

Compositional Bias is a Major Determinant of the Distribution Pattern and Abundance of Palindromes in *Drosophila melanogaster*

Guoqing Liu · Jia Liu · Bingjie Zhang

Received: 22 June 2012 / Accepted: 22 October 2012 / Published online: 9 November 2012
© Springer Science+Business Media New York 2012

Abstract Palindromic sequences are important DNA motifs related to gene regulation, DNA replication and recombination, and thus, investigating the evolutionary forces shaping the distribution pattern and abundance of palindromes in the genome is substantially important. In this article, we analyzed the abundance of palindromes in the genome, and then explored the possible effects of several genomic factors on the palindrome distribution and abundance in *Drosophila melanogaster*. Our results show that the palindrome abundance in *D. melanogaster* deviates from random expectation and the uneven distribution of palindromes across the genome is associated with local GC content, recombination rate, and coding exon density. Our data suggest that base composition is the major determinant of the distribution pattern and abundance of palindromes and the correlation between palindrome density and recombination is a side-product

of the effect of compositional bias on the palindrome abundance.

Keywords Palindrome · GC content · Recombination · Gene conversion

Introduction

Palindrome, an inverted repeat without spacer, is a double-strand DNA, 5'- to -3' sequence of which is identical on each DNA strand. Palindromes are ubiquitous among the DNA of natural plasmids, viral and bacterial genomes, and eukaryotic genomes and organelles. Palindromes have various roles in the regulation of cellular processes, such as gene expression, DNA replication and recombination (Thukral et al. 1991; McClarin et al. 1986; Hiratsu et al. 2000). For example, many palindromes contained in *cis*-acting elements are binding sites for regulatory proteins (Thukral et al. 1991); the recognition sites for most restriction endonucleases are short palindromes of 4–8 bp in prokaryotic genomes; the palindromes located in telomeric regions can initiate DNA replication (Hiratsu et al. 2000). Another important aspect of palindromes is that they have the potential to fold back on themselves to form secondary structures, known as hairpin or cruciform, which are putative substrates for several structure-specific nucleases and mismatch repair enzymes, and the action of such an enzyme may result in a double-strand break (DSB) in the DNA. The hairpin structures formed by palindromes immediately preceding the terminators could increase RNA polymerase transcriptional activity by promoting more efficient template clearance (Chu et al. 1997). Hairpin structures can also be formed in RNA transcripts by palindromic sequences, and they may inhibit gene expression

Electronic supplementary material The online version of this article (doi:10.1007/s00239-012-9527-y) contains supplementary material, which is available to authorized users.

G. Liu
The Institute of Bioengineering and Technology,
Inner Mongolia University of Science and Technology,
Baotou 014010, China

G. Liu (✉) · J. Liu · B. Zhang
School of Mathematics, Physics and Biological Engineering,
Inner Mongolia University of Science and Technology,
Baotou 014010, China
e-mail: gqliu1010@163.com

G. Liu
Inner Mongolia Key Laboratory of Biomass-Energy Conversion,
Baotou 014010, China

probably by impeding the proceeding of the ribosome along the mRNA transcript (Zhang et al. 2002).

Giant palindromes that are scarce in the genome may play important roles and evolve in a unique way in eukaryotic Y chromosomes. The giant palindromes have evolved in the Y chromosome, showing remarkable divergence in their content between Chimpanzee and human Y chromosomes (Hughes et al. 2010). Abundant gene conversion occurred between arms of palindromes in human and ape Y chromosomes (Rozen et al. 2003). A large palindrome is involved in interchromosomal gene duplications in the pericentromeric region of the *Drosophila melanogaster* Y chromosome (Méndez-Lago et al. 2011).

The abundance of palindromes in the genome has been a focus owing to its diverse functions in the cell. The abundance of palindromes varies with their size in the genome. It has been shown that short palindromes are under-represented in some bacteriophages and their bacterial hosts, as well as in some archaea, in comparison with random sequences (Gelfand and Koonin 1997; Rocha et al. 2001). In *Saccharomyces cerevisiae*, short palindromes (<14 bp) are also under-represented while long palindromes are over-represented (Lisnić et al. 2005). Long, AT-rich palindromes were also over-represented in the chromosomes III and X of *Caenorhabditis elegans* (LeBlanc et al. 2000). In the human genome palindromic sequences are over-represented in introns and gene upstream regions (Lu et al. 2007). Several possible explanations for the short palindrome avoidance have been proposed. First, the avoidance of short palindromes (<8 bp) that usually serve as recognition sites for type II restriction/modification systems (RMSs) in prokaryotes was thought to be caused by the activity of RMSs (Rocha et al. 2001). Such an interpretation, however, could not explain the paucity of short palindromes in the *S. cerevisiae* genome, where no RMS has been detected. Second, a strong bias against palindromic dinucleotides, which is probably caused by intrinsic mutational bias associated with neighbor-dependent mutations, replication/repair and recombination, may explain the observed short palindrome avoidance (Lisnić et al. 2005; Karlin et al. 1997).

The over-representation of long palindromes (10–50 bp) in genomes can be explained by their selective advantages (or functional importance) in genomes, such as roles in initiation of DNA replication, meiotic recombination and transcription termination. Relatively long palindromes, however, were also known to generate genomic instabilities by forming hairpin structures (Nag and Kurst 1997; Nasar et al. 2000; Lobachev et al. 1998; Lisnić et al. 2009; Leach 1994). For example, palindromes stimulate deletions during the replication of DNA (Lobachev et al. 1998); Palindromes stimulate interchromosomal recombination between homologous sequences (Nag and Kurst 1997; Nasar et al. 2000);

Palindromes can also stimulate intrachromosomal recombination between homologous sequences leading to the loss of the intervening sequence (Lisnić et al. 2009). Hence, the over-representation of long palindromes may reflect their population advantages over disadvantages as a whole at genome level.

