Deficiency of X-Linked Inverted Duplicates with Male-Biased Expression and the Underlying Evolutionary Mechanisms in the *Drosophila* Genome

Zhen-Xia Chen,[1] Yong E. Zhang,[2] Maria Vibranovski,[2] Jingchu Luo,[1] Ge Gao,*,[1] and Manyuan Long*,[2]

[1]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, PR China

[2]Department of Ecology and Evolution, University of Chicago

***Corresponding author:** E-mail: mlong@midway.uchicago.edu; gaog@mail.cbi.pku.edu.cn.

Associate editor: Aoife McLysaght

Abstract

Inverted duplicates (IDs) are pervasive in genomes and have been reported to play functional roles in various biological processes. However, the general underlying evolutionary forces that maintain IDs in genomes remain largely elusive. Through a systematic screening of the *Drosophila melanogaster* genome, 20,223 IDs were detected in nonrepetitive intergenic regions, far more than expectation under the neutrality model. 3,846 of these IDs were identified to have stable hairpin structure (i.e., the structural IDs). Based on whole-genome transcriptome profiling data, we found 628 unannotated expressed structural IDs, which had significantly different genomic distributions and structural properties from the unexpressed IDs. Among the expressed structural IDs, 130 exhibited higher expression in males than in females (i.e., male-biased expression). Compared with sex-unbiased ones, these male-biased IDs were significantly underrepresented on the X chromosome, similar to previously reported pattern of male-biased protein-coding genes. These analyses suggest that a selection-driven process, rather than a purely neutral mutation-driven mechanism, contributes to the maintenance of IDs in the *Drosophila* genome.

Key words: inverted duplicates, noncoding RNA, sex evolution, MSCI, meiotic drive, *Drosophila melanogaster*.

Introduction

An inverted duplicate (ID) consists of two perfect or nearly perfect duplicates (here called "arms") of a particular DNA sequence that are located next to each other in reverse orientation. An ID is termed as "palindrome" when the distance between the arms (here called "spacer length") is zero. IDs have been observed in excess amounts in various organisms, including bacteria (Ladoukakis and Eyre-Walker 2008), yeast (Strawbridge et al. 2010), and humans (Wang and Leung 2009). Functional roles have been reported for several individual IDs (Tao, Masly, et al. 2007; Gleghorn et al. 2008; Larson et al. 2008; Okamura et al. 2008; Randau et al. 2009; Bussmann et al. 2010; Geraldes et al. 2010). However, sequence instability caused by IDs has also been documented (Mizuno et al. 2009; Tanaka and Yao 2009; Darmon et al. 2010). A functionless ID would be expected to accumulate deletions until its complete loss from the genome (Yang et al. 2008). Thus, an excess number of IDs indicates functionality as a maintaining force. Despite individual case reports, the general mechanism of ID function is unknown. One possibility is that IDs function through their encoded hairpin RNAs, or stem loops, which consist of a double-stranded RNA stem and a terminal loop (supplementary fig. S1, Supplementary Material online). The formation of hairpins in microRNA (miRNA) precursors is important for their biogenesis and regulatory functions (Ruby et al. 2007).

Differences in chromosome composition between males and females provide opportunities to understand the function of hairpin RNAs. Sex chromosomes often show distinct evolutionary patterns related to their specific genetics and biology. Due to the hemizygosity of the X chromosome in males, nonneutral mutations in autosomes and the X chromosome are subject to different selective dynamics from each other and are consequently expected to evolve at different rates, which leads to an uneven distribution of evolutionary changes between the X chromosome and the autosomes (Rice 1984; Charlesworth et al. 1987; Vicoso and Charlesworth 2006; Ellegren and Parsch 2007). This prediction is supported by mounting experimental evidence, which has demonstrated that many genes expressed exclusively or preferentially in one sex are distributed unevenly between the sex chromosomes and autosomes in *Drosophila melanogaster* (Parisi et al. 2003; Ranz et al. 2003; Vibranovski, Zhang, and Long 2009; Zhang, Vibranovski, Krinsky, and Long 2010), *Caenorhabditis elegans* (Reinke et al. 2004), mammals (Lercher et al. 2003; Zhang, Vibranovski, Landback, et al. 2010), and birds (Storchova and Divina 2006). Notably, studies in *D. melanogaster* have revealed that male-biased genes tend to be underrepresented on the X chromosome (Parisi et al. 2003; Ranz et al. 2003; Ellegren and Parsch 2007; Vibranovski, Zhang, and Long 2009). However, most of these observations have focused on protein-coding genes, whereas little is known about the evolution of noncoding genes.

Given the mounting evidence for the functionality of noncoding RNAs (Hildebrandt and Nellen 1992; Avner and Heard 2001; Dai et al. 2008), we took advantage of the whole-genome tiling array expression data for *D. melanogaster* (Gao G, Vibranovski M, Zhang L, et al. unpublished data) to investigate the distribution of intergenic IDs encoding hairpin RNAs on autosomes and the X chromosome. We found a nonrandom distribution of expressed IDs: Intergenic IDs encoding male-biased hairpin RNAs were underrepresented on the X chromosome. Our observations not only suggest the general functionality of noncoding hairpin RNAs but also indicate that meiotic drive and/or more likely, meiotic sex chromosome inactivation (MSCI) play an important role in the evolution of noncoding genes.

Materials and Methods

Initial Detection of IDs

We searched for IDs in the entire genome of *D. melanogaster* (dm3, April 2006, downloaded from University of California at Santa Cruz (UCSC) Genome Browser website, http://hgdownload.cse.ucsc.edu/) using the program Inverted Repeats Finder (IRF) version 3.05 (Warburton et al. 2004). The IRF program assesses IDs using a scoring function in which positive scores are given for complementary pairs in the arms and penalties are assigned for mismatches and insertions/deletions (indels). Here, we assigned a score of +2 to each Watson–Crick base pair (A-T or G-C) match, −3 to each mismatch, and −5 to each indel. Therefore, the score of the ID shown in supplementary figure S1, Supplementary Material online, for example, is 8 because it has 8 matches, 1 mismatch, and 1 indel in the stem.

