Chromosoma (2010) 119:589–600
DOI 10.1007/s00412-010-0280-y

RESEARCH ARTICLE

Gene density profile reveals the marking of late replicated domains in the *Drosophila melanogaster* genome

Stepan N. Belyakin · Vladimir N. Babenko · Daniil A. Maksimov · Viktor V. Shloma · Evgeny Z. Kvon · Elena S. Belyaeva · Igor F. Zhimulev

Received: 4 March 2010 / Revised: 4 June 2010 / Accepted: 7 June 2010 / Published online: 3 July 2010

Abstract Regulation of replication timing has been a focus of many studies. It has been shown that numerous chromosomal regions switch their replication timing on cell differentiation in *Drosophila* and mice. However, it is not clear which features of these regions are essential for such regulation. In this study, we examined the organization of late underreplicated regions (URs) of the *Drosophila melanogaster* genome. When compared with their flanks, these regions showed decreased gene density. A detailed view revealed that these regions originate from unusual combination of short genes and long intergenic spacers. Furthermore, gene expression study showed that this pattern is mostly contributed by short testis-specific genes abundant in the URs. Based on these observations, we developed a genome scanning algorithm and identified 110 regions possessing similar gene density and transcriptional profiles. According to the published data, replication of these regions has been significantly shifted towards late S-phase in two *Drosophila* cell lines and in polytene chromosomes. Our results suggest that genomic organization of the underreplicated areas of *Drosophila* polytene chromosomes may be associated with the regulation of their replication timing.

Communicated by B. Calvi

Electronic supplementary material The online version of this article (doi:10.1007/s00412-010-0280-y) contains supplementary material, which is available to authorized users.

S. N. Belyakin · D. A. Maksimov · V. V. Shloma · E. S. Belyaeva · I. F. Zhimulev (✉)
Department of Molecular and Cellular Biology,
Institute of Chemical Biology and Fundamental Medicine SD RAS,
Lavrentyev ave, 10,
Novosibirsk 630090, Russia
e-mail: zhimulev@bionet.nsc.ru

V. N. Babenko · E. Z. Kvon
Institute of Cytology and Genetics SD RAS,
Novosibirsk, Russia

Present Address:
E. Z. Kvon
Research Institute of Molecular Pathology,
Vienna, Austria

Introduction

DNA replication prior to cell division is strictly regulated in eukaryotes. According to Zhimulev (1998), this signifies that some genomic regions are replicated early, while others start replication close to the end of S-phase of the cell cycle. Differential replication timing allows formation of early and late replication domains. As observed in many cell types, genes from late replicated regions are usually silent in these cells, while those from early replicated regions are more active (Hiratani and Gilbert 2009; Hiratani et al. 2008; Kalisch and Hagele 1976; MacAlpine et al. 2004; Schuebeler et al. 2002; Schwaiger et al. 2009).

Replication program is accurately transmitted in succession of cell generations (Berezney et al. 2000; Ma et al. 1998). However, it has been shown that replication timing depends on cell type-specific regulation. The well-known example of such a regulation in mammals is the cluster of β-globin genes, which replicates early in erythroblasts, but demonstrates late replication in other cells (Cimbora et al. 2000; Simon et al. 2001). A recent study on mice demonstrated that replication timing may change on differentiation of cells, suggesting its involvement in the developmental regulation of large chromosomal domains comprising many genes (Hiratani and Gilbert 2009; Hiratani et al. 2008).

Recently, the plasticity of replication program in two cell cultures of *Drosophila melanogaster* was demonstrated (Schwaiger et al. 2009). It was found that initiation of

replication correlates with Lys 16 acetylation of histone H4, rather than with the gene activity. Accordingly, local variation of this chromatin mark was found to lead to the shift in the replication timing. Reported differences affected about 20% of the autosomal part of the genome. However, details of the regulation allowing particular genomic regions to switch replication timing essentially remain uncovered, and are apparently fulfilled through the epigenetic status change in different cell types.

The replication timing of *Drosophila* salivary gland polytene chromosomes was visualized using pulse incorporation of labeled nucleotides (Kalisch and Hagele 1976; Zhimulev and Belyaeva 2003). Replication was found to start in puffs and other decondensed regions, and subsequently involve new regions, and the whole chromosome was labeled (stage of continuous labeling or early replication). At the following stage (discontinuous replication), early regions had already completed replication, while the replication was still active in the late regions. As a result, the chromosomes demonstrated discrete pattern of labeling. At this stage, about 240 reproducible regions could be observed (Zhimulev et al. 2003). These regions demonstrated a delay in the completion of replication. As the S-phase progressed, fewer regions remained involved in replication, and finally, only about 60 regions completed replication at the very end of the S-phase (Zhimulev 1998; Zhimulev and Belyaeva 2003).

Regions that replicated last in the *Drosophila* polytene chromosomes often failed to complete the process of replication in the recurring endocycles, and remained underreplicated to some extent (Moshkin et al. 2001). These regions can be observed through microscope, because they often show characteristic morphology: in the wild-type strains, about 60 regions of the 240 late replicated regions demonstrated chromosomal breaks (Zhimulev et al. 2003). Underreplication of these domains was suppressed in the *Suppressor of UnderReplication* (*SuUR*) mutants, while two transgenic extra copies of the wild-type *SuUR* allele increased their replication delay, and consequently, about 150 late replicated regions manifested chromosomal breaks (Belyaeva et al. 1998; Zhimulev et al. 2003).

In our previous study, we identified 52 underreplicated regions (URs) in *D. melanogaster* salivary gland polytene chromosomes using microarrays (Belyakin et al. 2005). These regions contained more than 1,000 genes, and most of them replicated late in the cultured Kc cells (Schuebeler et al. 2002), resided in the late replicated chromosome bands of the polytene chromosomes (Zhimulev et al. 2003), and demonstrated binding to the SuUR protein and genetic repression factors, such as Polycomb and HP1 (Pindyurin et al. 2007). These regions represent the most profoundly underreplicated part of approximately 240 late replicated regions of polytene chromosomes (Zhimulev et al. 2003). It was also shown that these late replicated regions are enriched with testis-specific genes (Belyakin et al. 2005) that are known to cluster in the genome (Boutanaev et al. 2002; Shevelyov et al. 2009).

In this study, we examined the organization of URs. Importantly, it must be noted that replication of all these regions is affected by the *SuUR* gene, suggesting a common regulatory mechanism. Thus, our analysis was aimed to determine the possible peculiarities that may be related to the regulation of their replication timing. The gene-density pattern in URs was studied and compared with that in the flanking areas. The analysis revealed characteristic peculiarities in the length of the genes and intergenic regions in the URs. Furthermore, expression profiling of 93% *D. melanogaster* genes during development was carried out to examine transcriptional specificity of the known URs. A strong transcriptional bias in the group of very short underreplicated genes was observed. Subsequently, we tested the use of the discovered characteristics to predict the late replicated regions in the *D. melanogaster* genome.