The evolutionary forces exposed on the abundance of palindromes in the genome are far from thorough elucidation, although several factors, such as the activity of RMSs and palindromic dinucleotide bias, are thought to affect palindrome abundance. In this paper, we investigated the possible effects of base composition and recombination on the abundance and genomic distribution of palindromes in the *D. melanogaster* genome. The results show that compositional bias is a major determinant of the abundance and distribution of palindromes in the *D. melanogaster* genome.

Materials and Methods

Genomic Sequences

Complete genome sequences, several types of genomic sequences, including 5' UTR, 3' UTR, introns, intergenic regions, and gene coding sequences (CDS) of *D. melanogaster* and their corresponding chromosomal positions were all retrieved from Flybase (ftp://ftp.flybase.net/releases/FB2006_01/dmel_r5.1/). In order to reduce redundancy, only the sequences that correspond to the first transcripts were selected when a gene locus matches more than one transcript. Repeat-masked genome of *D. melanogaster* was downloaded from UCSC (<http://www.genome.ucsc.edu>). Repeat-masked intergenic sequences, introns, and CDS were taken from the repeat-masked genome using their chromosomal positions. For statistical significance in palindrome counting and relevant correlation analysis, only the sequences with effective length, i.e., the length of sequence excluding the unknown nucleotides and the fraction of repeats greater than 1,000 bp were retained for CDS, introns, and intergenic regions. The ultimate data analyzed includes 5,948 introns, 5,795 intergenic regions, and 7,963 CDS.

Palindrome Scoring

Palindromes were counted by a program named Spinnaker (Lisnić et al. 2005). As indicated in Fig. 1, palindromes in a given sequence can be scored in different ways. First, only the palindromes that do not share common base pairs are scored (Fig. 1a). Second, all palindromes including non-overlapping palindromes, partially overlapping palindromes, and shorter palindromes entirely embedded within longer palindromes are scored (Fig. 1b). Third,

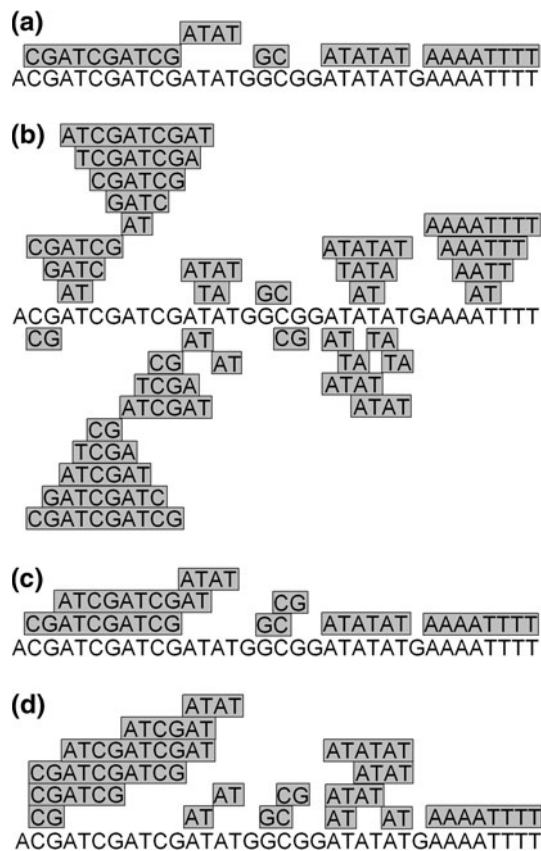


Fig. 1 Palindrome scoring. Different numbers of palindromes can be counted in the same sequence depending on the scoring criterion. **a** Non-overlapping palindromes do not share common base pairs. **b** All palindromes including non-overlapping palindromes, partially overlapping palindromes, and short palindromes contained within long palindromes. **c** Palindromes including non-overlapping and partially overlapping palindromes but not short palindromes contained within a long palindrome. **d** Palindromes including non-overlapping palindromes, partially overlapping palindromes, and the short palindromes contained within long palindromes but having a different center of symmetry

palindromes including non-overlapping palindromes, partially overlapping palindromes but not shorter palindromes entirely embedded within longer palindromes are scored (Fig. 1c). Fourth, palindromes including non-overlapping palindromes, partially overlapping palindromes, and the shorter palindromes entirely embedded within longer palindromes but having a different center of symmetry are

scored (Fig. 1d). The first approach would result in an underestimation of the actual number of palindromes. The second approach would identify too many short palindromes embedded within long palindromes in which the embedded palindromes sharing the same center of symmetry cannot represent those possessing different functional potentials. The third approach is unable to identify any short palindrome embedded within long palindromes. Short palindromes embedded within long palindromes but sharing a different center of symmetry may have different functional potentials, so we decided to count palindromes by the fourth approach.

Recombination Rate Estimation

The data of *D. melanogaster* cytogenetic bands mapping to the genome (Release 5.1) were taken from Flybase (<http://flybase.bio.indiana.edu>). We used a procedure similar to that described by Kliman and Hey (1993). A fourth-order polynomial curve was fitted to the genetic position as a function of physical position for each chromosome arm. The middle position of a sequence was chosen to represent its physical position, and the recombination rates (cM/Mb) for genomic sequences were estimated as the derivative of the polynomial function. Recombination rate on the fourth chromosome was considered to be zero.

Palindrome Density

The genome of *D. melanogaster* was split into non-overlapping fragments of 100 kb, and corresponding random sequences, length and base composition of which are the same with the 100-kb fragments of *D. melanogaster* genome were generated by computer. Then, the palindrome density, defined as the palindrome occurrence divided by the effective sequence length, was computed for each of the 100-kb fragments of the *D. melanogaster* genome and corresponding random sequences. The sequences with effective length smaller than 50 kb were not included in our analysis. For the genomic sequences including CDS, introns, and intergenic sequences, palindrome density was defined in the same way. Sample sizes of the data analyzed in this study were listed in Table 1.

