterochromatic properties throughout the genome in order to minimize spurious transcription, i.e. to suppress transcription that is not the result of deliberate gene activation. According to this model, enhancers and LCRs are utilized by tissue-specific genes to force the repressive chromatin to be remodeled into an 'open' state that is readily accessible to transcription factors. A prediction of this model is that loss-of-function mutations in chromatin remodeling proteins would cause spurious transcription rates to rise. This prediction could be tested in future studies.

References

- Adams MD, Celniker SE, et al: The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195 (2000).
- Alter O, Golub GH: Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc Natl Acad Sci USA* 101:16577–16582 (2004).
- Bailey JA, Carrel L, et al: Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci USA* 97:6634–6639 (2000).
- Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300 (1995).
- Bernard P, Allshire R: Centromeres become unstuck without heterochromatin. *Trends Cell Biol* 12:419–424 (2002).
- Chadwick BP, Willard HF: Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome. *Proc Natl Acad Sci USA* 101:17450–17455 (2004).
- Chambers JM: Linear models, in Chambers SJM, Hastie T (eds): *Statistical Models*, pp 96–138 (Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, 1992).
- Chan SR, Blackburn EH: Telomeres and telomerase. *Phil Trans R Soc Lond B Biol Sci* 359:109–121 (2004).
- Cherbas P, Cherbas L, et al: Induction of acetylcholinesterase activity by beta-ecdysone in a *Drosophila* cell line. *Science* 197:275–277 (1977).
- Cohen SM, Brylawski BP, et al: Same origins of DNA replication function on the active and inactive human X chromosomes. *J Cell Biochem* 88:923–931 (2003).
- Davison AC, Hinkley DV: *Bootstrap Methods and Their Application* (Cambridge University Press, Cambridge 1997).
- Dorer DR, Henikoff S: Expansions of transgene repeats cause heterochromatin formation and gene silencing in *Drosophila*. *Cell* 77:993–1002 (1994).
- Ferguson BM, Fangman WL: A position effect on the time of replication origin activation in yeast. *Cell* 68:333–339 (1992).
- Fox CA, Loo S, et al: A transcriptional silencer as a specialized origin of replication that establishes functional domains of chromatin. *Cold Spring Harb Symp Quant Biol* 58:443–455 (1993).
- Garrick D, Fiering S, et al: Repeat-induced gene silencing in mammals. *Nat Genet* 18:56–59 (1998).
- Gilbert DM: Replication timing and transcriptional control: beyond cause and effect. *Curr Opin Cell Biol* 14:377–383 (2002).
- Gottschling DE, Aparicio OM, et al: Position effect at *S. cerevisiae* telomeres: reversible repression of Pol II transcription. *Cell* 63:751–762 (1990).
- Haaf T, Warburton PE, et al: Integration of human alpha-satellite DNA into simian chromosomes: centromere protein binding and disruption of normal chromosome segregation. *Cell* 70:681–696 (1992).
- Hansen RS: X inactivation-specific methylation of LINE-1 elements by DNMT3B: implications for the Lyon repeat hypothesis. *Hum Mol Genet* 12:2559–2567 (2003).
- Hansen RS, Canfield TK, et al: Association of fragile X syndrome with delayed replication of the *FMR1* gene. *Cell* 73:1403–1409 (1993).
- Hansen RS, Canfield TK, et al: A variable domain of delayed replication in FRAXA fragile X chromosomes: X inactivation-like spread of late replication. *Proc Natl Acad Sci USA* 94:4587–4592 (1997).
- Harrington JJ, Van Bokkelen G, et al: Formation of de novo centromeres and construction of first-generation human artificial microchromosomes. *Nat Genet* 15:345–355 (1997).
- Henikoff S: Dosage-dependent modification of position-effect variegation in *Drosophila*. *Bioessays* 18:401–409 (1996a).
- Henikoff S: Position-effect variegation in *Drosophila*: recent progress, in Russo VEA, Martienssen RA, Riggs AD (eds): *Epigenetic Mechanisms of Gene Regulation*, pp 319–334 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor 1996b).
- Henikoff S, Jackson JM, et al: Distance and pairing effects on the brown Dominant heterochromatic element in *Drosophila*. *Genetics* 140:1007–1017 (1995).
- Kim DD, Dubey DD, et al: Early-replicating heterochromatin. *Genes Dev* 17:330–335 (2003).
- Kim GD, Choi YH, et al: Sensing of ionizing radiation-induced DNA damage by ATM through interaction with histone deacetylase. *J Biol Chem* 274:31127–31130 (1999).
