Published in final edited form as:

J Mol Biol. 2015 January 30; 427(2): 521–536. doi:10.1016/j.jmb.2014.11.008.

Bidirectional promoters of insects: genome-wide comparison, evolutionary implication and influence on gene expression

Susanta K. Behura* and David W. Severson

Eck Institute for Global Health and Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA

Abstract

Bidirectional promoters are widespread in insect genomes. By analyzing 23 insect species, we show that the frequency of bidirectional gene pairs varies according to genome compactness and density of genes among the genomes. The expected density of bidirectional genes based on number of genes per megabase of genome explains the observed density indicating that bidirectional pairing of genes may be due to random event. We identified specific transcription factor binding motifs that are enriched in bidirectional promoters across insect species. Furthermore, we observed that bidirectional promoters may act as transcriptional hotspots in insect genomes where protein coding genes tend to aggregate around bidirectional promoters in significantly biased ($p < 0.001$) manner compared to unidirectional promoters. Natural selection seems to have an association with the extent of bidirectionality of genes among the species. The rate of non-synonymous to synonymous (dN/dS) changes shows a second-order polynomial distribution with bidirectionality between species indicating that bidirectionality is dependent upon evolutionary pressure acting on the genomes. Analysis of genome-wide microarray expression data of multiple insect species suggested that bidirectionality have a similar association with transcriptome variation across species. Furthermore, bidirectional promoters show significant association with correlated expression of the divergent gene pairs depending upon their motif composition. Analysis of gene ontology further showed that bidirectional genes tend to have a common association with ‘binding’ (including ion binding, nucleotide binding and protein binding) related functions across genomes. Such functional constraint of bidirectional genes may explain their widespread persistence in genome of diverse insect species.

Keywords

gene regulation; insect genome; divergent transcription; genome structure; *cis*-regulatory motif

© 2014 Elsevier Ltd. All rights reserved.

*Correspondence: Susanta K. Behura, Phone: 574-904-2794, FAX: 574-631-7413, sbehura@nd.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

When two genes are localized within 1 kb distance but arranged on opposite strands with their 5' ends close to each other, they are called bidirectional gene pairs, and the intervening DNA that serves as the common upstream sequence of these gene pairs is referred to as the bidirectional promoter [1–2]. Bidirectional gene arrangement is a common feature of eukaryote genomes [3–4]. Nearly one in every eleven genes in the human genome is associated with bidirectional promoters [2]. In human, bidirectional promoters are associated more often with DNA repair and oncogenes than normal genes [1, 5]. Bidirectional promoters have been identified in several mammalian genomes [6] as well as in some lower eukaryotes [4]. Although it is not known why bidirectional gene arrangements are widespread in eukaryotes, functional constraints on genome seem to play a role in preserving such structures across species [4].

During bidirectional transcription of genes, the promoter activity is regulated at multiple points giving rise to diverse types of transcripts [7]. The strand-specific sequence composition of bidirectional promoters is critical to binding of appropriate regulatory proteins [8–9]. Thus, rational design of bidirectional promoters may be implemented, for example, in improving stability of synthetic genetic circuits [10]. Moreover, natural selection of bidirectional promoters acts as an important driver in modulating transcriptomes in a species-specific manner [11]. Hence, understanding structure, function and evolution of bidirectional promoters is important to better understand evolution of gene regulation among organisms. In insects, although a great deal of information has been obtained on architecture of promoters of the model organism *D. melanogaster* [12–14], genome-scale studies of promoter sequences in non-model insects are very limited [15–17]. Particularly, information on bidirectional promoters in insect genomes, other than *D. melanogaster*, is non-existent. This is a major gap in knowledge for appreciating the diversity and evolution of gene expressions in insects at a time when increasing interests are focused on genomics [18–19] and transcriptomics [20] of various insects. The major aims of this study are to 1) identify bidirectional promoters of coding genes in diverse insects, 2) characterize sequence composition of bidirectional promoters, 3) understand evolution of bidirectional arrangement of genes among species and 4) investigate how bidirectionality may influence gene expression variation in different species.

Results

Frequency of bidirectional genes in insects

The number of bidirectional gene pairs varies from a few hundred to more than a thousand among the 23 insects. They represent as low as 1.9% to as much as 25.6% of all the coding genes annotated from the genome assemblies of these insects (Table 1). Although these estimates account for bidirectional genes that are protein coding, our data also shows that in some cases specific non-coding genes are localized in a bidirectional manner with protein coding genes. However, such gene pairs are significantly lower in genomic abundance as compared to that of coding gene pairs (Figure 1). Figure 1 further shows that the drosophiloid species generally have higher abundance of bidirectional gene pairs compared to the non-drosophiloid insects. To determine if the observed variation of bidirectional genes

in insect genomes is due to random or non-random process, we calculated the expected density of bidirectional gene pairs in each species (Figure 1). The expected density was calculated based on the genome size, frequency and average length of coding and non-coding genes localized in either plus or minus strand in each genome (Supplementary Table 1). By comparing the observed and the expected variation of bidirectional gene pairs among these species, it was found that when higher number of genes are localized in a genome in opposite orientation, the expected density of bidirectional genes also increases. This pattern was clearly reflected in the observed density of bidirectional genes across species (Figure 1) suggesting that bidirectional arrangement of genes may be due to random chance.