Materials and methods

Gene sets compilation

FlyBase 5.1 genome database (ftp://ftp.flybase.net/releases/FB2007_03/) was used in our analysis. A sample of 51 previously described (Belyakin et al. 2005) URs was employed for the assessment of UR features. The longest annotated transcript in a locus was considered as a reference gene. All the sets compiled were nonoverlapping. The intergenic length was calculated between the neighbor genes independent of their orientation.

Method of URs prediction de novo

1. Discriminating gene density skew in the putative URs prediction

The search routine for the putative URs was based on the direct chromosome scan. We used overlapping windows of floating length (150–500 kb) in the course of the scan with 20-kb subsequent shift. The exact putative UR length was ascertained by local optimum length search procedure. The number of gene starts within each window was compared with that in the flanking regions, which were half the size of that window from both the sides.

The target functional was defined as the P value of the one-sided binomial test:

$$F(x;n,p) = \Pr(X \le x) = \sum_{j=0}^{\mathrm{Floor}(x)} \binom{n}{j} p^j (1-p)^{n-j}$$

where, p is the a priori probability for a gene to fall into either the UR or flanks (p=0.54, based on the sample sizes), n is the total number of genes in the UR and flanks, and X is the gene number in the flanks region.

2. Discriminating intergenic length values in the putative URs prediction

The one-sided t test (unequal sample sizes, equal variance) was used to compare the intergenic length values in the predicted regions and their flanks:

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

where

$$S_{X_1X_2} = \sqrt{\frac{n_1 S_{x1}^2 + n_2 S_{x2}^2}{n_1 + n_2}}$$

where, X_1 is a log-transformed intergenic length values array for the UR and n_1 is the array size, while X_2 is the log-transformed intergenic length values array for the flanks region with n_2 size. In the target prediction function, 95 percentile for t value calculated in this way was used (see further on).

3. Discriminating testis/ovary expression ratio in putative URs prediction

The one-sided t test (1) was used to detect the expression bias: the arrays were the log-transformed testis/ovary ratios in the putative URs and their flanks calculated for genes shorter than 2 kb.

We used the linear combination of the three above-mentioned tests as the discrimination function:

$$\text{Score} = P \times 29 + I \times 30 + E \times 18,$$

where P, I, and E are the power values (1-P value) of the gene density discrimination (first), testis/ovary expression ratio discrimination (second), and accumulative intergenic length discrimination (third) tests, respectively. The coefficients were ascertained empirically from the Receiver Operating Characteristic (ROC) curves presented in Online Resource 1. As the three characteristics that we employed for the discrimination were not strictly dependent on a total genomic scale, we assessed each of the tests independently for the optimum score.

Subsequently, the quality of our predictions was tested. The sample of 51 URs was used as a positive sample, while their flanks were used as a negative sample. Sensitivity and specificity were calculated according to the following formulas:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{number of false negatives}}$$

$$\text{Sensitivity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives} + \text{number of false positives}}$$

The region was considered “True Positive” if it overlapped with the experimentally determined UR by more than 50% of the length of that UR. On the other hand, the region was considered “False Positive” if it overlapped with the flank of known UR by more than 50% of the length of that flank.

Developmental profiling of gene expression

Biological samples from 10 developmental stages were collected: 0–1-h embryos, 5–6-h embryos, 11–12-h embryos, first instar larvae, second instar larvae, third instar larvae, 0–1-h prepupae, 13–14-h pupae, 3-day-old adult males, and 3-day-old females. The embryos were collected as follows: flies were allowed to lay eggs for 45 min on the agar/apple juice medium with some yeast. The embryos were either collected immediately and frozen at −80°C, or allowed to develop at 25°C to the stage of interest and then collected and frozen. The first instar larvae were collected right after hatching, while the second and third instar larvae were collected after the corresponding molts, 24 and 48 h after hatching, respectively. The 0–1-h prepupae were picked each hour from the tube walls and either frozen at −80°C immediately or allowed to develop at 25°C for 13 h to obtain 13–14-h pupae. Adult males and females were collected twice a day, kept for 3 days at 25°C, and then separated and frozen at −80°C. The females collected in this way were mated. Total RNA isolation from the frozen samples was performed with TRIzol reagent (Invitrogen, #155-96-011). The RNAs from 10 stages were isolated in parallel and diluted to the same concentration (4 μg/μl). The reference sample was prepared by mixing the same amount of total RNA from each stage. Subsequently, 25 μg of the sample and reference RNAs were labeled with either Cy-3 or Cy-5 dUTPs (Amersham, #PA53022 and #PA55022) using SuperScript Direct cDNA Labeling System (Invitrogen, #L1015-02), according to the manufacturer’s protocol. The labeled cDNAs were purified using QIAquick polymerase chain reaction (PCR) purification kit (QIAGEN, #28104) and combined in one tube, and 20 μg of sonicated salmon sperm DNA was added to it. The DNA mixture was dried in the stream of nitrogen and dissolved in 60 μl of hybridization buffer (50% deionized formamide, 6× SSC, 0.5% SDS, and 5× Denhardt’s).

Microarray slides (DGRC-2 transcriptome slides received through Drosophila Genome Resource Center, GEO platform GPL3880) were prehybridized in 6× SSC, 0.5% SDS, and 1% BSA at 42°C water bath for 1 h, rinsed with deionized water, and dried in bucket rotor centrifuge in 50-ml Falcon tubes (5 min, 1,000×*g*). Labeled cDNAs were denatured at 100°C heat block for 2 min, cooled at room temperature, pipetted onto the surface of the microarray, and covered with 25×60 mm LifterSlip (Erie Scientific Company). The assembled microarray was hybridized for 12–14 h in the VersArray hybridization chamber (Bio-Rad) at 42°C water bath. After hybridization, the arrays were placed into the beaker with 1× SSC and 0.1% SDS to remove the LifterSlips. Subsequently, three washes were performed in the same solution (10, 5, and 5 min), and the microarrays were rinsed with 1× SSC (30 s), 0.1× SSC (30 s), and dried in centrifuge as described earlier. The dry microarrays were scanned at 10-μm resolution on ScanArray Lite instrument (Perkin Elmer). The scans were analyzed using TIGR Spotfinder program (www.tm4.org). The data were lowess normalized and filtered using MIDAS software (www.tm4.org). The final data sets were compiled in Microsoft Excel. All the data are available at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/), with the accession number: GSE16531.

Real-time PCR assay

The underreplication level of target regions was determined by real-time PCR using SYBR Green PCR Master Mix (Syntol, Russia) according to the manufacturer's instructions. The list of primers is provided in the Online Resource 2. Real-time PCR was performed using the ABI Prism® 7900 Sequence Detection System using standard conditions (95°C for 5 min, 95°C for 15 s, 60°C for 20 s, and 72°C for 30 s, with the latter three steps repeated 40 times). The final concentration of each primer was 500 nM. The primers were designed using the Primer Express 2.0 software (Applied Biosystems). All the data were analyzed with SDS software v2.3 (Applied Biosystems). To ensure that only a single product was amplified in each reaction, dissociation curve analysis was performed. The curves with only one peak were considered in the analysis.