- Kimura H, Sugaya K, et al: The transcription cycle of RNA polymerase II in living cells. *J Cell Biol* 159:777–782 (2002).
- Kitsberg D, Selig S, et al: Replication structure of the human beta-globin gene domain. *Nature* 366:588–590 (1993).
- Lica LM, Narayanswami S, et al: Mouse satellite DNA centromere structure and sister chromatid pairing. *J Cell Biol* 103:1145–1151 (1986).
- Lima-de-Faria A, Jaworska H: Late DNA synthesis in heterochromatin. *Nature* 217:138–142 (1968).
- Lyon MF: X-chromosome inactivation: a repeat hypothesis. *Cytogenet Cell Genet* 80:133–137 (1998).
- Marahrens Y: X-inactivation by chromosomal pairing events. *Genes Dev* 13:2624–2632 (1999).
- Milot E, Strouboulis J, et al: Heterochromatin effects on the frequency and duration of LCR-mediated gene transcription. *Cell* 87:105–114 (1996).
- Muller H: Types of visible variations induced by X-rays in *Drosophila*. *J Genet* 22:299–334 (1930).
- Osborne CS, Chakalova L, et al: Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet* 36:1065–1071 (2004).
- Pal-Bhadra M, Bhadra U, et al: Cosuppression in *Drosophila*: gene silencing of alcohol dehydrogenase by white-Adh transgenes is Polycomb dependent. *Cell* 90:479–490 (1997).
- Platero JS, Ahmad K, et al: A distal heterochromatic block displays centromeric activity when detached from a natural centromere. *Mol Cell* 4:995–1004 (1999).
- Rivier DH, Rine J: An origin of DNA replication and a transcription silencer require a common element. *Science* 256:659–663 (1992).
- Ross IL, Browne CM, et al: Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. *Immunol Cell Biol* 72:177–185 (1994).
- Rubin GM, Hong L, et al: A *Drosophila* complementary DNA resource. *Science* 287:2222–2224 (2000).
- Schmidt DR, Schreiber SL: Molecular association between ATR and two components of the nucleosome remodeling and deacetylating complex HDAC2 and CHD4. *Biochemistry* 38:14711–14717 (1999).
- Schubeler D, Scalzo D, et al: Genome-wide DNA replication profile for *Drosophila melanogaster*, a link between transcription and replication timing. *Nat Genet* 32:438–442 (2002).
- Selker EU: Gene silencing: repeats that count. *Cell* 97:157–160 (1999).
- Shechter D, Costanzo V, et al: ATR and ATM regulate the timing of DNA replication origin firing. *Nat Cell Biol* 6:648–655 (2004).
- Smit A, Hubley R, et al: RepeatMasker Open-3.0 (1996–2004).
- Storey JD: A direct approach to false discovery rates. *J R Stat Soc B* 64:479–498 (2002).
- Sugawara O, Takagi N, et al: Allocyclic early replicating X chromosome in mice: genetic inactivity and shift into a late replicator in early embryogenesis. *Chromosoma* 88:133–138 (1983).
- Sun FL, Haynes K, et al: *cis*-Acting determinants of heterochromatin formation on *Drosophila melanogaster* chromosome four. *Mol Cell Biol* 24:8210–8220 (2004).
- Vogelauer M, Rubbi L, et al: Histone acetylation regulates the time of replication origin firing. *Mol Cell* 10:1223–1233 (2002).
- Wakimoto BT: Beyond the nucleosome: epigenetic aspects of position-effect variegation in *Drosophila*. *Cell* 93:321–4 (1998).
- Webb T: Delayed replication of Xq27 in individuals with the fragile X syndrome. *Am J Med Genet* 43:1057–1062 (1992).
- Wevrick R, Willard HF: Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. *Proc Natl Acad Sci USA* 86:9394–9398 (1989).
- Wijgerde M, Grosveld F, et al: Transcription complex stability and chromatin dynamics in vivo. *Nature* 377:209–213 (1995).
- Wilson AJ, Velcich A, et al: Novel detection and differential utilization of a c-myc transcriptional block in colon cancer chemoprevention. *Cancer Res* 62:6006–6010 (2002).
- Yu JK, La Rota M, et al: EST derived SSR markers for comparative mapping in wheat and rice. *Mol Genet Genomics* 271:742–751 (2004).
- Zappulla DC, Sternglanz R, et al: Control of replication timing by a transcriptional silencer. *Curr Biol* 12:869–875 (2002).