Our data also indicates that gene duplication may influence the number of bidirectional genes in some species. We found gene pairs whose duplicated copies are also localized in bidirectional positions in the genome (Supplementary Table 2). The frequency of duplication of bidirectional genes was variable among species. To know if bidirectional genes are differentially distributed within genomes, we also determined the number of bidirectional genes in individual chromosomes of specific insect species (total five) where genome sequences have been assigned to chromosomes. The total number of coding genes in the corresponding chromosomes were also determined for comparison with number of bidirectional genes (Table 2). We observed that, in each species, the proportion of bidirectional genes to the total number of genes of X chromosome is always lower than the proportion of genes that are bidirectional in all the autosomes. - However, significant differences in number of bidirectional genes between sex chromosome and autosomes were observed only in some species but not all (Table 2).

Motif composition of bidirectional promoters

A total of 6,216 motifs (16.7 ± 0.03 bp in length) were identified in the sequences of bidirectional promoters of the insect species. Each motif was highly conserved among multiple bidirectional promoters ($n = 7$) either within or between species. To know if these conserved motifs represented the binding sites of eukaryotic transcription factors (TF), the known *cis*-regulatory motifs curated in JASPAR database were searched. It was found that majority ($n = 5,141$) of the motifs in bidirectional promoters are novel (Supplementary Table 3). The characteristic features of these novel motifs are low GC content (average value = 17.7%), low sequence palindromes (only 1.2% of motifs contain such sequences) and low TATA box sequences (0.09%). Apart from these novel motifs, we also identified motifs representing known binding sites of 36 different transcription factors (Supplementary Table 4). Based on Clover analysis [21], we further observed that motifs of six of these transcription factors were enriched in the bidirectional promoters in most of the insects (Figure 2). These motifs are associated with transcription factors hunchback (*Hb*), forkhead box protein A1 (*Foxa1*), chorion factor 2-ii (*Cf2-ii*), LIM homeodomain transcription factor3 (*Lhx3*), myocyte enhancer factor-2a (*Mef2a*) and Fork head domain transcription factor *slp1*. Although these transcription factor (TF) binding motifs represented only about 17% of all the motifs identified in the bidirectional promoters, they are found localized closest to transcription start sites (TSS) of majority of the bidirectional gene pairs across species (Figure 3). In fact, ~70% of the bidirectional promoters had the TF-binding motifs closest to the TSS of the bidirectional genes. On the other hand, the bidirectional promoters

where the closest motifs to the TSS of genes were unknown (novel motifs) represented no more than 5% of all the bidirectional promoters. The other ~25% of the promoters showed combinations of both TF binding motif and novel motif as the closest motifs of TSS of the bidirectional genes (Figure 3). It was further observed that the TSS-proximal motifs are often oriented in opposite orientation. In *D. melanogaster*, we found 130 bidirectional promoters in which TSS-proximal motifs are oriented in opposite orientation similar to that of the bidirectional gene pairs. Such patterns were also observed in bidirectional promoters of other species as well (for example, 92 in *T. castaneum*, 37 in *P. humanus* and 11 in *A. aegypti*). Such strand-specific arrangement of motif sequences may play a role in appropriate binding of protein factors for regulating the expression bidirectional gene pairs as suggested from other work [8].

Aggregation of protein coding genes around bidirectional promoters

We observed that multiple protein coding genes are often positionally aggregated around bidirectional promoters. A gene aggregation was defined as the cluster of coding genes around bidirectional promoter wherein the intergenic distance between neighboring genes within the cluster was always less than 1kb. To determine if gene aggregation is a characteristic feature of bidirectional gene pairs or not, we compared the aggregation frequency around bidirectional gene pairs with that of unidirectional gene pairs in genome-wide manner in each species (Table 3). Similar to aggregation around bidirectional promoters, a gene aggregation around unidirectional promoter was defined as the cluster of coding genes around unidirectional gene pair wherein each intergenic distances within the cluster was less than 1 kb. The unidirectional gene pairs in this comparison represented the gene pairs that are localized within 1 kb distance similar to that of bidirectional genes but are localized in same orientation (both the genes are either in plus strand or minus strand) in each genome. The frequency of aggregation around the bidirectional and unidirectional gene pairs was compared by hypergeometric tests, the results of which revealed a significant bias ($p < 0.001$) of gene aggregation towards bidirectionality. In fact, nearly a 3-fold higher amount of bidirectional gene pairs is associated with gene aggregations compared to the unidirectional gene pairs (Table 3). Such bias aggregation of protein coding genes around bidirectional gene pairs suggested that bidirectional promoters may act as hotspots of gene activity (transcription) of genomes.