The relative UR ratio of the target region in the salivary glands polytene chromosomes was measured using the standard curve method. Representation of the target DNA sequence in each unknown sample was determined by comparing the Ct value of each sample with a standard curve generated from four-point serial dilution (dilution factor×6) of the genomic DNA from the diploid tissues (larval brain). All PCR reactions were performed in triplicate, and the results were obtained from at least two independent biological samples.

Results

Gene density correlates with late replication in URs

In our previous study, we used a microarray approach to identify 52 URs out of 240 cytologically observed late replicated regions (Belyakin et al. 2005) (the tracks in UCSC genome browser BED format are available as Online Resource 3 for the whole genome polytenization profile, and as Online Resource 4 for URs positions). These 52 regions were found to replicate last in the S-phase of the cell cycle in the salivary glands and remain underreplicated (Zhimulev et al. 2003). Other late replicated regions were either fully polytenized or their underreplication was beyond the sensitivity of our method. Here, we tried to find common characteristics of the 52 previously described regions and use them to determine similar regions in the *D. melanogaster* genome. The highly atypical region 39E, containing the repeated histone genes cluster, was removed from our analysis, and we further analyzed 51 URs. The URs were routinely compared with their flanking regions. The flanks were defined as the lengths of chromosome adjacent to each UR, spanning from both the sides for half the number of the genes present in the corresponding UR. In a few cases, the URs were located close to each other and shared one of the flanks; therefore, the total flanking genes (n=1,059) was somewhat lower than the number of underreplicated genes (n=1,164). The gene lists are available as Online Resource 5 for URs and as Online Resource 6 for flanks.

Careful examination of the gene densities within the vicinity of the URs revealed an intriguing pattern: the gene density was low in the URs, whereas the flanking areas were enriched with genes. Indeed, simple calculation of the gene densities demonstrated a high difference: in the URs, we observed an average of 77.3 genes per megabase DNA, whereas in the flanks, we found 187.8 genes per megabase DNA, as calculated for 51 URs and 101 flanks (Mann-Whitney U test, $P = 4.9 \times 10^{-10}$, Fig. 1a).

To check whether the low gene density is a common attribute of all late replicated regions in the genome, we performed a coarse estimate of gene density in the cytologically defined late replicated regions. Therefore, we used an annotation of cytological bands presented in UCSC Genome Browser for the *Drosophila* Genome Release 5.12. According to our previous observations of replication timing in polytene chromosomes (Zhimulev et al. 2003), we determined the gene density in those bands that replicated early and those that replicated late (Fig. 1b). Despite the discrepancies in cytological annotation, the general tendency remained in the cytologically detected late replicated regions. Indeed, Mann-Whitney U test showed highly significant decrease in the gene density in the late

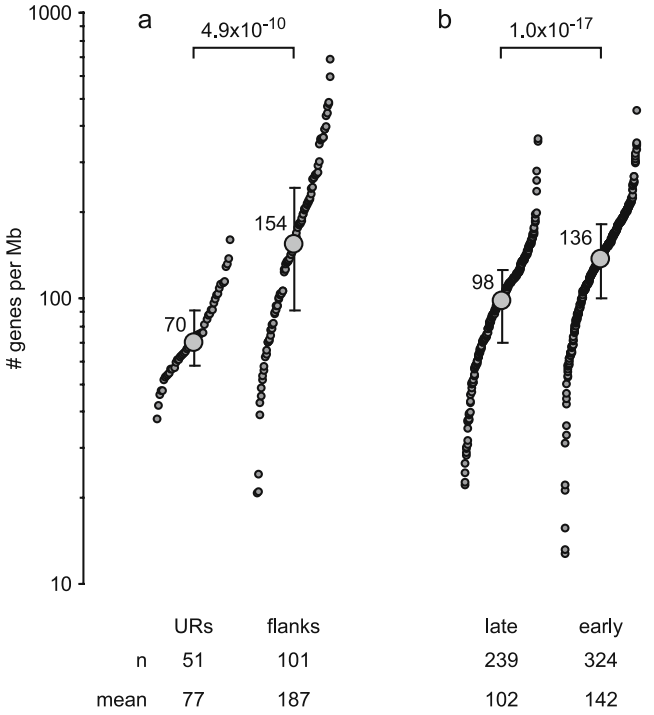

Fig. 1 Gene density and replication timing. **a** Number of genes per megabase DNA in 51 URs and 101 flanks. The median position is shown with increased circles. Error bars represent the data between 25 and 75 percentile (50% of data points). Mann-Whitney probability is presented. **b** The same analysis was preformed for the cytologically defined late and early replicated areas in polytene chromosomes (Zhimulev et al. 2003). Under the figures: *n* number of regions on each graph, *mean* average gene density

replicated regions when compared with the early replicated regions ($P = 1.0 \times 10^{-17}$) (Fig. 1b). This difference remained significant even after we removed those late replicated bands that overlapped with the control set of 51 URs ($P = 1.6 \times 10^{-12}$).

To investigate the nature of low gene density in the URs, we performed an analysis of the distribution of gene lengths inside the URs and in their flanks. The maximal transcribed region length from the very first transcription start site to the end of the farthest known transcript was used for each gene. Surprisingly, we found that URs were enriched in short genes of less than 2 kb (667 vs. 477, binomial test, α=0.52, $P = 3.3 \times 10^{-5}$), with the highest significance for those shorter than 900 bp (291 in URs, 143 in flanks, 25% and 14% of the total number of genes, binomial test, α=0.52, $P = 3.8 \times 10^{-10}$, Fig. 2a).

This result suggests that the strong reduction in the gene density in the URs should be owing to the longer intergenic regions. The analysis of the intergenic regions within the 51 URs and their flanks showed a striking general reduction in the sizes of the intergenic regions in the flanks. Indeed, the portion of intergenic regions longer than 8 kb was 2.7 times lower in the flanks than in the URs (12% in the flanks vs. 33% in the URs, binomial test, α=0.54, $P = 2.0 \times 10^{-21}$, Fig. 2b). Moreover, the portion of short intergenic regions less than 1 kb was 1.8 times higher in the flanks (32% in the URs vs. 58% in the flanks, binomial test, α=0.54, $P = 2.4 \times 10^{-16}$).

Intergenic regions are usually presumed to be regulatory to nearby genes. To check whether there is any preference for long intergenic regions in the upstream or downstream areas of the genes, we examined the upstream against downstream intergenic region lengths for the genes from the URs and flanks (Fig. 2c, d). All the nested genes were removed from the analysis and the longest transcribed areas were used to determine the span of each gene. Surprisingly, the genes in the URs tended to be surrounded with long intergenic spacers from both the sides—45% of them neighbored at least 1 kb of upstream and downstream noncoding sequences; only 21% of the UR genes had less than 1-kb spacers to their sides (Fig. 2c). The distribution was opposite in the flanks and the whole genome, where 46% of the genes were surrounded by short (<1 kb) and 22% of the genes were surrounded by the long (>1 kb) intergenic regions (Fig. 2d). These results indicate that the observed pattern of gene density in the URs comes from unusual combination of short genes surrounded by long intergenic spacers.