Table 1 Sample size of the data analyzed in the present study

	Non-masked	Masked	Masked CDS	Masked intron	Masked intergenic
Autosomes	978	937	6,639	4,890	4,763
X	223	220	1,324	1,058	1,032
Total	1,201	1,157	7,963	5,948	5,795

Note: Non-masked, the count of non-overlapping fragments of 100 kb obtained by splitting the non-masked genome, in which the fragments with effective length smaller than 50 kb were not included. Masked, the count of non-overlapping fragments of 100 kb obtained by splitting the repeat-masked genome, in which the fragments with effective length smaller than 50 kb were not included

Results

Comparison of Palindrome Abundance in the *D. melanogaster* Genome and Random Sequences

In order to examine the possible non-random usage of palindromes in the *D. melanogaster* genome, the palindrome abundance in *D. melanogaster* and that in randomly generated sequences were compared. As shown in Fig. 2, the short palindromes of 2–4 bp are present at lower frequency in the natural *D. melanogaster* genome than random sequences, while the situation of palindromes larger than 4 bp is just the reverse. Moreover, for the palindromes larger than 4 bp, the larger the palindromes are, the more pronounced the difference of their abundance between the natural and random sequences are. We can also see that palindromes are enriched more on the fourth chromosome than on the autosomes and long palindromes larger than 10 bp are distributed more densely on the X chromosome than on the autosomes. Palindromes in repeated sequences in the genome are of less interest in this study. When repeats were masked, the difference between natural sequences and random sequences still existed, but the palindrome occurrences decreased much, and the difference in the abundance of long palindromes between the autosomes and the X chromosome became less evident (Supplementary Fig. 1). This indicated that the over-representation of long palindromes on the X chromosome is attributed largely to the repeats distributed on the X chromosome.

Palindrome Distribution is Associated with Base Composition

To explore the possible effects of base composition and recombination on the genomic distribution of palindromes in *D. melanogaster*, the correlations of palindrome density with local GC content and recombination rate for non-overlapping sliding windows of 100 kb across the genome was analyzed. Due to the difference in patterns of molecular evolution between the autosomes and the X chromosome in *D. melanogaster* (Singh et al. 2005), our analysis was done separately for the two chromosome sets.

As shown in Table 2, palindrome density is negatively correlated with local GC content on both the autosomes and the X chromosome for palindromes, except for palindromic dinucleotides. The relationship between palindrome density and recombination rate differs between the X chromosome and the autosomes: palindrome density correlates negatively with recombination rate on the autosomes, but positively on the X chromosome. We also investigated the correlation of palindrome density with coding exon density. On the autosomes, palindrome density

is negatively correlated with coding exon density for palindromes larger than 4 bp. On the X chromosome, palindrome density is positively correlated with exon density, particularly for the short palindromes smaller than 8 bp (Table 2).

It is worth noting that microsatellite repeats, such as (ATT) n and (CCG) n , could contain many palindromes, but are likely to be evolving under a different mutational model than most of the genome. So, to exclude the repeat-associated effect, we further analyzed the correlation between palindrome density and the variables (GC content and recombination rate) based on the repeat-masked genome of *D. melanogaster*. As shown in Table 3, the results based on repeat-masked genome are similar with those listed in Table 2, indicating that those correlations were not caused by repeats. Partial correlation analysis shows that most significant correlations between palindrome density and recombination in Table 3 become non-significant after controlling for genomic GC content, and the correlations between palindrome density and GC content are not affected by recombination rate, suggesting that the correlation between palindrome density and recombination is mediated by GC content.

Apart from the negative correlation between palindrome density and local GC content in the fly genome, a similar stronger negative correlation between the two variables is also observed for random sequences which have the same base composition with the 100-kb genomic fragments of *D. melanogaster* (data not shown), suggesting that the correlation is unlikely to be caused by selection. Could the compositional bias of sequences modulate palindrome abundance? The GC contents of both the 100-kb genomic fragments of *D. melanogaster* and composition-constrained random sequences are lower than 50 %. If we suppose that the negative correlation resulted from the compositional bias of the sequences, i.e., the more biased a sequence is, the more palindromes can be detected in it no matter whether the bias is toward AT- or GC-richness, it is then expected that palindrome density is negatively correlated with GC content for sequences with GC content lower than 50 % and positively correlated for sequences with GC content higher than 50 %. To test this, we generated 40 random sequences of 100 kb, which have GC content increasing from 31 to 70 % by an interval of 1 %, and correlated the palindrome densities of them with their GC contents. As expected, strong negative correlations between palindrome density and GC content were observed when GC contents are lower than 50 %, and strong positive correlations were observed when GC contents are higher than 50 % (Table 4), indicating that compositional bias of a sequence has significant influence upon its palindrome abundance. The richness of palindromes on the fourth chromosome shown in Fig. 2 is thus likely due to its low GC content.