The IRF program reports IDs that satisfy threshold values for minimum score (MinScore), maximum spacer length (MaxLoop), and maximum arm length (MaxLength). We specified MaxLoop as 80 to find IDs with spacers not more than 80 nt in length. We specified MaxLength as 10,000 and MinScore as 20 so that all IDs with arm lengths not longer than 10,000 nt and scores not lower than 20 would be reported. The shortest IDs we detected had perfectly complementary arms as long as 10 nt with zero-length spacer.

In summary, we searched for IDs in the assembled chromosome arms (Chr2L, Chr2R, Chr3L, Chr3R, and ChrX) using IRF with the following parameters: 2, 3, 5, 70, 10, 20, 10000 nt, 80, –d, –h, and –a3. All of the 80,348 IDs that we found are here referred to as “whole-genome IDs.”

Subsequent Processing of IDs

Because we were interested in unannotated noncoding IDs, we filtered the initial data set as described below.

The gene annotation of *D. melanogaster* was downloaded from FlyBase (r5.23, http://flybase.org/) (Wilson et al. 2008). IDs that did not overlap with any coding genes (exons and introns) were considered to be intergenic IDs.

Because repetitive regions cause redundancy in the search for IDs, we excluded all IDs that had any overlap with regions annotated by RepeatMasker (Smit et al. 1996–2010) or Tandem Repeats Finder (Benson 1999).

We defined a structural ID as an ID capable of folding into a stem–loop structure in both strands, with the stem formed by the arms of the ID through base pairings and the loop formed by the spacer between the arms. To identify structural IDs, the sequences of all ID regions (including both arms and the spacer for each ID) were extended by 10 nt in both flanking regions (to obtain longer stems) and then folded using the RNAfold program (Zuker and Stiegler 1981; McCaskill 1990; Hofacker et al. 1994). To make sure that the two arms of each ID were complementary to each other in a stable secondary structure, we retained only those IDs whose extended sequences had structures that satisfied two criteria: 1) more than five pairings (A-U, C-G, G-U), centered in the middle of the ID, were formed between the regions that corresponded to the ID arms; and 2) the arms were not shorter than 21 nt, based on the length distribution of known small regulatory RNAs (ca. 21–23 nt for small interfering RNA [siRNA], Zamore et al. 2000 and ca. 22 nt for miRNA, Bartel 2004). Unfolded regions at both terminals were discarded, and the sequences were refolded (supplementary fig. S2, Supplementary Material online). Those IDs whose sense and antisense transcripts had folded stem–loop structures with a minimum free energy (MFE) not greater than −15 kcal/mol (two alternative cutoff values, −10 and −20 kcal/mol, were also used) were defined as structural IDs, whereas IDs that were not capable of forming stem–loop structures in either strand were defined as nonstructural IDs.

We have previously performed expression profiling of both male and female whole body fruit flies, together with the reproductive organs, including the testis, ovary, and accessory glands, using Affymetrix tiling arrays (Gao G, Vibranovski M, Zhang L, et al. unpublished data). Thousands of male-biased (higher expression in males), female-biased (higher expression in females), and unbiased (no significant difference between whole body males and females) transcribed fragments (transfrags) were identified (Gao G, Vibranovski M, Zhang L, et al. unpublished data). Taking advantage of this data set, we classified the 634 structural IDs that overlapped with any transfrags as expressed IDs and the remaining 3,212 structural IDs as unexpressed IDs. The expressed structural IDs potentially encoded hairpin RNAs. Within this data set, only six IDs overlapped with known noncoding genes, including one FlyBase (Wilson et al. 2008) noncoding gene (supplementary table S1, Supplementary Material online, CR32314-RA) and five miRNAs in miRBase (Griffiths-Jones et al. 2008) (supplementary table S1, Supplementary Material online). Our data set did not overlap with the remaining 81 intergenic miRNAs because their stems had more mismatches or indels than our pipeline allowed. We retained the remaining 628 expressed IDs that encoded unknown hairpin RNAs for further analyses. We referred to structural IDs that overlapped with male-biased but not female-biased transfrags as “male-biased IDs,” to those that overlapped with female-biased but not male-biased transfrags as "female-biased IDs," and to all others as "unbiased IDs."

We further integrated Manak's tiling array data over the first 24 h of *D. melanogaster* development (Manak et al. 2006) and White's RNA-sequencing (RNA-Seq) data across the entire life cycle of *D. melanogaster* (http://www.ncbi.nlm.nih.gov/geo/, GSE18068) to confirm whether our unexpressed IDs had no expression in those data sets, either. The majority of the unexpressed IDs (2,791/3,212 = 87%) were not expressed in either data set (supplementary fig. S3, Supplementary Material online), which demonstrated that they were truly not transcribed.

Simulations

We generated 5,000 randomized fly genomes to assess statistical significance of our data. We shuffled only the nonrepetitive intergenic regions because differences in base compositions between protein-coding regions and repetitive intergenic regions would potentially distort the ID frequency in the regions of interest. In other words, because we were focusing on nonrepetitive intergenic regions, we randomly shuffled these regions on the basis of their original sequence composition while keeping other genomic regions unchanged to maintain the individual genomic context of each nonrepetitive intergenic region. Therefore, compared with the observed intergenic regions, the corresponding nonrepetitive intergenic regions in the simulated genomes had different DNA sequences but the same nucleotide compositions, relative positions, and region lengths.

Results

Abundance of IDs in the *D. melanogaster* Genome

We identified IDs (defined as two ID copies of sequences nearly complementary to each other and separated by a spacer not more than 80 nt in length) in the genome sequence of *D. melanogaster* (dm3) using IRF 3.05 (Warburton et al. 2004). In total, 80,348 IDs were identified (see Materials and Methods), and the overall genome-wide density was as high as 676 IDs per million bases (Mb). The longest ID, whose arms were 856 nt in length, was found within intron 2 of the protein-coding gene *Cip4* on chromosome 3L (supplementary fig. S4, Supplementary Material online).

To investigate IDs with unknown functions, we focused on the 34,937 IDs located in intergenic regions that did not overlap with any exons or introns of coding genes annotated in FlyBase r5.23. We further excluded all IDs that overlapped with repeats annotated by RepeatMasker (Smit et al. 1996–2010) and Tandem Repeats Finder (Benson 1999). We retained a data set of 20,223 IDs with an average density of approximately 527 IDs per Mb across nonrepetitive intergenic regions.