Evolution of bidirectional gene arrangement in insects

To understand evolution of bidirectional gene arrangement between species, the 1:1 orthology of *D. melanogaster* bidirectional gene pairs was identified in each of the other 22 insects. We chose *D. melanogaster* genes for identification of orthologs in other insects because *D. melanogaster* genome and gene models have been better annotated as a model organism compared to other insect genomes. To determine if bidirectional pairing of genes is dependent or independent of evolutionary pressure on genomes, we compared rate at which the orthologous genes retain bidirectionality to the average ratio of non-synonymous to synonymous changes (dN/dS) of the orthologous genes in a genome relative to *D. melanogaster*. Data in Table 4 shows the estimates of relative rate of gene bidirectionality expressed as the proportion of orthologous genes that retain bidirectional structure between the two species. When this rate of bidirectionality was compared with the average rate of

non-synonymous to synonymous mutations of the orthologous genes between species, a significant correlation (Spearman correlation = 0.741, two-tailed $p = 0.000079$) was observed (Figure 4). This suggested that bidirectional arrangement of genes is dependent on evolutionary pressure acting on genome of these insects. The increase in bidirectionality with increase in dN/dS further indicated that genesis of bidirectional gene pairs may be a random evolutionary process. However, change in dN/dS rate with that of bidirectionality shows a second-order polynomial distribution rather than a linear relationship suggesting that evolutionary selection may not be the only factor that dictate bidirectionality in insect genomes. Our data showed that evolution of genome size and the density of gene distribution may also influence bidirectionality in insects. The fold change in genome size and gene density of different insects relative to that of *D. melanogaster* were calculated and compared with bidirectionality (Figure 5). While bidirectionality tends to decrease with increase in genome size, it increases with increase in gene density of the genomes. However, they showed a logarithmic distribution pattern indicating that bidirectionality doesn't increase beyond certain fold change of gene density.

To further understand evolutionary pattern of bidirectional genes among fruit flies, we resorted to comparison of bidirectional genes among the 12 *Drosophila* species. By identifying single copy orthologs of bidirectional gene pairs among the 12 *Drosophila* genomes (using the database OrthoDB: <http://orthodb.org/orthodb7>), it was observed that several bidirectional gene pairs were species-specific in fruit flies (Table 5). However, they represented no more than 12% of the total number of bidirectional gene pairs of the respective genomes suggesting that majority of the bidirectional genes retain orthology between species. On the other hand, only few bidirectional gene pairs ($n=14$) showed orthology in 10 or more fruit fly species (Supplementary Figure 1). In fact, it was observed that only one gene pair has orthologous copies in all the 12 species (OrthoDB group ID: EOG7Q8R71 + EOG73G36V). In *D. melanogaster*, that represented the gene pair FBgn0031457 + FBgn0031458. In fact, for any given species of these fruit flies, nearly two-thirds of the bidirectional gene pairs showed lack of orthology in another species (Figure 6A). We also identified bidirectional genes where either one or both of the partner genes showed lack of orthology between species. It was observed that these patterns of evolutionary conservation of bidirectional gene pairs vary in highly correlated manner across the genomes (Figure 6B) suggesting that a common mechanism may account for evolution of bidirectional genes between species. We also asked if the bidirectional gene pairs whose orthologs are widespread in *Drosophila* are associated with specific functions unlike the gene pairs that show lack of orthology in another species. By predicting the gene ontology (GO) of these gene groups, we however observed that no specific gene functions were significantly overrepresented in one *versus* the other (data not shown).

Influence of upstream motifs on bidirectional localization of genes

We observed that the loss or gain of motifs in the bidirectional promoters is linked to the retention or rearrangement of bidirectional localization of genes between species. We identified the 1:1 orthologs of *D. melanogaster* bidirectional gene pairs that either retained the bidirectional structure or relocated to unidirectional positions in the genome of the other 22 species. Then we analyzed the upstream motif sequences in the bidirectional promoters