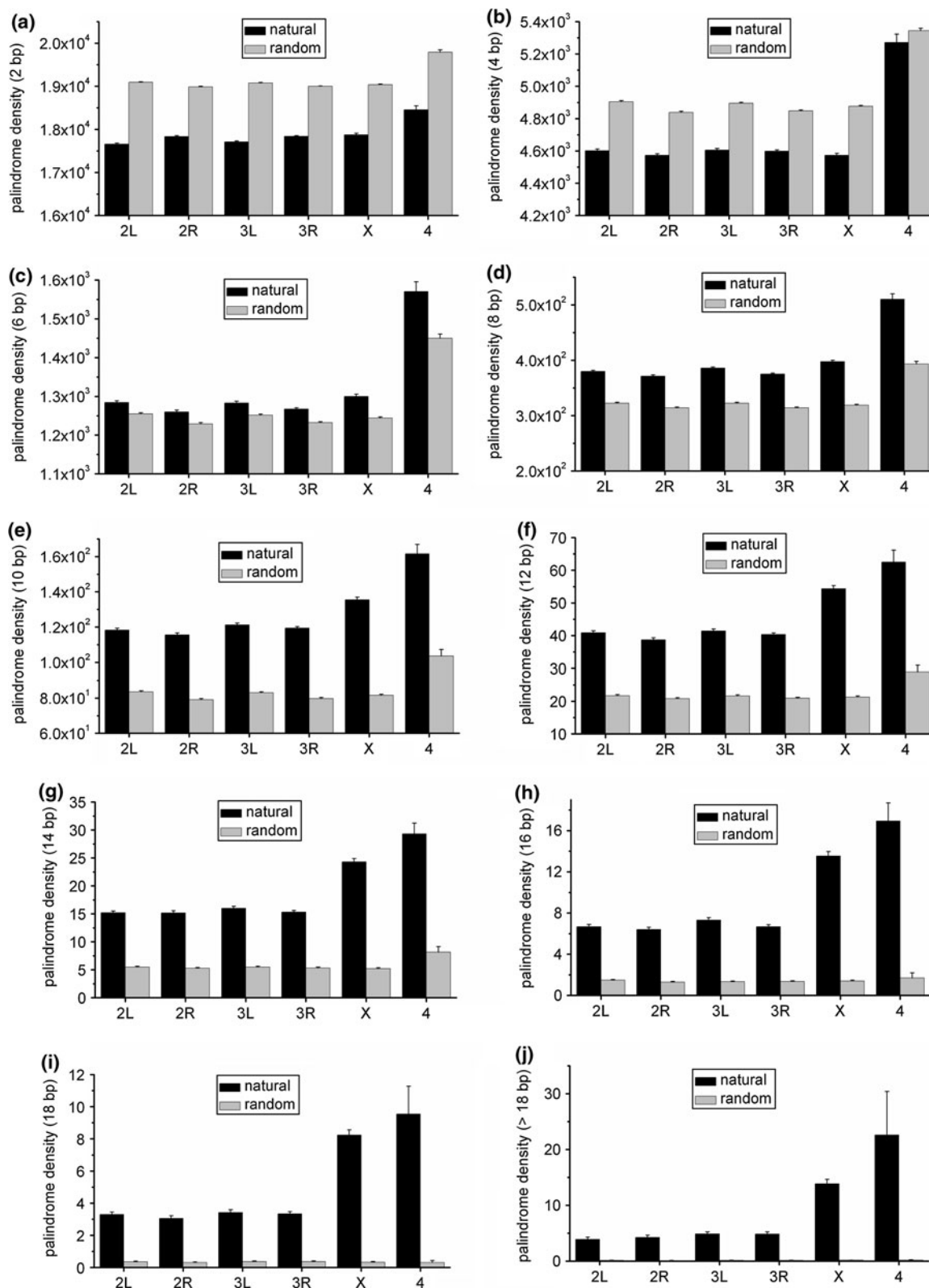


Fig. 2 Comparison of palindrome abundance in the non-masked *D. melanogaster* genome and random sequences. The size of palindromes is illustrated in the parentheses behind the title of ordinate. The ordinate represents the average of palindrome densities for both the non-overlapping 100-kb fragments of the

D. melanogaster genome and the corresponding composition-constrained random sequences (see “Materials and Methods” section for detail). Analysis of variance shows that the palindrome densities for natural sequences and random sequences are significantly different ($1 \times 10^{-46} < P < 1 \times 10^{-6}$)

Table 2 Spearman correlations of palindrome density with several factors in the non-masked genome of *D. melanogaster*

Size (bp)	Autosomes			X		
	Rec	GC	Exon	Rec	GC	Exon
2	-0.020	0.141***	0.358***	-0.025	0.358***	0.597***
4	-0.219***	-0.460**	0.009	-0.010	-0.259***	0.283***
6	-0.146***	-0.550***	-0.0342	0.193**	-0.383***	0.235**
8	-0.082**	-0.551***	-0.105**	0.294***	-0.504***	0.081
10	-0.091**	-0.450***	-0.122**	0.282***	-0.373***	0.089
12	0.010	-0.321***	-0.090**	0.429***	-0.287***	0.060
14	0.059	-0.227***	-0.097**	0.411***	-0.300***	0.006
16	-0.005	-0.221***	-0.086**	0.378***	-0.338***	-0.006
>16	-0.024	-0.158***	-0.024	0.295***	-0.235**	-0.056

Note: see the column “Non-masked” in Table 1 for the sample sizes of the data analyzed here. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

Table 3 Spearman correlations of palindrome density with several factors in the repeat-masked genome of *D. melanogaster*

Size (bp)	Autosomes			X		
	Rec	GC	Exon	Rec	GC	Exon
2	0.005(-0.023)	0.202***(0.204***)	0.389***	0.019(0.088)	0.342***(0.352***)	0.594***
4	-0.142***(-0.103**)	-0.330***(-0.317***)	0.084**	0.005(-0.030)	-0.197***(-0.199***)	0.343***
6	-0.083*(-0.021)	-0.471***(-0.466***)	-0.006	0.131*(0.085)	-0.280***(-0.263***)	0.272***
8	-0.036(0.037)	-0.492***(-0.492***)	-0.108**	0.220*(0.158*)	-0.452***(-0.430***)	0.072
10	-0.068*(-0.022)	-0.346***(-0.341***)	-0.125**	0.144*(0.085)	-0.365***(-0.348***)	0.035
12	0.006(0.042)	-0.247***(-0.250***)	-0.155***	0.111(0.082)	-0.177*(-0.161*)	0.093
14	0.058(0.082*)	-0.164***(-0.174***)	-0.143***	0.188**(0.163*)	-0.167*(-0.138*)	-0.078
16	0.018(0.038)	-0.139***(-0.143***)	-0.120**	0.092(0.075)	-0.101(-0.086)	-0.016
>16	-0.051(-0.030)	-0.157***(-0.151***)	-0.158***	0.021(-0.015)	-0.199***(-0.198**)	-0.113

Note: see the column “Masked” in Table 1 for the sample sizes of the data analyzed here. Data in the parentheses are partial correlation coefficients, and the others are pairwise correlation coefficients. In the partial correlation analysis between palindrome density and recombination rate, the control variable is local GC content, and in the partial correlation analysis between palindrome density and local GC content, the control variable is recombination rate. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