Characterization of IDs

To obtain an overview of repeat-masked IDs, we further characterized their sequence features relative to 5,000 randomized genomes simulated based on base composition (see Materials and Methods).

First, the overall percentages of matches and indels were similarly distributed in observed and simulated IDs (fig. 1*A* and *C*). However, as shown in fig. 1*E* and *F*, the observed IDs included a higher proportion with longer arms; the simulated IDs tended to be enriched for categories with arm lengths less than 20 nt (Fisher's exact test [FET], $P = 0.050$), whereas the observed IDs were enriched for almost all other categories (FET, $P = 0.044$). Known small endogenous regulatory RNAs are usually longer than 20 nt (Zamore et al. 2000; Bartel 2004; Kim 2005, 2006). If IDs are processed similarly to miRNA (i.e., generating mature regulatory RNAs from their arms), then this contrasting pattern suggests that the observed IDs are more likely to encode functional RNAs. Furthermore, IDs with longer arms ($\geq$20 nt) exhibited different patterns of percentages of matches and indels (fig. 1*B* and *D*). Approximately 30% and 65% of the masked IDs had perfectly complementary arms and matched arms with no indels, respectively (fig. 1*A* and *C*). However, these proportions dropped to about 0% and 25% (fig. 1*B* and *D*), respectively, for IDs with longer arms. These results were expected with our detection method: IDs with longer arms could tolerate more mismatches or indels while still satisfying our detection cutoffs.

Second, the observed IDs were relatively more A/T rich than the simulated IDs (fig. 1*G*). Again, if the stem–loop structure of an ID must be processed to generate a single-stranded RNA, a high G/C content may cause an energy barrier to unfolding such a structure. In other words, this different distribution of base composition suggests the functionality of the observed IDs.

Third, the simulated IDs were distributed almost evenly relative to the spacer length, whereas the frequency of observed IDs decreased with increasing spacer length. If the spacer is too long, a stable stem–loop structure may be difficult to form, again suggesting the functionality of the observed IDs.

Finally, among the 20,223 repeat-masked IDs, there were 3,846 structural IDs capable of forming stem loops in both strands and 11,524 nonstructural IDs incapable of forming stem loops in either strand (see Materials and Methods). We further classified the structural IDs into four types according to their structural motifs (fig. 2): 1) classical stem looped, 2) stick shaped, 3) pronged, and 4) watch shaped. This diversity suggests that IDs may have versatile functions.

Chromosomal Distribution of IDs

Notably, the number of repeat-masked IDs was three times the expected value based on our simulations (fig. 3; 7,016.3 ± 85.6). We then analyzed the chromosomal distribution, folding probability, and expression pattern of these IDs to better understand why the genome encodes so many of them.

Without repeats masked, IDs were overrepresented on the X chromosome (8,511 or 24%) relative to the total length of intergenic regions on all chromosomal arms (fig. 4*A* and table 1, FET, $P < 2.2 \times 10^{-16}$). However, with the repeats masked, this enrichment was not statistically significant (fig. 4*B* and table 1, FET, $P > 0.05$). Thus, this