of these gene pairs. In each of these 22 species, we observed that genes retaining the bidirectional structures had significantly ($p = 0.00002$) fewer numbers of upstream motifs compared to that of the genes that are rearranged to unidirectional position (Figure 7). To determine if motif numbers of bidirectional promoters are different than that of unidirectional promoters within genome, we performed genome-wide analysis of motifs of all coding genes in each species. If we assume that the motifs were distributed randomly in the upstream (1 kb) region of all genes irrespective of whether they are bidirectional or unidirectional, then the expected number of motifs in the bidirectional promoters can be determined from the total number of upstream motifs in the genome and the proportion of genes that are bidirectional. To determine these expected values, we counted the number of upstream (1 kb) motifs of all the protein coding genes. We chose 1 kb upstream regions as this is the limit of the intergenic distances between bidirectional gene pairs. When the expected numbers were compared with the observed number of motifs of bidirectional promoters of each species, the observed number of motifs was consistently lower than the expected values (Supplementary Figure 2). This figure also shows that bidirectional genes have lower number of upstream motifs compared to the unidirectional genes in each genome suggesting that bidirectional and unidirectional genes have differential property in maintaining the motif sequences in the promoter regions. We also find that specific transcription factor binding motifs such as PBX (Pre-B leukemia transcription factor 2), MEF2A (myocyte enhancer factor 2A) and Esrrb (Estrogen related receptor beta) are relatively more prevalent (at least 2-fold) in bidirectional promoters compared to unidirectional promoters in all the species. It is likely that evolutionary rearrangement of genes from unidirectional to bidirectional or *vice versa* is likely to have an association with accumulation or erosion of these upstream conserved motifs between species.

Mutation in bidirectional promoters

We analyzed sequence variation of bidirectional promoters among multiple inbred lines ($n=29$) of the *Drosophila* Genetic Resource Panel (DGRP). The bidirectional promoters associated with the gene pairs FBgn0029736-FBgn0029737 and FBgn0004656-FBgn0004657 were chosen for this purpose. These specific gene pairs were chosen as they show differences in retaining bidirectional structure among orthologs. While the orthologs of gene pairs FBgn0029736-FBgn0029737 retain bidirectional structure in *D. simulans* and *D. grimshawi*, the orthologs of FBgn0004656-FBgn0004657 are rearranged to unidirectional positions in these species. It was observed that the bidirectional promoter of FBgn0004656-FBgn0004657 carries a higher number of mutations than that of FBgn0029736-FBgn0029737 (Table 6). The phylogenetic analysis also showed that FBgn0004656-FBgn0004657 promoters are more variable in sequences than those of FBgn0029736-FBgn0029737 among the inbred lines (Supplementary Figure 3). It is likely that these mutational bias may have a role in accumulation or erosion of these the upstream motifs of bidirectional genes as shown in Figure 7.

Relevance of bidirectional promoters to gene expression and function

We investigated expression patterns of bidirectional gene pairs by analyzing published microarray data of different species. In *D. melanogaster*, we analyzed microarray data (GEO# GDS3639) of males and females of different strains [21]. The data was generated by

Baker et al [21] that investigated sex biased gene expression in *D. melanogaster*. Using the expression data of this study, we observed that expression variation between males and females is significantly associated with promoters of the genes (i.e. bidirectional or unidirectional). The genes showing sex biased expression differ significantly in association with bidirectionality compared to genes showing similar expression variation between males and females (Supplementary Figure 4). Furthermore, bidirectional genes differ significantly from the control genes (unidirectional) in showing either female biased or male biased expression in *D. melanogaster*. While the control genes are predominantly associated with male biased expression (~ 3 times more number of genes showing female biased expression), the bidirectional genes are associated with both male biased as well as female biased expression in nearly equal proportions (Supplementary Figure 4).

The expression analysis further revealed that many bidirectional gene pairs show correlated expression changes between males and females in *D. melanogaster* (Figure 8). We observed both positive and negative correlation in expression changes among different number of bidirectional gene pairs. The number of bidirectional gene pairs showing positive correlation in expression changes between males and females were ~ 12-fold more than the bidirectional gene pairs showing anti-correlated expressions. Furthermore, it was observed that different motif sequences are associated with the bidirectional promoters of these two types of gene pairs. The *hunchback*, *spl1*, *Lhx3* and *MEF2A* binding motifs were significantly (hypergeometric test, $p < 0.05$) associated with promoters of gene pairs showing correlated expression changes whereas *SRY*, *ABI4*, *Foxd3* and *Myc* motifs were significantly associated with promoters of gene pairs showing anti-correlated expression (data not shown). It is interesting to note that *hunchback*, *spl1*, *Lhx3* and *MEF2A* motifs, unlike *SRY*, *ABI4*, *Foxd3* and *Myc* motifs, are found in high abundance in bidirectional promoters of all insect genomes (Figure 2). It is a possibility that these specific binding motifs have a role to play in differential (positively or negatively correlated) expression of the bidirectional gene pairs.

Then we asked whether such co-regulation or anti-regulation of gene expression has any relationship with the positional aggregation of genes around bidirectional promoters (Table 3). To know that we identified bidirectional gene pairs where gene aggregation either occurred or didn't occur, and then determined if they are associated with co-regulated or anti-regulated expressions (Supplementary Figure 5). The result of the 2x2 contingency tests of number of co-regulated or anti-regulated gene pairs with or without gene aggregation showed that there is a significant relationship ($\chi^2 = 74.28$, $p = 0.001$) between expression and positional aggregation of genes around bidirectional promoters. While lack of gene aggregation was biased towards co-regulation of expression, the aggregation and non-aggregation patterns represented almost equally for anti-regulated gene pairs.