Table 4 Pearson correlation between palindrome density and GC content for random sequences

		2 bp	4 bp	6 bp	8 bp	10 bp	12 bp	14 bp	>14 bp
Approach4	GC < 50 %	-0.92***	-0.96***	-0.93***	-0.91***	-0.90***	-0.82***	-0.75**	-0.34
	GC > 50 %	0.95***	0.95***	0.92***	0.93***	0.89***	0.81***	0.61**	0.54**
Approach1	GC < 50 %	0.69**	-0.92***	-0.93***	-0.85***	-0.93***	-0.64**	-0.71**	-0.16
	GC > 50 %	-0.79***	0.94***	0.90***	0.90***	0.85***	0.64**	0.42	0.61**
Approach2	GC < 50 %	-0.96***	-0.95***	-0.94***	-0.92***	-0.91***	-0.83***	-0.77***	-0.37
	GC > 50 %	0.96***	0.95***	0.94***	0.96***	0.94***	0.86***	0.83***	0.42
Approach3	GC < 50 %	0.18	-0.94***	-0.93***	-0.90***	-0.89***	-0.82***	-0.74***	-0.34
	GC > 50 %	-0.35	0.95***	0.92***	0.92***	0.88***	0.81***	0.61**	0.54*

Note: both the high GC group and the low GC group consist of 20 random sequences of 100 kb. The GC content of the random sequences increases from 31 to 70 % by an interval of 1 %. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

Negative correlations between palindrome density and local GC content for introns and intergenic regions, and positive correlations for the CDS (Table 5) are also compatible with the aforementioned model for the compositional effect on the palindrome abundance, as majority of the CDS have GC content higher than 50 %, and majority

of the introns and intergenic regions have GC content lower than 50 %.

As introduced in methods, there are several ways of palindrome scoring, which differ in counting the overlapping palindromes. In this study, we identified palindromes by the fourth approach described in the method section. Is

Table 5 Spearman correlation between palindrome density and local GC content for the repeat-masked genomic sequences of different type in *D. melanogaster*

	Autosomes			X		
	CDS	Intron	Intergenic	CDS	Intron	Intergenic
2 bp	0.371***	-0.091***	-0.127***	0.403***	-0.057*	-0.103**
4 bp	0.073***	-0.384***	-0.416***	0.149***	-0.385***	-0.434***
6 bp	-0.004	-0.334***	-0.370***	0.127***	-0.352***	-0.305***
8 bp	0.006	-0.263***	-0.334***	0.051*	-0.240***	-0.330***
10 bp	0.004	-0.147***	-0.192***	0.020*	-0.156***	-0.194***
>10 bp	0.037**	-0.079***	-0.143***	0.096**	-0.155***	-0.131***

Note: see Table 1 for the sample size of the sequences analyzed here. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

Table 6 The second approach-based Spearman correlations of palindrome density with several factors in the repeat-masked genome of *D. melanogaster*

Size (bp)	Autosomes			X		
	Rec	GC	Exon	Rec	GC	Exon
2	-0.040	-0.033	0.299***	0.048	0.118	0.546***
4	-0.126***	-0.463***	0.016	0.070	-0.290***	0.312***
6	-0.082*	-0.563***	-0.086**	0.161*	-0.392***	0.217**
8	-0.043	-0.537***	-0.178***	0.221**	-0.458***	0.069
10	-0.040	-0.407***	-0.207***	0.162*	-0.357***	0.043
12	0.018	-0.305***	-0.220***	0.135*	-0.229**	0.012
14	0.043	-0.208***	-0.186***	0.158*	-0.209**	-0.073
16	-0.000	-0.179***	-0.165***	0.055	-0.167*	-0.055
>16	-0.038	-0.120**	-0.125***	-0.039	-0.121	-0.072

Note: see the column “Masked” in Table 1 for the sample sizes of the data analyzed here. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

it possible that the negative correlation between palindrome density and sequence GC content was caused by palindrome identifying method? To test this possibility, we identified palindromes in the above-mentioned 40 random sequences using the other approaches, and correlated their frequency with sequence GC content. As shown in Table 4, other approaches generated similar results as the fourth approach with the only exception of 2-bp palindromes, indicating that the negative correlation between palindrome density and sequence GC content cannot be attributed to the palindrome identifying method.

The fourth approach for identifying palindromes excludes internal shorter palindromes that share a center of

symmetry with a larger palindrome. Thus, it underestimates the frequencies of shorter palindromes and may generate a negative correlation between the densities of longer and shorter palindromes. By contrast, the second approach is to count any palindrome regardless of context, and does not underestimate shorter palindromes. However, when the second approach was employed for palindrome identification, similar results as the fourth approach were observed (Tables 4, 6), suggesting the choice of the fourth approach or the second approach has no bearing on our conclusion.

Palindrome GC content is negatively associated with recombination

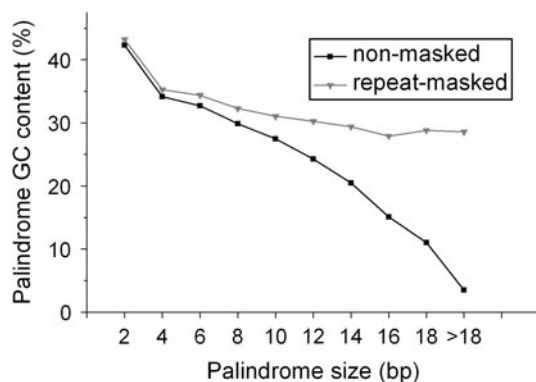


Fig. 3 The GC content of palindromes decreases with the increasing of palindrome size

As shown in Fig. 3, the GC content of palindromes strongly decreased with their size increasing. Large palindromes located in repeats are AT-rich and may affect the relationship between the GC content of palindromes and their size. However, the decreasing trend of GC content of palindromes with their size still exists when palindromes in repeats were excluded (Fig. 3).