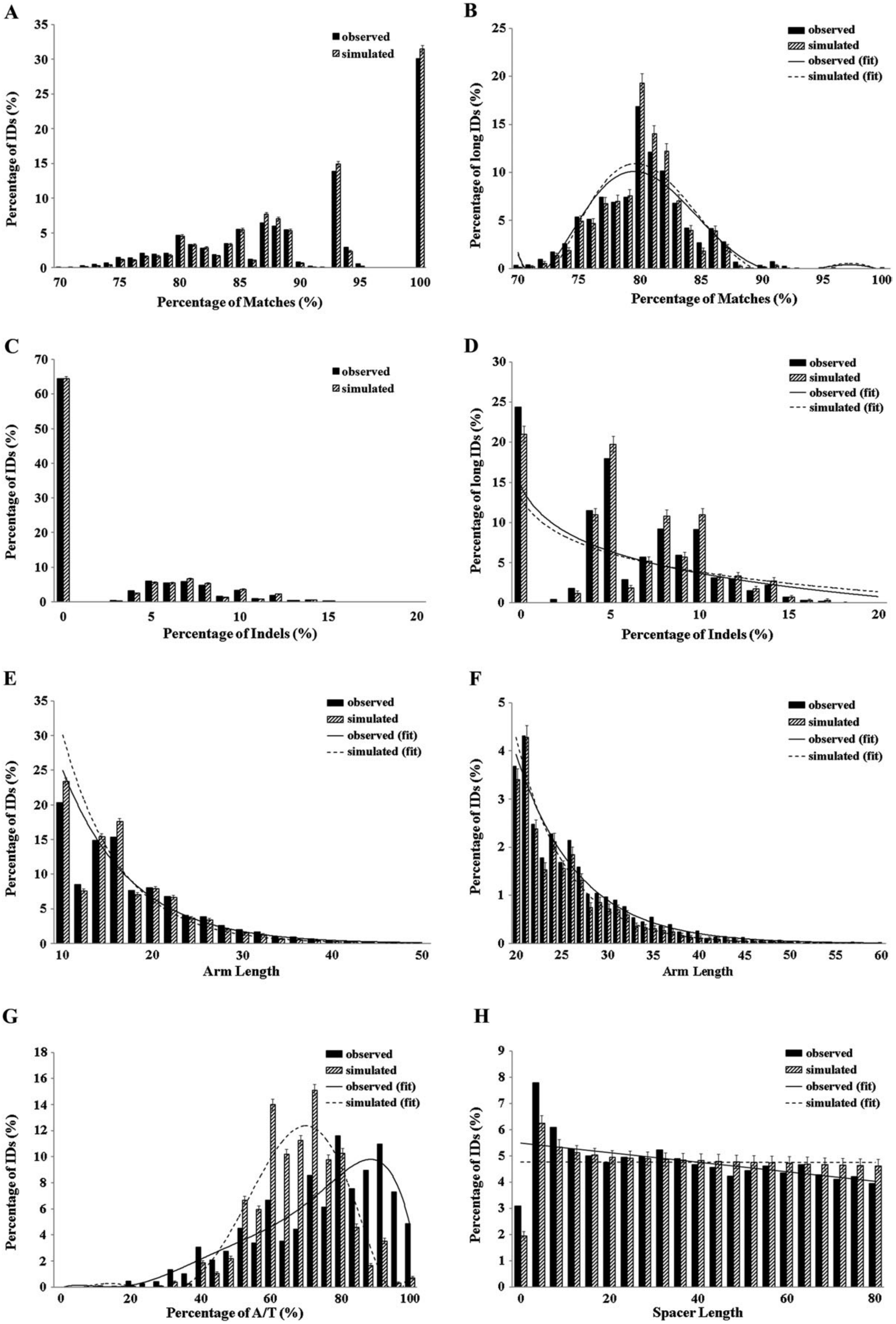

Fig. 1. Characterization of intergenic IDs after masking low-complexity regions. (*A*, *B*) percentage of matches between the two arms; (*C*, *D*) percentage of indels between the two arms; (*E*, *F*) arm length (average of two arms); (*G*) percentage of A/T in the arms; and (*H*) spacer length. Panels *A*, *C*, and *E* are based on all IDs, whereas Panels *B*, *D*, and *F* focus on relatively longer IDs (referred as long IDs, arm length ≥ 20 nt). We fitted the distribution of the numbers of observed and simulated IDs (solid and dashed curves, respectively) using the least-squares method in Excel 2010. The bars indicate the standard deviation for the simulated IDs.

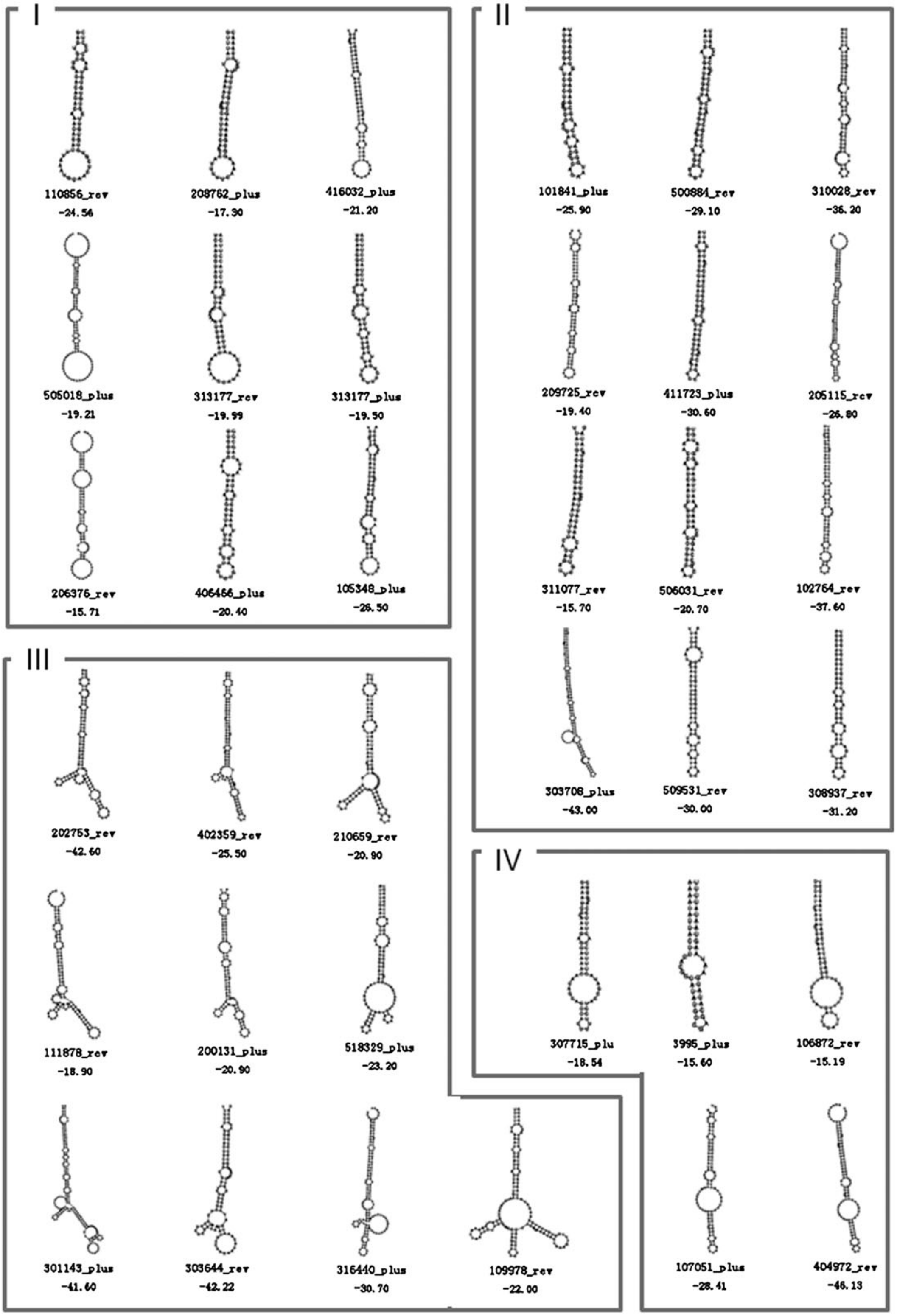

FIG. 2. Thirty-six randomly selected structural IDs in nonrepetitive intergenic regions. Each ID encodes a structure with arms longer than 20 nt and a loop not longer than 80 nt. The structures can be classified into four groups: I) classical stem loop, II) stick shaped, III) pronged, and IV) watch shaped. The name, including a unique assigned number for the ID and its corresponding strand (plus: plus strand; rev: reverse strand), and the minimum free energy (MFE; kcal/mol) are shown below each structure.

excess can be attributed mainly to the higher repetitive element content of the X chromosome. Moreover, structural and nonstructural IDs were similarly distributed between the autosomes and the X chromosome (table 2) across different structure prediction cutoffs. Thus, after accounting for repeat content, IDs appeared to be evenly distributed between the autosomes and the X chromosome. However, it is possible that only a small portion