We further observed that low or high transcriptome variation shows a similar relationship with bidirectionality of genes even among unrelated insects such as fruit fly (*D. melanogaster*), malaria mosquito (*A. gambiae*) and honey bee (*A. mellifera*). For comparison of expression variation, we determined genes that show either low or high expression variation in males versus females of *D. melanogaster*, developmental times of *A. gambiae*, and queen versus worked destined larvae of *A. mellifera*. In *A. gambiae*, we analyzed

microarray data (GEO# GDS3844) generated for different time points (total 56 time points varying from 2hr to 52hr) during embryonic development [22]. In *A. mellifera*, we analyzed the array data (GEO# GSE52291) generated to study variation of developmental differences in gene expressions between worker-destined and queen-destined bees [23]. In each case, we consistently observed significant association of low/high expression variation with bidirectional/unidirectional gene arrangement (Figure 9). The ratio of bidirectional genes associated with low expression variation to genes associated with high expression variation was 14.3 for *D. melanogaster*, 12.5 for *A. gambiae* and 4.2 for *A. mellifera*. But for the unidirectional genes, these ratios were always lower (9.5 for *D. melanogaster*, 6.8 for *A. gambiae* and 1.5 for *A. mellifera*). Also, the ratios of unidirectional genes to bidirectional genes associated with high expression variation were 5.1 for *D. melanogaster*, 12.0 for *A. gambiae* and 14.2 for *A. mellifera* whereas the corresponding ratios for low expression variation were 3.3 for *D. melanogaster*, 6.5 for *A. gambiae* and 5.3 for *A. mellifera*. These results clearly suggest that bidirectional genes are biased toward low expression variation across the three species. The consistent patterns across these unrelated species further suggest that bidirectional promoters may have similar association with regulation of gene expression variation (low variation) across the genomes. Besides gene expression analysis, we also analyzed gene ontology (GO) of bidirectional genes to further understand function of bidirectional genes in different species. Our results show that although many GO terms showed differential association with bidirectional genes in different species, specific GO terms are commonly enriched (hypergeometric $p < 0.05$) in bidirectional genes across the species. Among the top three ranking GO terms associated with bidirectional genes, the binding activity such as ion binding, nucleotide binding and protein binding are largely common across the species (Supplementary Table 5).

Discussion

The present study provides insights into the characteristics of bidirectional promoters in diverse insect species. The 23 insect species compared in this investigation represented 5 different orders, thus providing comparative genomic analysis of bidirectional genes in a broad phylogenetic context. The frequency and distribution of bidirectional gene pairs reveal wide variation among these species. Although the numbers of protein coding genes predicted from genome assemblies of these insects are very similar (average $15,800 \pm \text{S.E. } 1,000$), the proportion of genes that are localized in bidirectional manner is highly variable (Table 1). The total number of genes doesn't explain the variation of number of bidirectional genes among genomes. To give an example, the pea aphid genome is annotated to have 33,816 coding genes that is more than 3-fold higher the number of coding genes predicted from the human louse genome, yet both have nearly the same number of bidirectional genes (Table 1). Our analysis showed that it is not the total number of genes but the gene density of a genome that explains the variation of genomic abundance of bidirectional genes among species (Figure 1). The random expectation of bidirectional gene frequency based on density of genes in the genome correlates with the observed variation of bidirectional genes among the 23 species. This further suggests that bidirectional localization of genes is due a random event. Gene density of these insects varies largely due to changes in genome size as the protein coding genes are highly comparable in number among the species. The malaria

mosquito *A. gambiae* and dengue mosquito *A. aegypti* genomes have 12, 810 and 15, 998 coding genes respectively but their genome size differs more than 5-fold (273.1 Mb genome of *A. gambiae* versus 1383.9 Mb genome of *A. aegypti*). Hence, the per Mb density of the protein coding genes varies substantially in their genome that in turn explains higher density of bidirectional gene pairs in *A. gambiae* compared to *A. aegypti* (Figure 1). Although it is a random chance that genes are localized within 1 kb in divergent manner in genome, we observed that a specific length of intergenic distance (~300–600 bp) is associated with majority of the bidirectional promoters, and this pattern is consistent across genomes (data not shown). If these specific lengths are optimal for bidirectional transcription of these genes pairs, although a possibility, is not known from this study.