In differentiating mice cells, the replication timing was adjusted to Cytosine and Guanine (CG) content profile (Hiratani et al. 2008). We checked if such correlation existed in the URs. Indeed, the CG content in the URs was found to be systematically lower than in the whole genome (41.3% vs. 42.5%). Interestingly, the CG content was also different between the cytologically defined late and early replicated regions (41.9% vs. 42.6%). However, these differences were statistically insignificant.

Sex-biased transcriptional specificity of URs

Previously (Belyakin et al. 2005), we showed that URs are enriched with genes showing male-specific expression pattern during fly development: they were downregulated in the embryos and larvae, and gradually activated through metamorphosis in the males, but not in females. This pattern disappeared in *tudor* mutants, where the germline formation was arrested (Arbeitman et al. 2002). However, this was observed on the limited gene set encompassing about 4,000 genes, and only about 200 of them were found in the URs (Belyakin et al. 2005). To obtain information about the other genes in the URs, we performed our own analysis of gene expression during *Drosophila* development. We used printed oligonucleotide arrays (DGRC-2 Oligonucleotide Transcriptome Microarrays) representing about 93% of *Drosophila* genes. The relative expression of the genes was studied at 10 time points along the fly

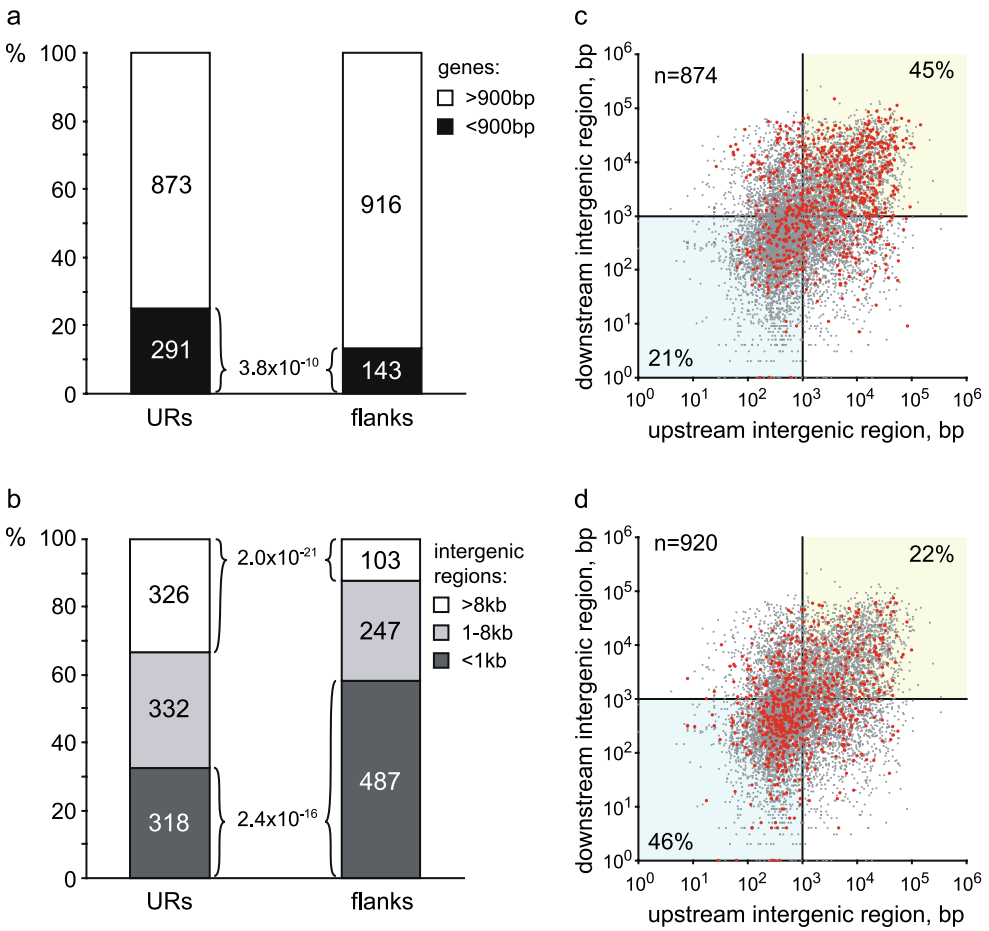

Fig. 2 URs tend to accumulate extremely short genes in combination with long intergenic regions when compared with their flanks. **a** The portion of genes shorter than 900 bp is 1.8 times higher in the URs than in the flanking regions. Binomial test was used to estimate the significance. Gene numbers are different in the URs and flanks owing to the fact that some regions share their flanks (shown in Fig. 4 for the regions 36C and 36E). **b** Intergenic regions shorter than 1 kb are underrepresented in the URs, while those longer than 8 kb are overrepresented. Binomial probability is presented. **c** Long intergenic regions show no preference to be located upstream or downstream of the genes in the URs (*red dots*). The distribution for all genes is shown with *gray dots*; 45% of the genes in the URs were surrounded with intergenic regions longer than 1 kb and only 21% of the genes had both the intergenic regions shorter than 1 kb. **d** The same analysis was performed for flanking regions. The tendency for the genes with either long or short intergenic spacers was opposite to the URs, and corresponded to the whole genome distribution

development: 0–1-h embryos, 5–6-h embryos, 11–12-h embryos, first instar larvae 0–1-h after hatching, second instar larvae after molt, third instar larvae after molt, 0–1-h prepupae, 13–14-h pupae, adult males, and adult females. The experimental design has been described in detail in the "Materials and methods" section. According to our data on the differences in the gene lengths between the URs and their flanks, we inspected the expression patterns in two groups: genes shorter than 2 kb and genes longer than 2 kb. Comparison of the averaged expression profiles revealed that enrichment of the URs with male-specific genes, upregulated during metamorphosis in adult males, but not in females, comes from the set of short genes (*t* test, $P = 8.8 \times 10^{-15}$, Fig. 3a). However, short genes in the flanks did not show this specificity. It was previously demonstrated that the observed pattern corresponds to the gene activity during male gonads development (Arbeitman et al. 2002). This was confirmed using the recent whole genome gene expression study of 14 adult fly organs (Chintapalli et al. 2007). The comparison of the averaged expression profiles of the short genes in the URs and flanks revealed a highly pronounced testis-specific pattern in the URs (Fig. 3b). In contrast, short genes from the flanks did not show any specificity patterns.