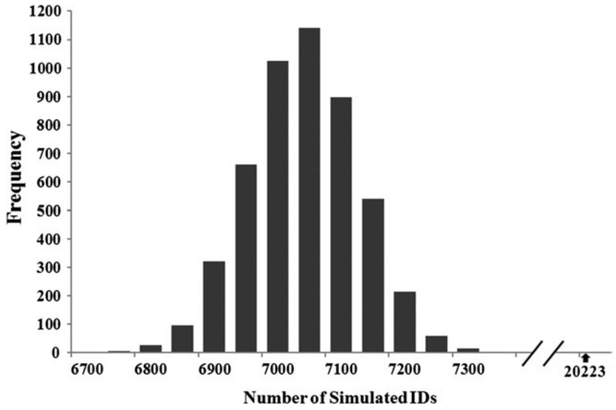

Fig. 3. Histogram of the number of IDs derived from the locally randomized nonrepetitive intergenic regions of 5,000 genomes based on nucleotide composition (see Materials and Methods). The frequency follows a distribution with mean of 7,016.3 and a standard deviation of 85.6. The arrow on the right indicates the observed number of IDs (20,223) in the nonrepetitive intergenic regions of the *Drosophila melanogaster* genome.

of the 3,846 structural IDs are functional. Thus, any nonrandom pattern for functional IDs, even if it exists, would be overlooked in this overall analysis.

Therefore, we integrated genome-wide transcriptional data (Gao G, Vibranovski M, Zhang L, et al. unpublished data) to identify 628 unannotated expressed IDs. We examined how these expressed IDs were distributed among the X chromosome and autosomes (see Materials and Methods). If IDs do function at the RNA level, we would expect unexpressed IDs to be distributed differently between the autosomes and the X chromosome compared with expressed IDs because nonneutral mutations on the autosomes and the X chromosome evolve at different rates, thus causing an uneven distribution of functional IDs (Charlesworth et al. 1987; Vicoso and Charlesworth 2006). Furthermore, considering that sex-biased transcription is subject to various contrasting forces (Ellegren and Parsch 2007; Zhang, Vibranovski, Krinsky, and Long 2010 ; Zhang, Vibranovski, Landback, et al. 2010), we compared unbiased IDs rather than expressed IDs with unexpressed IDs. We found that unbiased IDs were enriched on the X chromosome (FET, $P < 0.05$, table 3). However, compared with unbiased structural IDs, male-biased structural IDs were underrepresented on the X chromosome (FET, $P < 0.05$, table 4), whereas female-biased structural IDs were not (table 4). As expected, sex-biased ID expression was mainly contributed by the reproductive organs. For example, up to 88 (68%) IDs that encoded male-biased hairpin RNAs were expressed in the testis, whereas only 42 (32%) and 48 (37%) IDs that encoded male-biased hairpin RNAs were expressed in the ovary and accessory gland, respectively (table 5).

The nonrandom chromosomal distribution of expressed IDs suggests their potential functionality. We further compared the expressed IDs with the unexpressed structural IDs to identify additional signals of functionality for the former group. We found that the expressed IDs had longer arms than the unexpressed IDs (with an average length of 23.9 vs. 21.5, Wilcoxon rank test $P = 0.002$, table 6). This difference could be interpreted as either a bias created by our method of detecting expressed IDs, in which longer IDs were more likely to overlap with the expression data, or as a result of purifying selection to maintain the structure-dependent function of IDs. However, under the first scenario, the expressed IDs would be expected to have longer spacers, which was not consistent with our observations (expressed IDs vs. unexpressed IDs: 40.6 vs. 40.8, Wilcoxon rank test $P = 0.905$, table 6). Together with the minimum length of the structural IDs (21 nt; see Materials and Methods) and the length distribution of known small regulatory RNAs (21~23 for siRNA, Zamore et al. 2000 and ~22 for miRNA, Bartel 2004), this evidence suggested that the expressed IDs were more likely to be functional. In other words, the stem–loop structure is functionally important and maintained by purifying selection.

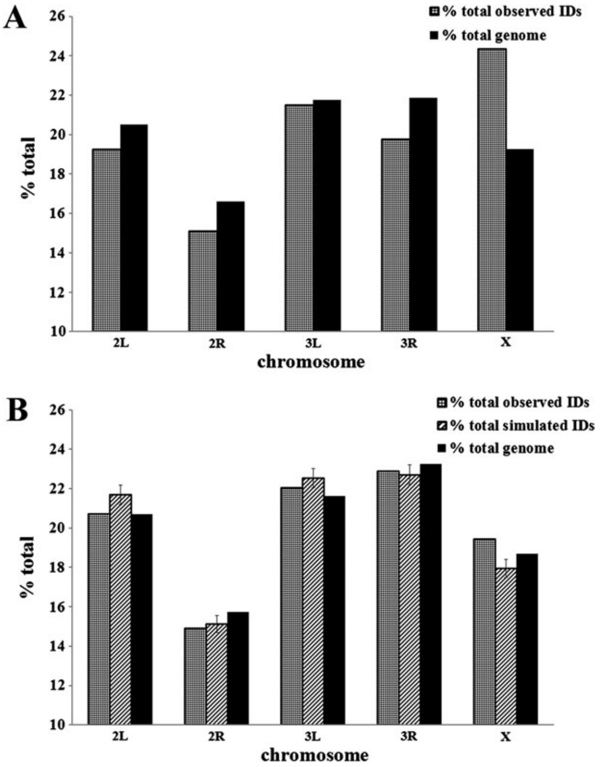

Fig. 4. Intergenic chromosomal distribution of IDs. (*A*) ID distribution without repeat masking. Gridded bars indicate the percentages of observed IDs on each chromosome arm; black bars indicate the length distributions of intergenic regions. (*B*) ID distribution with repeat masking. Bars filled with slashes indicate the percentages of simulated IDs on each chromosome arm. Standard deviations for the number of simulated IDs are shown.

Table 1. The Chromosomal Distribution of Identified Intergenic IDs.