A previous study indicated that genome size and gene counts of eukaryotes show a common relationship wherein the gene counts increase with genome size in logarithmic scale [24]. Although, various factors such as body size, developmental rate, organ complexity as well as cell size, cell division and metabolic rates may confound the gene number and genome size relationship [25], we observed that insects also show a positive logarithmic relationship between gene numbers and genome size [genome size $\sim 0.9 + 1.5 \times \log(\text{gene counts})$]. But, when the number of bidirectional genes were compared with genome size of insects, a negative relationship [genome size $\sim 18.7 - 7.2 \times \log(\text{bidirectional genes})$] was observed. This is in agreement with our random model of genesis of bidirectional genes where genome size has a negative relationship of gene density of these insects. To put this into the context of evolution of bidirectional genes, the 1:1 orthologous copies of bidirectional gene pairs of *D. melanogaster* in different insect species were analyzed. We quantified ‘bidirectionality’ as the proportion of orthologs that retain bidirectional structure between genomes. This is a relative measurement (relative to *D. melanogaster*) that is appropriate for the comparative analysis between species. When we compared the relative fold change of genome size of each of the 22 insect species compared to that of *D. melanogaster*, a negative logarithmic relationship was observed between the relative fold changes in genome size and their corresponding bidirectionality values (Figure 5). On the other hand, when bidirectionality was plotted against gene density, a positive logarithmic relationship was observed (Figure 5). These results suggested that evolution of genome compactness may be the critical factor that modulates bidirectionality among these insect species.

Additionally, we discovered that bidirectional gene pairs follow a similar pattern in retaining or losing orthology among species (Figure 6). By comparing orthology of bidirectional genes among the 12 *Drosophila* species, it was observed that the number of gene pairs where both genes retain orthology varies in correlation with the number of gene pairs where only one gene was orthologous between species. This suggested that evolution of bidirectional genes may be governed by a common mechanism across the species. Furthermore, by analyzing sequences of specific bidirectional promoters among multiple inbred lines of *D. melanogaster*, we observed that the promoter sequences have reduced numbers of mutations if the bidirectional structure is evolutionary stable (orthologs are also bidirectional). Higher number of mutations was observed in the promoter where the orthologs were rearranged between species. This suggested that mutations in the flanking

sequences of genes are critical in determining evolutionary stability of bidirectional arrangements of genes.

The coupling of transcriptional regulation of coding genes with neighboring non-coding RNA genes have been documented in many eukaryotes [26]. The results from our current study show that although several coding and non-coding gene pairs are arranged in bidirectional manner, their frequency is exceptionally low in all the insects (Figure 1). It is known that regulation of bidirectional transcription can be controlled by various factors such as chromatin modification, DNA looping, three dimensional structure of transcription, flanking sequences of genes and others [7]. All these factors are likely to influence the functional consequences of bidirectional genes among different species. It is also known that variation in gene expression follows a power-law distribution where expression changes are proportional to expression level of genes, a pattern that is universal across all organisms [27]. Our results from current study showed that bidirectional promoters influence gene expression in similar pattern across species (Figure 9). It is a possibility that bidirectional promoters have an evolutionary link with the universality of gene expression pattern across species.

We further observed that multiple coding genes are often aggregated around bidirectional promoters in insect genomes. This seems to have an underlying evolutionary cause/consequence of bidirectionality as they are significantly biased towards bidirectional promoters over the unidirectional promoters in each species. Furthermore, these bidirectional gene pairs show a significant bias towards correlated expression changes as evident from microarray data of *D. melanogaster* (Supplementary Figure 5). Our analysis shows that the bidirectional gene pairs showing correlated expression changes contain specific transcription factor binding sites in their promoter sequences which are not found in the promoters of gene pairs showing anti-regulated expressions. It is likely that aggregation of multiple protein coding genes around bidirectional promoters may be associated with correlated transcriptional regulation of genes by binding of appropriate regulatory proteins to these specific motifs of the bidirectional promoters.

We identified gene ontologies (GO) associated with the bidirectional genes and observed that different GO terms such as protein binding, structural molecules, oxidation-reduction process, ion binding, catalytic activity, peptidase activity, membrane activity, reproduction, hydrolase activity, transferase activity, nucleotide binding and several housekeeping and cellular metabolic activities were significantly overrepresented (hypergeometric $p < 0.05$) among the bidirectional genes of these insects. The basic housekeeping and cellular metabolic genes are also known to be regulated by bidirectional promoters in human [2]. In *D. melanogaster*, we observed that 'reproduction' was the most frequent GO term associated with the bidirectional genes (Supplementary Table 5). This prediction is consistent with our analysis of gene expression data showing that bidirectional genes have a strong association with the sex-biased expression changes in *D. melanogaster*. In spite of the fact that different species show differential enrichment of GO terms for the bidirectional genes, 'binding' related gene functions such as ion binding, nucleotide binding and protein binding were commonly found among the top three ranking functions of the bidirectional genes among different insects (Supplementary Table 5). It suggested that this specific functional

association of bidirectional genes may be evolutionary conserved across these insects. Considering that maintenance of bidirectional genes in eukaryotes is associated with functional constraints [4], it is likely that association of bidirectional genes with 'binding' functions in insects may be an important factor for maintenance of such gene pairs across phylogenetically diverse insects.