However, the averaged profiles could be strongly biased if a small group of highly regulated genes is present in the data set. To prove that the observed testis-specific pattern is not caused by few genes, but represents an important general characteristic of the URs, we distinguished testis- and ovary-specific genes using the adult fly expression data set (Chintapalli et al. 2007). Testis-specific genes were determined as those showing upregulation in the testis and

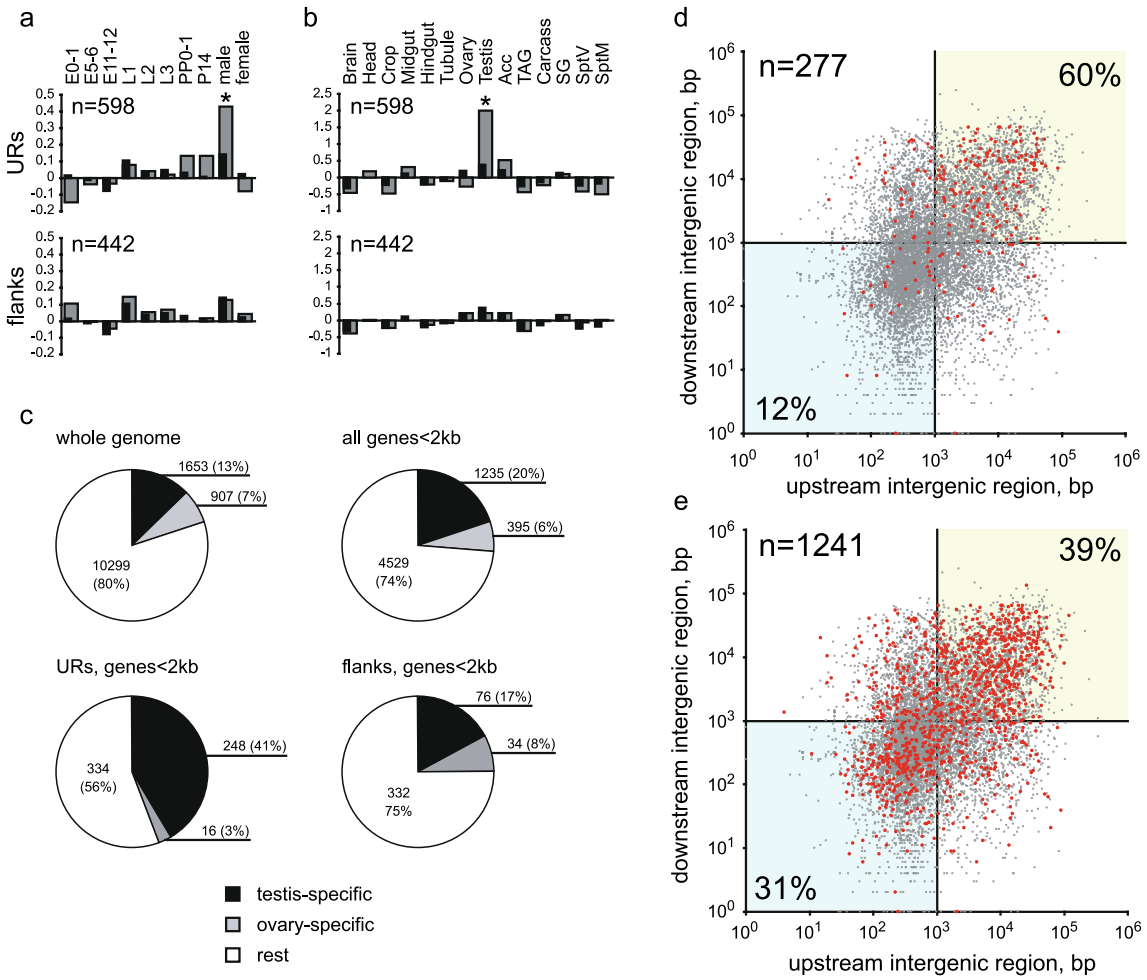

Fig. 3 Transcriptional specificity of the URs. **a** During *Drosophila* development, genes from URs (*gray bars*) shorter than 2 kb demonstrated a significant male-specific expression (*asterisk*) pattern when compared with the whole data set (*black bars*; *t* test, $P = 8.8 \times 10^{-15}$). The averaged expression of genes less than 2 kb from the URs, their flanks, and the total in the genome are shown (ordinates are averaged log2-transformed expression data). Male-specific pattern is strongly reduced in the flanks and shows no specificity. **b** Averaged expression of same genes in the organs of adult flies, according to FlyAtlas (Chintapalli et al. 2007). URs are enriched with genes expressed in testis (asterisk, *t* test, $P = 1.3 \times 10^{-31}$). This pattern is not present in the flanks. **c** Portions of testis- and ovary-specific genes in the whole *D. melanogaster* genome, among all the genes less than 2 kb and genes less than 2 kb in the URs and flanks. Testis-specific genes (*black*) are strongly overrepresented (41%), while ovary-specific genes (*gray*) are underrepresented (2%) in the URs. In flanks, the portions are close to those among all the short genes. **d** 60% of the testis-specific genes in the URs are surrounded by intergenic regions longer than 1 kb (*red dots*). All nonoverlapping *Drosophila* genes are shown with gray dots. **e** The length of the intergenic regions upstream and downstream of all nonoverlapping testis-specific genes (*red dots*): 39% of them are surrounded by long spacers (>1 kb), and substantial part (31%) has both intergenic regions shorter than 1 kb. On all panels: *n* number of genes in each category

downregulation in all the other organs. The same criterion was used to choose ovary-specific genes. As a result, we distinguished 1,653 testis-specific genes, a number comparable with those previously reported (Boutanaev et al. 2002; Shevelyov et al. 2009). The total number of ovary-specific genes was 907 (Fig. 3c). The lists of testis- and ovary-specific genes that we used are presented in Online Resources 7 and 8, respectively. Gene length analysis showed that the majority of testis-specific genes (1,235 out of 1,653; 75%) are shorter than 2 kb (Fig. 3c). However, the portions of short (<2 kb) testis-specific genes in the URs and their flanks were drastically different: 41% in the URs vs. 17% in the flanks. Moreover, the portion of short ovary-specific genes was found to be strongly reduced in the URs: 3% vs. 8% in the flanks. These differences in both the classes were highly significant and specific for the URs (χ^2 test, $P = 3.0 \times 10^{-17}$).

Genes longer than 2 kb showed upregulation in the ovaries and downregulation in the testis. The same observation was noted in the flanks, but not in the URs, where some testis-specific pattern was still distinguishable (Online Resource 9).

These observations suggest that URs are specialized domains of the *D. melanogaster* genome, which contain testis-specific genes and are purposed for their proper regulation. However, the question of what part of the testis-specific genes has similar genomic features and may be found in other late replicated domains needs to be addressed. To answer this question, we first examined the length of the upstream against downstream intergenic spacers of the testis-specific genes from the URs. As expected, most of the testis-specific genes (166; 60%) were surrounded by intergenic spacers longer than 1 kb (Fig. 3d). Furthermore, only 33 genes (12%) in this group had both spacers shorter than 1 kb. When we performed the same analysis for the whole testis-specific gene set, we observed that 482 (39%) of them had both intergenic regions longer than 1 kb. However, there was a substantial portion (382; 31%) that was flanked with short spacers of less than 1 kb (Fig. 3e). This suggests that the URs encompass a specific subpopulation of testis-specific genes that are surrounded by long intergenic regions (χ^2 test, $P = 4.6 \times 10^{-12}$). As the total number of such genes was 482, after subtraction of 166 genes from the known URs, we obtained an estimation of about 316 testis-specific genes that might be organized into similar domains in the *D. melanogaster* genome.