	Autosomes		X		
	Observed	Expected	Observed	Expected	*P* value
Unmasked	26,426	28,211	8,511	6,726	$<2.2 \times 10^{-16}$***
Masked	16,291	16,443	3,932	3,780	0.056

Note.—Two-tailed FET. Expected values were calculated in terms of the sequence length in the autosomes and the X chromosome. Masked: intergenic IDs that remained after masking repetitive genomic regions.
***$P < 0.001$.

Table 2. Structural and Nonstructural IDs in Nonrepetitive Intergenic Regions Are Similarly Distributed Between the Autosomes and the X Chromosome.

	Structural		Nonstructural		
MFE Limit (kcal/mol)	Autosomes	X	Autosomes	X	*P* value
MFE ≤ −10	4,008	959	7,786	1,853	0.912
MFE ≤ −15	3,104	742	9,298	2,226	0.981
MFE ≤ −20	2,178	549	11,244	2,690	0.328

NOTE.—Two-tailed FET. Structural: IDs that could form stem–loop structures in both strands, given the indicated minimum free energy (MFE) threshold. Nonstructural: IDs that could not form stem–loop structures in either strand, given the indicated MFE threshold. The MFE value for each ID was calculated using RNAfold.

The series of comparisons described above is summarized in figure 5.

Discussion

The D. melanogaster Genome Encodes an Excess of IDs, Which Suggests Their Potential Functionality as Noncoding RNAs

Our analysis revealed a significant excess of IDs in the intergenic regions of the *D. melanogaster* genome (fig. 3), which raises the question of why they are maintained in the genome. Our analyses indicated two major reasons for the large excess of IDs. First, the genomic environment (i.e., repetitive sequences) may have generated an excess of X-linked IDs (fig. 4), which suggests a mechanistic force for ID creation. Second, purifying selection has apparently protected expressed IDs from degeneration and eventual deletion from the genome, owing to their functionality.

In nonrepetitive intergenic regions, the role of repetitive sequences in creating IDs can be ignored. Furthermore, considering the genomic instability that IDs are expected to cause (Mizuno et al. 2009; Tanaka and Yao 2009; Darmon et al. 2010) and the strong selection in the fly genome against deleterious mutations (Yang et al. 2008), the excess of IDs in the fly genome would be unlikely to be maintained if they did not play functional roles. Therefore, natural selection may act to eliminate mutations that destroy the structures of these IDs. To detect possible functions of IDs, we compared structural IDs with nonstructural IDs and expressed IDs with unexpressed IDs.

Interestingly, the differences in arm length (table 6) and distribution (fig. 5 and table 3) between unexpressed and expressed IDs suggest that expression is an important factor affecting the distribution and structure of IDs. Because the expressed structural IDs that we investigated are located in intergenic regions, they are more likely to function at the RNA level through their encoded hairpin RNAs. Moreover, because the hairpin RNAs encoded by these IDs have stems longer than 20 nt (which could be processed into miRNAs or siRNAs), they may have regulatory functions. Additionally, most (>70%) of the structural IDs, which have structures different from classical stem loops (fig. 2) and cannot encode any known noncoding RNAs, might represent new types of noncoding RNAs with novel functions.

Table 3. Unbiased Structural IDs That Encode Hairpin RNAs Show a Higher Enrichment on the X Chromosome Than Unexpressed Structural IDs.

	MFE ≤ −10 kcal/mol		MFE ≤ −15 kcal/mol		MFE ≤ −20 kcal/mol	
	Autosomes	X	Autosomes	X	Autosomes	X
Unexpressed	3,415	787	2,606	606	1,805	441
Unbiased	397	131	333	103	245	84
P value	0.001**		0.020*		0.016*	

NOTE.—Two-tailed FET. Unexpressed: structural IDs with no overlap with any transfrags. Unbiased: structural IDs that overlap with unbiased transfrags. MFE, minimum free energy.
*P < 0.05, **P < 0.01.

Table 4. Structural IDs That Encode Male-Biased Rather Than Female-Biased Hairpin RNAs are Underrepresented in the X Chromosome Compared with Those That Encode Unbiased Hairpin RNAs.

	MFE ≤ −10 kcal/mol			MFE ≤ −15 kcal/mol			MFE ≤ −20 kcal/mol		
	A	X	*P* value	A	X	*P* value	A	X	*P* value
Unbiased	397	131		333	103		245	84	
Female biased	56	15	0.558	49	13	0.749.	36	11	0.859
Male biased	135	25	0.017*	111	19	0.029*	87	13	0.009**

NOTE.—Two-tailed FET. IDs that encoded sex-biased (female-biased or male-biased) hairpin RNAs were compared with unbiased IDs. A: autosomes; X: X chromosome; MFE, minimum free energy.
*P < 0.05, **P < 0.01.

Notably, the X chromosome is enriched for unbiased IDs relative to unexpressed IDs (fig. 5 and table 3). This pattern may have some mechanistic cause; for example, the X chromosome may be transcriptionally permissive for IDs. Further functional study of these IDs is necessary to elucidate why they are often X linked.

Underrepresentation of Male-Biased IDs on the X Chromosome

Compared with IDs that encode unbiased hairpin RNAs, IDs that encode male-biased hairpin RNAs are underrepresented on the X chromosome (fig. 5 and table 4), whereas those that encode female-biased hairpin RNAs are not (fig. 5 and table 4), which suggests selection related to sex evolution. The demasculinization of the X chromosome for protein-coding genes has been observed in mice (Khil et al. 2004) and flies (Sturgill et al. 2007), but the mechanisms involved are just beginning to be understood. The analogous paucity of IDs that encode male-biased hairpin RNAs on the X chromosome suggests that selective properties, rather than the consequences of mutational

Table 5. Structural IDs (MFE ≤ −15 kcal/mol) That Encode Male-Biased Hairpin RNAs Are Preferentially Expressed in the Testis.

Organ	Expressed	Unexpressed	*P* value
Testis	88	42	
Ovary	42	88	1.72×10^{-8}***
Accessory gland	48	82	1.07×10^{-6}***

NOTE.—Two-tailed FET. Male-biased IDs in the ovary and accessory gland were compared with those in the testis. MFE, minimum free energy.
***P < 0.001.

Table 6. Expressed and Unexpressed Structural IDs Have Different Arm Lengths but Similar Spacer Lengths.