In conclusion, our study shows that bidirectional promoters are widespread in insects and act as regulatory regions around which protein coding genes tend to aggregate in a significantly biased manner in the genome. Moreover, the results further show that bidirectional promoters play a universal role in controlling transcriptomes across the insect species. The study will prompt investigations to further elucidate role of bidirectional promoters in functional evolution of insect genomes.

Materials and Methods

Identification of bidirectional gene pairs

The bidirectional gene pairs were identified from the official genes annotated from the genome sequences of different insects. The insect species, genome assembly and gene build versions that were used in this investigation are as follows: *Atta cephalotes* (Attacep1.0: 2013-08-AntGenomes Portal), *Acyrtosiphon pisum* (Acyr_2.0: 2013-07-AphidBase), *Aedes aegypti* (AegL2: 2013-12-VectorBase), *Anopheles darlingi* (AdarC1: 2012-06-LNCC), *Anopheles gambiae* (AgamP3: 2012-10-VectorBase), *Apis mellifera* (Amel4.0: 2013-07-BeeBase), *Bombyx mori* (ASM15162v1: 2010-12-SilkDB), *Culex quinquefasciatus* (CpipJ1: 2013-02-VectorBase), *Drosophila ananassae* (dana_caf1: 2008-07-FlyBase), *Drosophila erecta* (dere_caf1: 2008-07-FlyBase), *Drosophila grimshawi* (dgri_caf1: 2008-07-FlyBase), *Drosophila melanogaster* (BDGP 5: 2012-04-FlyBase), *Drosophila mojavensis* (dmoj_caf1: 2008-07-FlyBase), *Drosophila persimilis* (dper_caf1: 2008-07-FlyBase), *Drosophila pseudoobscura* (BCM-HGSC 2: 2012-06-FlyBase), *Drosophila sechellia* (dsec_caf1: 2008-07-FlyBase), *Drosophila simulans* (WUGSC 1: 2012-03-FlyBase), *Drosophila virilis* (dvir_caf1: 2008-07-FlyBase), *Drosophila willistoni* (dwil_caf1: 2008-07-FlyBase), *Drosophila yakuba* (dyak_r1.3_FB2008_07: 2008-07-FlyBase), *Nasonia vitripennis* (Nvit_2.1: 2010-12-NasoniaBase), *Pediculus humanus* (PhumU1: 2009-07-VectorBase) and *Tribolium castaneum* (Tcas3: 2012-09-24). The gene annotation features (chromosome or scaffold name, transcription start coordinate, orientation strand and gene biotype) in each of these genome assemblies were obtained from the *EnsemblMetazoa* database. They were then used to identify gene pairs that are localized within 1 kb distance but in opposite orientation to each other in individual genome.

Gene density and bidirectional gene pairs

The gene density of each species was calculated as the number of genes per megabase (Mb) of genome size. To calculate the expected bidirectional gene pairs between coding-coding and coding-noncoding genes, first the total number of genes (coding or non-coding) localized in either orientation (plus strand or minus strand) was determined in genome-wide manner in each species. The average length and gene counts were used to calculate their occurrence per Mb of genome size. Then the expected density of bidirectional gene pairs

was determined from the joint probability of occurrence of coding/coding or coding/non-coding genes in a genomic region that is sum of their average lengths plus 1 kb (the maximum distance between bidirectional gene pairs).

To compare bidirectional gene pairs between sex chromosome and autosomes, we choose five selected species (*A. gambiae*, *D. melanogaster*, *D. pseudoobscura*, *D. simulans* and *D. yakuba*) where genome sequences have been assigned to chromosomes. The number of bidirectional genes as well as the total number of genes were determined for each chromosome of these species. To know if bidirectional genes have biased representation in sex chromosome, hypergeometric test was performed with the number of bidirectional gene in sex chromosome, total number of bidirectional genes in the autosomes, total number of genes in sex chromosome and the total number of genes in the autosomes in each species.

Analysis of sequence homologies among bidirectional promoters

The sequences of bidirectional promoters were analyzed to determine if they contain motifs that are highly conserved within and between species. The sequences of all bidirectional promoters from the 23 species were included together to make a local database for blastn analysis. The blastn analysis was performed with each bidirectional promoter as query sequence. The output of the blast search was then used to identify sequence homologies present in multiple bidirectional promoters both within and between species. When the sequence of a bidirectional promoter showed homologies to multiple bidirectional promoters within genome, we separated those hits based on two criteria: 1) the query and subject had more than 90% sequence identity along with an alignment length higher than 90% of both query and subject sequence lengths, and 2) the gene pairs associated with those homologous sequences are annotated as paralogs in that species (annotations downloaded from *BiomartMetazoa*). Those sequences satisfied these two criteria represented different duplications of bidirectional promoters. When the homologies were short in length (cutoff value 50 bp), we wanted to know if that sequence was present in multiple bidirectional promoters (at least 7) either within or between species. They represented the motif sequences of bidirectional promoters those were further investigated throughout this study. Based on the multiple alignment of motif sequences of different bidirectional promoters, the consensus sequences and the position frequencies of A, T, C and G were determined for each motif.