Search for putative late replicated regions in the *D. melanogaster* genome

Based on these results, we attempted to identify the genomic regions that display similar gene density and transcriptional pattern. These could represent novel putative URs, not discovered in previous studies. Using a sliding window (100 kb, step 10 kb), we built a profile showing a local average of the number of gene starts (see "Materials and methods" section) along each chromosome. The corresponding UCSC track is available in Online Resource 10. The wide variations in the gene densities calculated in this manner are found to be in close accordance with those observed in previous studies (Adams et al. 2000; Ashburner et al. 1999). Subsequently, we developed a program that searches this profile for genomic regions (160–440 kb, based on the lengths of 51 previously characterized URs) with a gene density significantly lower than that in the adjacent regions. This program also accounted the length of intergenic regions and the bias towards testis-specific expression defined by the integrated expression profile (see "Materials and methods" section).

By using this method on the whole genome, we identified 110 regions with the defined patterns (the track with the positions of predicted regions is available in Online Resource 11). The sensitivity of our algorithm was estimated to be 80%, as 41 of the known URs overlapped with the predicted regions by more than 50% of their span. The flanks were used as a negative sample to assess the false positive rate: 28 flanks overlapped with the predicted regions by more than 50% of their span, giving the estimate for the specificity of about 72% (see "Materials and methods" section). Furthermore, 45 (60%) out of 74 *de novo* detected regions were found to overlap by at least 50% of their span with the cytologically annotated late replicated bands. Only eight of them (10%) were completely early replicated in this test. A part of 2 L chromosomal arm with gene density, polytenization profile, known, and predicted regions is presented in Fig. 4.

Previously, we reported a high correlation between the URs in the polytene chromosomes and the regions of late replication in the Kc cells (Belyakin et al. 2005; Schuebeler et al. 2002). To obtain more information about the accuracy of our predictions, we compared the numbers of early and late replicated genes within the predicted regions with the recent whole genome data on replication timing in Kc and Cl8 cells (Schwaiger et al. 2009). All the genes from this experiment were classified as early (normalized log2 ratio of microarray signals above 0) and late (log2 ratio below 0). In addition, 51 experimentally determined URs representing a positive control sample were also assessed. The replication timing of the genes from the predicted regions and 51 control regions significantly shifted towards the end of the S-phase in both the Kc and Cl8 cells (Fig. 5a, b, χ^2 test, $P = 4.0 \times 10^{-37}$ for autosomal-predicted regions in Kc cells and $P = 1.6 \times 10^{-43}$ for those in Cl8 cells). As the Cl8 cells were genetically male, their X chromosome was subjected to dosage compensation. This led to a global shift in its replication timing towards the beginning of S-phase (Schwaiger et al. 2009). A similar shift was observed in the predicted regions on the X chromosome (Fig. 5a, b), and the portion of late replicated genes in these regions were not different from the entire dosage-compensated X chromosome (Fig. 5b, χ^2 test, P=0.57). On the other hand, in the female Kc cells, these regions were essentially late replicated (Fig. 5a, χ^2 test, $P = 1.1 \times 10^{-10}$).

The number of testis-specific genes was calculated in the predicted regions. A total of 370 genes that we selected as testis-specific were found in these regions, and 285 of them were not overlapping and their intergenic regions' length was studied. Strikingly, 141 (49%) of them were surrounded by long intergenic regions (>1 kb) and only 47 (16%) had both adjoining intergenic regions shorter than 1 kb (Fig. 5c). This distribution was very similar to that in the experimentally validated URs (60% and 12%, respectively, Fig. 3d), and is significantly different from the distribution of all testis-specific genes (Fig. 3e, 39% and 31%, respectively, χ^2 test, $P = 1.2 \times 10^{-6}$).

To validate the accuracy of our approach directly, a standard curve real-time PCR method was applied to eight randomly picked predicted regions to measure their DNA

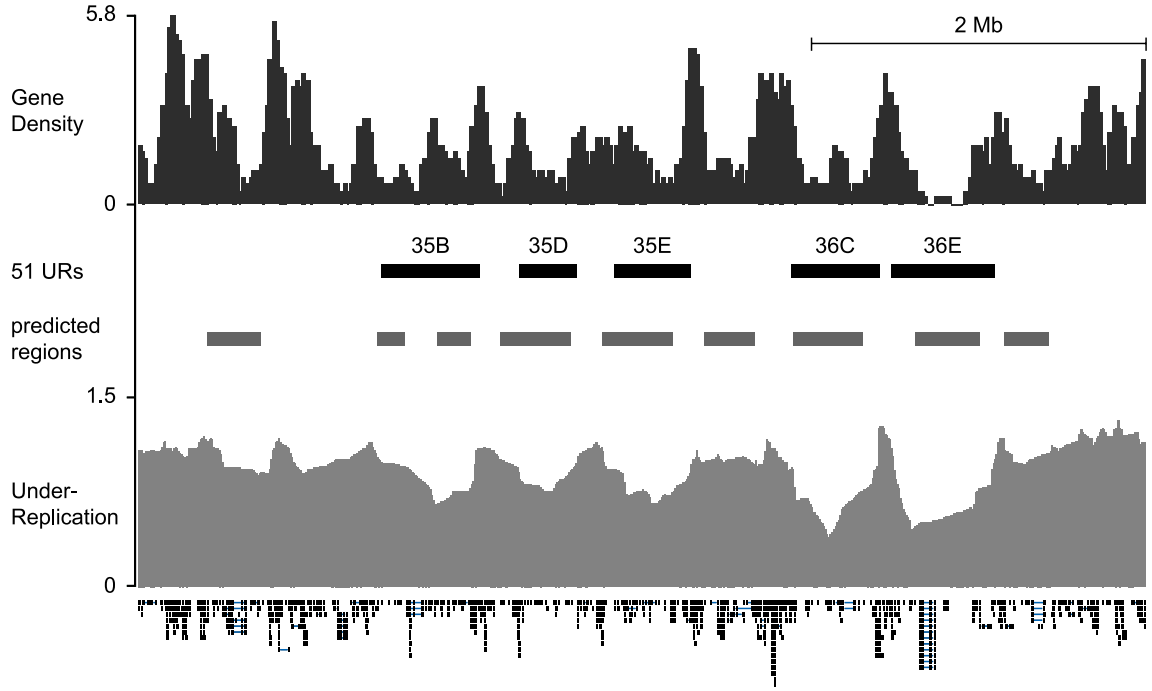

Fig. 4 Gene-density profile and positions of the predicted and experimentally discovered URs. A part of the 2 L chromosome is presented using UCSC genome browser. *Upper profile* shows gene density as calculated in this study (see text). The lowest profile represents the polytenization as determined earlier (Belyakin et al. 2005). The span of the predicted regions is shown with *gray horizontal bars*; previously discovered URs are shown with *black bars*. Protein-coding genes are shown below

polytenization in the salivary glands ("Materials and methods" section). Region 34C was taken as a negative control, because it represents a clear false positive, being early replicated in all our tests. We expected that some of these regions would show underreplication in the polytene chromosomes when extra copies of *SuUR* gene are added. As expected, region 34C showed virtually equal polytenization level in both the strains close to 100% of *Actin 42A* control (Fig. 6). In addition, three other regions—25F, 99A, and 90A—also did not show any differences in polytenization. However, four regions—63A, 97A, 33CD, and 18A—demonstrated significant reduction in polytenization level in *4xSuUR*$^{+}$ strain (two to three biological replicates were examined, three technical replicates each, significance levels were calculated using *t* test, $P=0.006$ for 33CD and 97A regions, and $P<10^{-3}$ for 63A and 18A, Fig. 6).