	N	Mean	*P* value
Arm length			
Unexpressed	3212	21.50	0.002**
Expressed	628	23.89	
Spacer length			
Unexpressed	3212	40.78	0.905
Expressed	628	40.57	

Note.—Two-tailed Wilcoxon rank-sum test with continuity correction for arm length and spacer length between unexpressed and expressed IDs.

mechanisms, are responsible for the autosomally biased distribution. Three selection-based hypotheses which have been proposed previously can explain this interesting phenomenon: sexual antagonism (Rice 1984; Vicoso and Charlesworth 2006), MSCI (Vibranovski, Lopes, et al. 2009; Vibranovski et al. 2010), and meiotic drive (Tao, Araripe, et al. 2007; Tao, Masly, et al. 2007).

According to the sexual antagonism model (Rice 1984; Vicoso and Charlesworth 2006), the fixation of sex-biased genes (either male or female biased) on the X chromosome depends on the level of dominance of fitness. Sexually antagonistic alleles with dominant or partially dominant advantageous effects in males and deleterious effects in females could be accumulated on the autosomes by fixation under the joint forces of selection and genetic drift. If sexual antagonism accounts for the underrepresentation of IDs that encode male-biased hairpin RNAs on the X chromosome, most mutations would be expected to be dominant or partially dominant. In this case, we would also expect IDs that encode female-biased hairpin RNAs to be enriched on the X chromosome. However, this prediction is not supported by our data (table 4 and fig. 5).

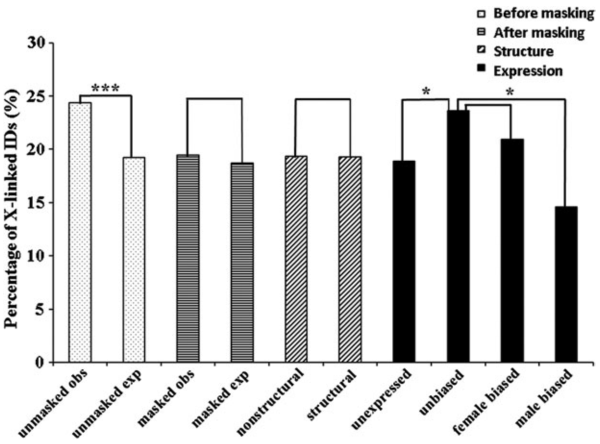

Fig. 5. Factors associated with the distribution of intergenic IDs between the autosomes and the X chromosome. Percentage of X-linked IDs is the percentage of X-linked IDs in the fly genome, including the four assembled autosome arms and the X chromosome. Three factors were tested: 1) genomic context, as shown in the left-most two columns filled with dots (before masking) and the two columns filled with horizontal lines (after masking) (Unmasked: IDs in intergenic regions before masking of repetitive regions; Masked: IDs in nonrepetitive regions where repeats were masked; obs: observed number of IDs; and exp: expected number of IDs based on the sequence length of the genomic context); 2) structure, as shown in the central two columns filled with slashes; and 3) expression, as shown in the right-most four columns filled in with black. Six pairs of columns (linked with lines) were compared. The asterisks above the lines show the significance of the differences in the comparisons. *$P < 0.05$, ***$P < 0.001$.

According to MSCI, the X chromosome is transcriptionally silenced during meiosis in the male, as recently demonstrated (Vibranovski, Lopes, et al. 2009; Vibranovski et al. 2010). Therefore, genes that function during meiotic prophase should escape the X chromosome to avoid the fate of being functionally inactivated. Previous studies in both flies (Betran et al. 2002; Vibranovski, Lopes, et al. 2009; Vibranovski, Zhang, and Long 2009) and mammals (Emerson et al. 2004; Potrzebowski et al. 2008) have shown that new gene duplicates escaped from the X chromosome under the selective force of MSCI, as confirmed by a recent study of gene expression profiles at different stages in the fly testis (Vibranovski, Lopes, et al. 2009). If MSCI is responsible for the paucity of IDs that encode male-biased hairpin RNAs on the X chromosome, those IDs should be preferentially expressed in the testis, where MSCI occurs. Our observations indicate that most IDs that encode male-biased hairpin RNAs are expressed in the testis, far more than are expressed in the ovary or accessory gland (table 5), which suggests that MSCI might have played a role in the evolution of male-biased IDs.

According to the meiotic drive hypothesis (Tao, Araripe, et al. 2007; Tao, Masly, et al. 2007), there are intragenomic conflicts over sex ratio because sex-linked genes would be disproportionately represented in the next generation if they shifted the sex ratio to more female or more male offspring by favoring their carrier sex chromosome. Autosomal ID–induced RNA silencing has been reported to be a mechanism that suppresses X-linked sex ratio distorters (Tao, Araripe, et al. 2007; Tao, Masly, et al. 2007). If many male-biased IDs have evolved to suppress potential X-linked distorters, we would expect male-biased IDs to be enriched on autosomes. Moreover, male-biased IDs would be expected to be preferentially expressed in the testis, where meiotic drive occurs, relative to other reproductive organs. Therefore, our observations are also consistent with the meiotic drive hypothesis.

However, MSCI can be a mechanism of suppressing the potential meiotic drive that results from intragenomic conflicts over sex ratio because it silences the expression of sex-linked genes, including sex ratio distorters. In this case, autosomal IDs must play other regulatory functions in spermatogenesis. However, MSCI is often an incomplete process. Some genes escape from the inactivation process, thus presenting the biological issue predicted by the meiotic drive hypothesis. Therefore, autosomal suppressors would still be selected for silencing the expression of sex ratio distorters on sex chromosomes. Meanwhile, MSCI might also be enhanced and extended to suppress the active or newly evolved distorters (Meiklejohn and Tao 2010). Therefore, meiotic drive may facilitate the evolution of and work together with MSCI, contributing to the excess of autosomal IDs that encode male-biased RNAs.

Supplementary Material

Supplementary table S1 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Shuqi Zhao and Jiongtang Li for suggestions on computational techniques. This work was supported by National Key Basic Research Program of China (2007CB946904, 2011CBA01102), two grants of National Institutes of Health and National Science Foundation in USA (R01GM078070-01A1 and MCB-1051826 to M.L.), CBC (Chicago Biomedical Consortium) support and the grant 111.