Since both the GC content of palindromes and the relationship between palindrome density and recombination rate depend on their size (see Fig. 3 and Table 2), the correlation analysis between palindrome GC content and recombination rate was carried out for individual palindrome size to exclude the size effect. The correlation

between the mean GC content of palindromes in genomic 100-kb fragments of *D. melanogaster* and recombination rate is shown in Table 7. A different correlation pattern between the autosomes and the X chromosome was observed, although the correlations on the X chromosome were non-significant. Palindrome GC content is positively correlated with recombination on the autosomes and negatively on the X chromosome, which is in agreement with the observation that local GC content correlates positively with recombination rate on the autosomes and negatively on the X chromosome of *Drosophila* (Singh et al. 2005). Furthermore, the correlation between palindrome GC content and recombination rate on autosomes becomes weak with the increasing size of palindromes.

We also examined how palindrome GC content is affected by recombination in the sequences of different type, such as gene coding regions, introns, and intergenic regions. As shown in Table 8, the correlation differs between the autosomes and the X chromosome. Regardless of the sequence type, palindrome GC content is positively

correlated with recombination on the autosomes and negatively on the X chromosome, which is consistent with the finding illustrated in Table 7.

The chromosome 4 is very AT-rich, and therefore one may suspect that its contribution to the observed correlations between palindrome density and local GC content and between palindrome GC content and recombination rate is quite strong. Our analysis, however, shows that the correlations and the decreasing trend of palindrome GC content with the palindrome size increasing are not affected much by chromosome 4 (data not shown).

Several major findings can be summarized as follows: (1) In the *D. melanogaster* genome, short palindromes (<6 bp) are under-represented, while relatively long palindromes (>6 bp) are over-represented as compared to those in random sequences; (2) The density of palindromes is negatively correlated with local GC content on both the autosomes and X chromosome; (3) The correlation between the density of palindromes and recombination rate is mediated by local GC content; (4) Coding exon density is positively correlated with the density of short palindromes (<6 bp) and negatively with the density of long palindromes (>6 bp); (5) The GC content of palindromes decreases with their size increasing, and it correlates with recombination rate positively on the autosomes and negatively on the X chromosome.

Table 7 Spearman correlation between the GC content of palindromes and recombination rate in the repeat-masked genome of *D. melanogaster*

Palindrome size (bp)	Autosomes	X
2	0.230***	-0.058
4	0.199***	-0.089
6	0.208***	-0.078
8	0.205***	-0.097
10	0.186***	-0.096
12	0.093**	-0.125
14	0.111**	-0.095
16	0.074*	0.115
>16	-0.009	-0.118

Note: see the column “Masked” in Table 1 for the sample sizes of the data analyzed here. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

Discussion

Palindrome Abundance in *Drosophila*

Lisnić et al. proposed that a strong bias against palindromic dinucleotides may explain the observed short palindrome avoidance in *S. cerevisiae* (Lisnić et al. 2005). In order to test if the under-representation of short palindromes of 2-4 bp in the *D. melanogaster* genome is caused by biased usage of palindromic dinucleotides, we counted the short palindromes center of symmetry of which are composed of

Table 8 Spearman correlation between the GC content of palindromes and recombination rate for the repeat-masked genomic sequences of different type in *D. melanogaster*

Size (bp)	Autosomes			X		
	CDS	Intron	Intergenic	CDS	Intron	Intergenic
2	0.013	0.082***	0.170***	-0.093**	0.025	-0.174***
4	0.027*	0.075***	0.130***	-0.081*	-0.009	-0.203***
6	0.042**	0.053**	0.121***	-0.092**	0.010	-0.176***
8	0.033**	0.044*	0.102***	-0.080**	0.037	-0.203***
10		0.033*	0.055***		-0.068	-0.101**
>10		0.022	0.054**		0.021	-0.074*

Note: see the column “Masked” in Table 1 for the sample sizes of the sequences analyzed here. Two-tailed significance: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.0001$

Table 9 Comparison of the abundances of short palindromes (2–4 bp) between the non-masked fly genome and random sequences

	AT	TA	CG	GC	AT-centered	TA-centered	CG-centered	GC-centered
Fly	5,903	4,336	3,190	4,353	1,509	1,295	743	1,040
Random	6,131	6,136	3,378	3,380	1,568	1,572	864	862
Fly-random	−228***	−1800***	−188***	973***	−59***	−277***	−121***	178***

Note: see the column “Non-masked” in Table 1 for the sample sizes of the data analyzed here. The values in the first two lines represent the average of palindrome densities (number/100 kb). Analysis of variance shows that AT, TA and CG-centered palindromes are under-represented while GC-centered palindromes are over-represented in the fly genome, as compared with random sequences. Two-tailed significance: *** $P < 0.0001$

the palindromic dinucleotides. Consistent with our previous findings (Liu and Li 2008), the AT, TA, and CG-centered palindromes are significantly rare in *D. melanogaster*, while GC-centered palindromes are rich (Table 9). These indicate that the under-representation of short palindromes of 2–4 bp is, at least partially, caused by the biased usage of palindromic dinucleotides. The biased usage of palindromic dinucleotides was probably shaped by intrinsic mutational bias associated with neighbor-dependent mutations, replication/repair and recombination (Arndt et al. 2003; Karlin et al. 1997; Liu and Li 2008).

By contrast with short palindromes, long palindromes are over-represented in both *S. cerevisiae* and *D. melanogaster*. Keeping the beneficial effects of long palindromes in various cellular processes and their deleterious effects associated with genomic instabilities together in mind, we speculate that the actual occurrence of long palindromes in the genome is the consequence of the balance between two selective forces acting oppositely on them. The force associated with beneficial effects tend to increase long palindromes, while the force associated with deleterious effects decrease them. Compared with very long palindromes (>50 bp), the moderately long palindromes (10–50 bp) are less easy to result in DSBs (Nasar et al. 2000). So we propose that the over-representation of palindromes of moderate size (10–50 bp) detected in *D. melanogaster* may reflect their population advantages over disadvantages at genome level. The negative correlation between the GC content of palindromes and their size may be caused by a positive selection of high A + T content for long palindromes, because high A + T content may facilitate local DNA melting and adoption of secondary structures. The long AT-rich palindromes may be generated through the insertion of AT or TA dinucleotides into AT-rich palindromes by slippage during DNA replication (Lisnic et al. 2005; Toth et al. 2000).