Discussion

Our results demonstrate that gene-density variation may be connected to replication timing regulation in *Drosophila*. As a model, we used the URs of the salivary gland polytene chromosomes, which were late replicated and regulated by *SuUR* gene.

The control set of 51 URs was obtained from our previous study (Belyakin et al. 2005). These were the regions that were found to be differentially polytenized in the polytene chromosomes of *SuUR* mutants and 4x*SuUR* strain bearing two transgenic copies of the *SuUR* wild-type allele. As we know, about 240 regions in the polytene chromosomes showed late incorporation of labeled nucleotides in the S-phase in the polytene chromosomes. However, only a part of them demonstrated high frequency of chromosomal breaks, which is a cytological evidence for underreplication (Zhimulev et al. 2003) and only these were detected as underreplicated in our experiments. Addition of extra copies of *SuUR* wild-type allele increased the number of regions manifesting reproducible chromosomal breaks (Zhimulev et al. 2003), suggesting that apart from the detected URs, there may be other regions sharing similar features.

To dissect possible peculiarities in the URs organization, we performed a number of tests comparing the 51 URs with their flanks. In mammals, a correlation exists between low CG percentage and late replication (Hiratani and Gilbert 2009; Huvet et al. 2007; Woodfine et al. 2004). We observed a similar tendency in the URs, but the variation was too slight to be significant.

In mammals, late replicated areas (Woodfine et al. 2004) and those regions that switch their replication timing on differentiation also showed low gene density (Hiratani et al. 2008). We observed abnormally low gene density in the URs originating from the peculiar combination of short

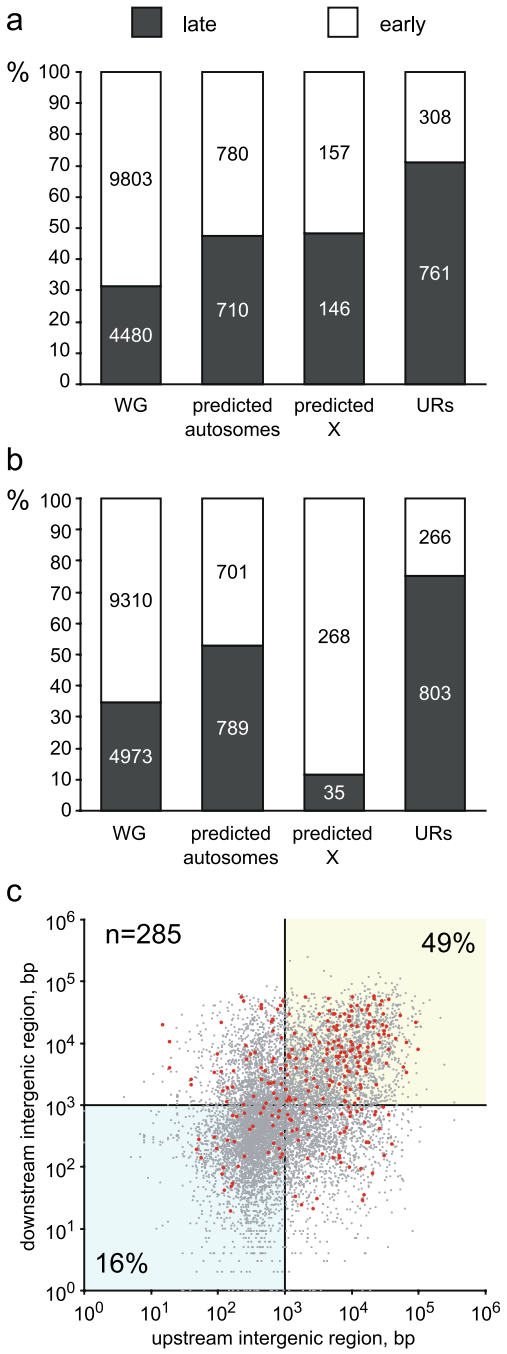

Fig. 5 Replication timing of the predicted regions. **a** Proportions of late (*black*) and early (*white*) replicated genes in the whole genome, 74 predicted regions, and 51 URs in the Kc cells (according to Schwaiger et al. 2009). **b** The same analysis for the Cl8 cells (according to Schwaiger et al. 2009). **c** Scatter plot of intergenic regions surrounding testis-specific genes (red dots) from 74 predicted regions (49% of these genes have both intergenic regions longer than 1 kb, 16% have shorter than 1 kb). All nonoverlapping genes are shown with *gray dots*

testis-specific genes and surrounding long intergenic areas. This may suggest that long intergenic regions could be the targets for the tissue-specific regulatory mechanism governing the expression and replication timing.

A large group of genes residing in the URs were found to be active in adult testis. According to the existing evidences (Schwaiger et al. 2009), mass activation of these genes should result in opening of the chromatin and shift in their replication time to the beginning of the S-phase. An indirect confirmation was presented by a study on testis-specific genes repression near the nuclear lamina in *D. melanogaster*. It was shown that in male germline stem cells, the cluster of genes in 60D region is attached to the lamina. However, later, in spermatocytes, it was found to dissociate from the lamina and relocate inside the nucleus (Shevelyov et al. 2009). Notably, the association with the nuclear lamina is one of the characteristic features of the late replicated regions (Zhimulev 1998).

Previously, enrichment with male-specific genes allowed examination of transcriptional territories in the URs (Belyakin et al. 2005). Although the organization of the URs is probably more complex than the co-regulated groups of the genes that were found previously (Spellman and Rubin 2002), we believe that our results revealed a part of the regulatory mechanism concerning testis-specific genes.

To prove the functionality of genetic organization and expression bias found in the URs, we attempted to predict more late replicated regions using these features. To test our predictions, we examined the replication features of 74 predicted regions according to previous studies (Schwaiger et al. 2009; Zhimulev et al. 2003). This comparison proved that most of the genes in the predicted regions are late

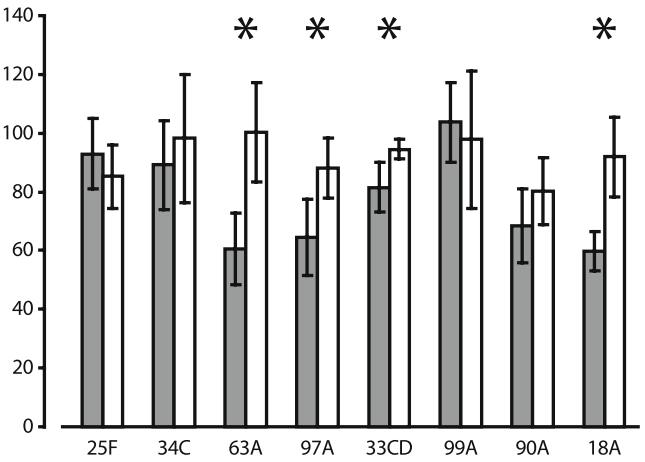

Fig. 6 Predicted regions have a potential to be underreplicated in the polytene chromosomes Underreplication in eight predicted regions was validated using real-time PCR. DNA from the salivary glands of the 4x*SuUR*$^{+}$ strain (*gray columns*) was compared with that from the *SuUR* mutants (*white columns*). The ordinate reflects the polytenization level in each region compared with the *Actin42A* gene (percent). Four regions that showed significant underreplication are marked with asterisks (*t* test, 63A: P=0.00088; 97A: P=0.0055; 33CD: P=0.0059; 18A: $P = 3.8 \times 10^{-5}$)

replicated in the salivary glands, in Kc and Cl8 cell cultures.