References

Avner P, Heard E. 2001. X-chromosome inactivation: counting, choice and initiation. *Nat Rev Genet*. 2:59–67.

Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–297.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 27:573–580.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res*. 12:1854–1859.

Bussmann M, Baumgart M, Bott M. 2010. RosR (Cg1324), a hydrogen peroxide-sensitive MarR-type transcriptional regulator of *Corynebacterium glutamicum*. *J Biol Chem*. 285:29305–29318.

Charlesworth B, Coyne J, Barton N. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat*. 130:113–146.

Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, Eyre-Walker A, Du W, Long M. 2008. The evolution of courtship behaviors through the origination of a new gene in Drosophila. *Proc Natl Acad Sci U S A*. 105:7478–7483.

Darmon E, Eykelenboom JK, Lincker F, Jones LH, White M, Okely E, Blackwood JK, Leach DR. 2010. E. coli SbcCD and RecA control chromosomal rearrangement induced by an interrupted palindrome. *Mol Cell*. 39:59–70.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet*. 8:689–698.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* 303:537–540.

Geraldes A, Rambo T, Wing RA, Ferrand N, Nachman MW. 2010. Extensive gene conversion drives the concerted evolution of paralogous copies of the SRY gene in European rabbits. *Mol Biol Evol*. 27:2437–2440.

Gleghorn ML, Davydova EK, Rothman-Denes LB, Murakami KS. 2008. Structural basis for DNA-hairpin promoter recognition by the bacteriophage N4 virion RNA polymerase. *Mol Cell*. 32:707–717.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. 36:D154–D158.

Hildebrandt M, Nellen W. 1992. Differential antisense transcription from the Dictyostelium EB4 gene locus: implications on antisense-mediated regulation of mRNA stability. *Cell* 69:197–204.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh Chem*. 125:167–188.

Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. 2004. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nat Genet*. 36:642–646.

Kim VN. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*. 6:376–385.

Kim VN. 2006. Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes. *Genes Dev*. 20:1993–1997.

Ladoukakis ED, Eyre-Walker A. 2008. The excess of small inverted repeats in prokaryotes. *J Mol Evol*. 67:291–300.

Larson MH, Greenleaf WJ, Landick R, Block SM. 2008. Applied force reveals mechanistic and energetic details of transcription termination. *Cell* 132:971–982.

Lercher MJ, Urrutia AO, Hurst LD. 2003. Evidence that the human X chromosome is enriched for male-specific but not female-specific genes. *Mol Biol Evol*. 20:1113–1116.

Manak JR, Dike S, Sementchenko V, et al. (11 co-authors). 2006. Biological function of unannotated transcription during the early development of *Drosophila melanogaster*. *Nat Genet*. 38:1151–1158.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers* 29:1105–1119.

Meiklejohn CD, Tao Y. 2010. Genetic conflict and sex chromosome evolution. *Trends Ecol Evol (Personal edition)*. 25:215–223.

Mizuno KI, Lambert S, Baldacci G, Murray JM, Carr AM. 2009. Nearby inverted repeats fuse to generate acentric and dicentric palindromic chromosomes by a replication template exchange mechanism. *Genes Dev*. 23:2876–2886.

Okamura K, Chung WJ, Ruby JG, Guo H, Bartel DP, Lai EC. 2008. The *Drosophila* hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* 453:803–806.

Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science* 299:697–700.

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jegou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol*. 6:e80.

Randau L, Stanley BJ, Kohlway A, Mechta S, Xiong Y, Soll D. 2009. A cytidine deaminase edits C to U in transfer RNAs in Archaea. *Science* 324:657–659.

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* 300:1742–1745.

Reinke V, Gil IS, Ward S, Kazmer K. 2004. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* 131:311–323.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* 38:735–742.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res*. 17:1850–1864.

Smit A, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0. Available from: http://www.repeatmasker.org.

Storchova R, Divina P. 2006. Nonrandom representation of sex-biased genes on chicken Z chromosome. *J Mol Evol*. 63:676–681.

Strawbridge EM, Benson G, Gelfand Y, Benham CJ. 2010. The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. *Curr Genet*. 56:321–340.

Sturgill D, Zhang Y, Parisi M, Oliver B. 2007. Demasculinization of X chromosomes in the *Drosophila* genus. *Nature* 450:238–241.

Tanaka H, Yao MC. 2009. Palindromic gene amplification—an evolutionarily conserved role for DNA inverted repeats in the genome. *Nat Rev Cancer*. 9:216–224.

Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL. 2007. A sex-ratio meiotic drive system in *Drosophila simulans*. II: an X-linked distorter. *PLoS Biol*. 5:e293.

Tao Y, Masly JP, Araripe L, Ke Y, Hartl DL. 2007. A sex-ratio meiotic drive system in *Drosophila simulans*. I: an autosomal suppressor. *PLoS Biol.* 5:e292.

Vibranovski MD, Chalopin DS, Lopes HF, Long M, Karr TL. 2010. Direct evidence for postmeiotic transcription during *Drosophila melanogaster* spermatogenesis. *Genetics* 186:431–433.

Vibranovski MD, Lopes HF, Karr TL, Long M. 2009. Stage-specific expression profiling of Drosophila spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. *PLoS Genet.* 5:e1000731.

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the *Drosophila* genus. *Genome Res.* 19:897–903.

Vicoso B, Charlesworth B. 2006. Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet.* 7:645–653.

Wang Y, Leung FC. 2009. A study on genomic distribution and sequence features of human long inverted repeats reveals species-specific intronic inverted repeats. *FEBS J.* 276:1986–1998.

Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.* 14:1861–1869.

Wilson RJ, Goodman JL, Strelets VB. 2008. FlyBase: integration and improvements to query tools. *Nucleic Acids Res.* 36: D588–D593.

Yang S, Arguello JR, Li X, et al. (12 co-authors). 2008. Repetitive element-mediated recombination as a mechanism for new gene origination in *Drosophila*. *PLoS Genet.* 4:e3.

Zamore PD, Tuschl T, Sharp PA, Bartel DP. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 101:25–33.

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res.* 20:1526–1533.

Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* 8:e1000494.

Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res.* 9:133–148.