Why Palindrome Distribution is Correlated with GC Content and Recombination?

The strong correlation between palindrome abundance and local GC content in the *Drosophila* genome can be predicted by the following random expectation. The probability that

two symmetrical bases of a palindrome are complementary is $P_2 = 2p_Ap_T + 2p_Gp_C$, where p_i is the proportion of base i ($i = A, G, C, T$). Thus, the expected probability of k -base palindromes is

$$P_k = (2p_Ap_T + 2p_Gp_C)^{\frac{k}{2}} \quad (1)$$

Then the expected frequency of k -base palindromes is $F_k = P_k(L - k + 1)$, where L is the sequence length. An assumption that $p_A = p_T$ and $p_G = p_C$, combined with the normalization $p_A + p_T + p_G + p_C = 1$, generates $p_A + p_G = \frac{1}{2}$. Then we have

$$\begin{aligned} P_k &= (2p_Ap_T + 2p_Gp_C)^{\frac{k}{2}} \\ &= (2p_A^2 + 2p_G^2)^{\frac{k}{2}} \\ &= 2^{\frac{k}{2}} \left[\left(\frac{1}{2} - p_G \right)^2 + p_G^2 \right]^{\frac{k}{2}} \\ &= 2^k \left[\left(p_G - \frac{1}{4} \right)^2 + \frac{1}{16} \right]^{\frac{k}{2}} \end{aligned} \quad (2)$$

It is clear that the probability of palindromes has the minimum when $p_G = \frac{1}{4}$. Given $\begin{cases} p_A + p_T + p_G + p_C = 1 \\ p_A = p_T, p_G = p_C \end{cases}$, we know that the minimum probability of palindromes occurs when $p_A = p_T = p_G = p_C = 25\%$. Accordingly, palindrome frequency should bottom when GC content is 50%. This means the probability of palindromes decreases as GC content approaches 50%. Given the overall AT bias in the *Drosophila* genome, a negative correlation between GC content and palindrome density is expected. Consistent with this expectation, we observed negative correlations between GC content and palindrome densities, as listed in Table 3. Although base composition has a great effect on palindrome abundance, it is worth noting that the compositional effect is not great enough to solely explain the distribution of palindromes across the genome, as indicated by the low correlation coefficients in Table 3.

Our data indicated that except palindromic dinucleotides the density of palindromes is negatively correlated with

recombination rate on the autosomes and positively on the X chromosome. Can selection explain these correlations? Recombination accelerates the evolution of sexual reproductive organisms by segregating advantageous and disadvantageous genes (Felsenstein 1974; Kliman and Hey 1993; Lewin 2004). In other words, selection efficiency is elevated in regions of increased recombination rates. This effect could modulate the distribution of genomic elements, such as transposons and pseudogenes, in the genome (Bartolome et al. 2002; Hua-Van et al. 2005; Liu et al. 2010). Similarly, it is possible that beneficial palindromes are preferentially retained and deleterious palindromes are efficiently eliminated in the regions of high recombination rates. Short palindromes might be selectively disadvantageous while longer ones might be selectively advantageous, as implicated in their abundance in the genome. If selection associated with recombination is responsible for the under-representation of short palindromes and over-representation of long palindromes, it would be expected that recombination rate correlates negatively with the density of short palindromes and positively with the density of long palindromes. No such consistent result was observed between the autosomes and the X chromosome, implying that such a selection cannot explain the correlation between palindrome distribution and recombination.

Partial correlation analysis suggested that the effect of base composition is stronger than the effect of recombination on palindrome abundance and the relationship between palindrome abundance and recombination is a mere side-product of the effect of compositional bias on palindrome density. One may claim that recombination affects palindrome abundance indirectly by influencing base composition, as GC content is affected by recombination and GC content in turn could affect the palindrome abundance. However, besides the recombination, there are other factors affecting GC content in the genome, and thus it cannot be concluded that recombination determines the palindrome abundance in a composition-associated manner.

Why Palindrome GC Content is Correlated with Recombination?

Recombination rate correlates positively with the palindrome GC content, genomic GC content, and the density of dinucleotide palindromes composed of nucleotide G and C on the autosomes, and negatively on the X chromosome. The GC-biased gene conversion of which frequency is proportional to recombination rate could be invoked to explain the positive correlations on the autosomes, but it could not explain the negative correlations on the X chromosome.

There are several hypotheses that could explain the negative correlation between GC content and recombination rate on the X chromosome. First, a background

substitutional bias toward decreased GC with increased recombination rate could be mediated by a recombination-associated mutational bias toward A and T on the X chromosome. However, this mutational explanation seems unlikely, as to date there are no empirical data to support such a difference in the underlying mechanisms of mutation between the autosomes and the X chromosome. Second, a background substitutional bias mediated by gene conversion could explain the observation. In general, recombination increases the GC content of genomic sequences including palindromic sequences through gene conversion that may frequently occur in regions of high recombination rates (Birdsell 2002; Galtier et al. 2001; Marais et al. 2003; Meunier and Duret 2004; Singh et al. 2005). Nevertheless, if rates of gene conversion on the X chromosome were high in areas of low recombination, then this could lead to an X-specific, negative correlation between GC content and recombination rate. It does seem possible that rates of gene conversion and rates of recombination are inversely related on the *Drosophila* X chromosome (Langley et al. 2000). Third, it is also possible that the negative correlation between recombination rate and GC content on the X chromosome is the result of historical recombination rates rather than current ones. The negative correlation on the X chromosome can be explained if the X chromosome underwent a recent rearrangement of the recombination map such that the currently highly recombining areas experienced historically low recombination rates and visa versa.