We applied a direct quantitative PCR approach to check the polytenization level of the observed regions, as late replicated regions may be underreplicated in polytene chromosomes. Polytenization was compared between the strain with enhanced underreplication (4x*SuUR*) and the *SuUR* mutant strain. Regions showing significant decrease in polytenization in 4x*SuUR* strain, when compared with the mutants, were considered as underreplicated. Four regions were found to be underreplicated in this test. However, this assay could not detect late replicated regions that were fully polytenized (or their underreplication is negligible), and never manifested chromosomal breaks even in the 4x*SuU*R strain. As underreplication in polytene chromosomes is not an ultimate feature of late replicated regions, this result should not be considered as decisive when evaluating our predictions.

Taken together, our results reveal that a dramatic gene-density variation along the genome may reflect the peculiarities of the late replicated regions: these are enriched with tissue-specific genes and long intergenic regions; they are flanked by the gene-dense areas. The discovered pattern allows prediction of new late URs. Thus, late replication domains represent specific structural units of the genome. Further exploration of these findings should ascertain the mechanisms that underlie the regulation of these regions.

Acknowledgments We thank Dr. Gareth Lycett and Dr. Igor Makunin for the discussion of the results and critical reading of the manuscript. The microarray slides were received through DGRC. This work was supported by Russian Foundation for Basic Research (Grant No. 08-04-01105a to S.N.B. and V.V.S.), Program of Presidium of Russian Academy of Sciences "Molecular and Cellular Biology" (Grant No. 22.4), and Interdisciplinary Integration Project of Siberian Branch of Russian Academy of Sciences (Grant No. 37) and Science Schools (Grant No. 5104.2008.4.). The manuscript was edited for proper English language by SPi Technologies, Inc (www.prof-editing.com).

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP (2002) Gene expression during the life cycle of *Drosophila melanogaster*. Science 297:2270–2275

Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N et al (1999) An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. Genetics 153:179–219

Belyaeva ES, Zhimulev IF, Volkova EI, Alekseyenko AA, Moshkin YM, Koryakov DE (1998) Su(UR)ES: a gene suppressing DNA underreplication in intercalary and pericentric heterochromatin of *Drosophila melanogaster* polytene chromosomes. Proc Natl Acad Sci U S A 95:7532–7537

Belyakin SN, Christophides GK, Alekseyenko AA, Kriventseva EV, Belyaeva ES, Nanayev RA, Makunin IV, Kafatos FC, Zhimulev IF (2005) Genomic analysis of Drosophila chromosome underreplication reveals a link between replication control and transcriptional territories. Proc Natl Acad Sci U S A 102:8269–8274

Berezney R, Dubey DD, Huberman JA (2000) Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma 108:471–484

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI (2002) Large clusters of co-expressed genes in the *Drosophila* genome. Nature 420:666–669

Chintapalli VR, Wang J, Dow JA (2007) Using Fly Atlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet 39:715–720

Cimbora DM, Schubeler D, Reik A, Hamilton J, Francastel C, Epner EM, Groudine M (2000) Long-distance control of origin choice and replication timing in the human beta-globin locus are independent of the locus control region. Mol Cell Biol 20:5581–5591

Hiratani I, Gilbert DM (2009) Replication timing as an epigenetic mark. Epigenetics 4:93–97

Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, Chang CW, Lyou Y, Townes TM, Schubeler D, Gilbert DM (2008) Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 6:e245

Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton-Carafa Y, Arneodo A, Thermes C (2007) Human gene organization driven by the coordination of replication and transcription. Genome Res 17:1278–1285

Kalisch WE, Hagele K (1976) Correspondence of banding patterns to 3 h-thymidine labeling patterns in polytene chromosomes. Chromosoma 57:19–23

Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, Meng C, Berezney R (1998) Spatial and temporal dynamics of DNA replication sites in mammalian cells. J Cell Biol 143:1415–1425

MacAlpine DM, Rodriguez HK, Bell SP (2004) Coordination of replication and transcription along a *Drosophila* chromosome. Genes Dev 18:3094–3105

Moshkin YM, Alekseyenko AA, Semeshin VF, Spierer A, Spierer P, Makarevich GF, Belyaeva ES, Zhimulev IF (2001) The bithorax complex of *Drosophila melanogaster*: Underreplication and morphology in polytene chromosomes. Proc Natl Acad Sci U S A 98:570–574

Pindyurin AV, Moorman C, de Wit E, Belyakin SN, Belyaeva ES, Christophides GK, Kafatos FC, van Steensel B, Zhimulev IF (2007) SUUR joins separate subsets of PcG, HP1 and B-type lamin targets in *Drosophila*. J Cell Sci 120:2344–2351

Schuebeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, Groudine M (2002) Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. Nat Genet 32:438–442

Schwaiger M, Stadler MB, Bell O, Kohler H, Oakeley EJ, Schubeler D (2009) Chromatin state marks cell-type- and gender-specific replication of the *Drosophila* genome. Genes Dev 23:589–601

Shevelyov YY, Lavrov SA, Mikhaylova LM, Nurminsky ID, Kulathinal RJ, Egorova KS, Rozovsky YM, Nurminsky DI (2009) The B-type lamin is required for somatic repression of testis-specific gene clusters. Proc Natl Acad Sci U S A 106:3282–3287

Simon I, Tenzen T, Mostoslavsky R, Fibach E, Lande L, Milot E, Gribnau J, Grosveld F, Fraser P, Cedar H (2001) Developmental regulation of DNA replication timing at the human beta globin locus. EMBO J 20:6150–6157

Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the *Drosophila* genome. J Biol 1:5

Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, Young BD, Debernardi S, Mott R, Dunham I, Carter NP (2004) Replication timing of the human genome. Hum Mol Genet 13:191–202

Zhimulev IF (1998) Polytene chromosomes, heterochromatin, and position effect variegation. Adv Genet 37:1–566

Zhimulev IF, Belyaeva ES (2003) Intercalary heterochromatin and genetic silencing. Bioessays 25:1040–1051

Zhimulev IF, Belyaeva ES, Makunin IV, Pirrotta V, Volkova EI, Alekseyenko AA, Andreyeva EN, Makarevich GF, Boldyreva LV, Nanayev RA et al (2003) Influence of the SuUR gene on intercalary heterochromatin in *Drosophila melanogaster* polytene chromosomes. Chromosoma 111:377–398