In this study, we also found that coding exon density tends to correlate positively with the density of palindromes shorter than 6 bp and negatively with the density of palindromes longer than 6-bp. The negative correlation is probably due to selection force that makes palindromes rare in or around genes to avoid unbeneficial effect on genes. One simple explanation for this is that the excess of large palindromic motifs in coding sequences is harmful to gene function and counter-selected. For example, palindromes in coding sequences might be selected against due to their potential to induce DSB or deletions. It is unclear that why the density of palindromes shorter than 6-bp correlates positively with coding exon density.

Conclusion

To conclude, we explored the possible evolutionary forces shaping the distribution pattern and abundance of palindromes in the *D. melanogaster* genome. Our results show that the palindrome occurrence in the genome deviates from random expectation and the uneven distribution of palindromes across the genome is associated with local GC content, recombination rate, and coding exon density. Our

data suggest that the correlation between palindrome abundance and recombination is a side-product of the effect of compositional bias on palindrome abundance, and the base composition is the main driver of the distribution pattern and abundance of palindromes in the *D. melanogaster* genome.

Acknowledgments We thank Xiu-Juan Zhao for her helpful discussions. This work was supported by grants from the National Natural Science Foundation (61102162), the Research Program of Higher Education of Inner Mongolia Autonomous Region (NJ10098) and the Innovation Fund of Inner Mongolia University of Science and Technology (2009NC005).

References

- Arndt PF, Burge CB, Hwa T (2003) DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol* 10:313–322
- Bartolome C, Maside X, Charlesworth B (2002) On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol* 19:926–937
- Birdsell JA (2002) Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol Biol Evol* 19:1181–1197
- Chu W, Ballard R, Schmind C (1997) Palindromic sequence preceding the terminator increase polymerase III template activity. *Nucl Acids Res* 25:2077–2082
- Felsenstein J (1974) The evolutionary advantage of recombination. *Genetics* 78:737–756
- Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159:907–911
- Gelfand MS, Koonin EV (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucl Acids Res* 25:2430–2439
- Hiratsu K, Mochizuki S, Kinashi H (2000) Cloning and analysis of the replication origin and the telomeres of the large linear plasmid pSLA2-L in *Streptomyces rochei*. *Mol Gen Genet* 263:1015–1021
- Hua-Van A, Rouzic AL, Maisonhaute C, Capy P (2005) Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenet Genome Res* 110:426–440
- Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SK, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, Trask BJ, Mardis ER, Warren WC, Repping S, Rozen S, Wilson RK, Page DC (2010) Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* 463:536–539
- Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* 179:1363–1370
- Kliman RM, Hey J (1993) Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol Biol Evol* 10:1239–1258
- Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM (2000) Linkage disequilibrium and the site frequency spectra in the su(s) and su(wa) regions of the *Drosophila melanogaster* X chromosome. *Genetics* 156:1837–1852
- Leach DR (1994) Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays* 16:893–900
- LeBlanc MD, Aspeslagh G, Buggia NP, Dyer BD (2000) An annotated catalog of inverted repeats of *Caenorhabditis elegans* chromosomes III and X, with observations concerning odd/even biases and conserved motifs. *Genome Res* 10:1381–1392
- Lewin B (2004) *Genes VIII*. Pearson Prentice Hall, Upper Saddle River
- Lisnić B, Svetec IK, Sarić H, Nikolić I, Zgaga Z (2005) Palindrome content of the yeast *Saccharomyces cerevisiae* genome. *Curr Genet* 47:289–297
- Lisnić B, Svetec IK, Stafa A, Zgaga Z (2009) Size-dependent palindrome-induced intrachromosomal recombination in yeast. *DNA Repair* 8:383–389
- Liu G, Li H (2008) The correlation between recombination rate and dinucleotide bias in *Drosophila melanogaster*. *J Mol Evol* 67:358–367
- Liu G, Li H, Cai L (2010) Processed pseudogenes are located preferentially in regions of low recombination rates in the human genome. *J Evol Biol* 23:1107–1115
- Lobachev KS, Shor BM, Tran HT, Taylor W, Keen JD, Resnick MA, Gordenin DA (1998) Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. *Genetics* 148:1507–1524
- Lu L, Jia H, Droge P, Li J (2007) The human genome-wide distribution of DNA palindromes. *Funct Integr Genomics* 7:221–227
- Marais G, Mouchiroud D, Duret L (2003) Neutral effect of recombination on base composition in *Drosophila*. *Genet Res* 81:79–87
- McClarín JA, Frederick CA, Wang BC, Greene P, Boyer HW, Grable J, Rosenberg JM (1986) Structure of the DNA-EcoRI endonuclease recognition complex at 3 Å resolution. *Science* 234:1526–1541
- Méndez-Lago M, Bergman CM, de Pablos B, Tracey A, Whitehead SL, Villasante A (2011) A large palindrome with interchromosomal gene duplications in the pericentromeric region of the *D. melanogaster* Y chromosome. *Mol Biol Evol* 28:1967–1971
- Meunier J, Duret L (2004) Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol* 21:984–990
- Nag DK, Kurst A (1997) A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast *Saccharomyces cerevisiae*. *Genetics* 146:835–847
- Nasar F, Jankowski C, Nag DK (2000) Long palindromic sequences induce double-strand breaks during meiosis in yeast. *Mol Cell Biol* 20:3449–3458
- Rocha EPC, Danchin A, Viari A (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res* 11:946–958
- Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423:873–876
- Singh ND, Davis JC, Petrov DA (2005) Codon bias and non-coding GC content correlate negatively with recombination rate on the *Drosophila* X chromosome. *J Mol Evol* 61:315–324
- Thukral SK, Eisen A, Young ET (1991) Two monomers of yeast transcription factor ADR1 bind a palindromic sequence symmetrically to activate ADH2 expression. *Mol Cell Biol* 11:1566–1577
- Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 10:967–981
- Zhang R, Guo S, Ren M (2002) Analysis the influence of palindrome structure to gene expression by constructing combination system. *Acta Microbiologica Sinica* 42:186–192 (in Chinese)