Enrichment of *cis*-regulatory motifs in bidirectional promoters

To know if bidirectional promoters are enriched with specific transcription factor binding sites, we used the JASPAR core database (downloaded in October 2013; <http://jaspar.genereg.net/>) that curates experimentally defined transcription factor binding sites for eukaryotes. The 'Clover' (Cis-eLement OVERrepresentation) tool [28] was used to determine if any of the known transcription factor binding motifs obtained from JASPAR were overrepresented in the bidirectional promoters of each species. The Clover algorithm generates motif incidences (homologies) in the given sequences. Based on those homologies, we determined if any of the JASPAR motifs matched to the motifs identified from our blastn analysis. If no homology was found, they were referred to as 'novel' motifs. For overrepresentation analysis, the Clover algorithm calculated a raw score that quantifies

the degree of motif incidences in the sequences. This score is calculated based on repeated averaging of likelihood ratios of motif incidence in the promoter sequence. In the second step, it estimates the p -value for obtaining a raw score that is either equal or greater than the threshold (user defined) by chance. The Clover algorithm has different approaches for calculation of p -values including nucleotide randomization and background sequences [28]. We used both these approaches to determine consensus p -values for overrepresentation of motifs. The background sequences, in our study, were randomly selected from the sequences representing 1 kb upstream regions of all the protein coding genes in the genome. These upstream sequences were first shuffled and then extracted randomly (using the *R* code available at <http://tata-box-blog.blogspot.com/>) to generate sequences of total size 10 times the total length of the bidirectional promoters in the genome. The default values of threshold score of 6 and that of significance ($p < 0.01$) were used to identify enriched motifs in the bidirectional promoters of each genome.

Orthology of bidirectional gene pairs and mutation in the common upstream sequences

As bidirectional genes are known to have association with selective pressure that favor synonymous changes over non-synonymous mutations [29], we used the average rate of non-synonymous to synonymous mutations (dN/dS) of orthologous genes between species to compare with bidirectionality. This comparison was performed to determine if bidirectionality is a random or non-random process relative to selective pressure on the genomes. Bidirectionality was calculated as the relative rate of evolutionary conservation of bidirectional genes. It was measured as the ratio of number of orthologous genes that were localized in bidirectional manner in both species relative to the total number of 1:1 orthologous genes between the two species. The 1:1 orthologies of *D. melanogaster* bidirectional gene pairs were identified in each of the other 22 insects using the *BiomartMetazoa* homology database. The dN and dS values of each ortholog pair were also obtained from *BiomartMetazoa* database. They were used to calculate the average dN/dS of orthologs of individual species relative to *D. melanogaster*. Bidirectionality of each species was also calculated relative to *D. melanogaster*. Apart from comparing bidirectionality with average dN/dS values of species, we also compared it with relative fold change of genome size and gene density of each species relative to *D. melanogaster*.

We also identified 'ortholog clusters' (orthologs that are present in multiple insects) from the OrthoDB database [30] to compare orthology of bidirectional gene pairs among the 12 *Drosophila* species. To investigate the polymorphism patterns in bidirectional promoters, we analyzed genome sequences of *D. melanogaster* inbred lines generated from the DGRP project [31]. We restricted this analysis to promoters of bidirectional gene pairs of chromosome X from 29 randomly selected inbred lines. The phylogenetic analyses of the bidirectional promoter sequences among these inbred lines were conducted using the Neighbor-Joining method implemented in MEGA4 [32]. The evolutionary distances, in the units of the number of base substitutions per site, were computed using the maximum composite likelihood method also implemented in MEGA4. All the sequence polymorphism analyses including calculation of total number of mutations, number of polymorphic sites, average number of nucleotide differences, and number of fixed mutations between sequences were conducted by DnaSP v 5.10 [33].

Analysis of gene aggregation around bidirectional promoters

We wanted to know if gene aggregation was a characteristic feature of bidirectional promoters. To test the null assumption that there is no bias in gene aggregation around bidirectional promoters compared to unidirectional promoters, we compared the aggregations around bidirectional gene pairs with aggregations around unidirectional gene pairs (control) that are localized within 1 kb distance similar to bidirectional genes but are in unidirectional manner. Then the frequency of gene aggregations around bidirectional and unidirectional gene pairs were compared by 2×2 contingency tests to determine the statistical significance of rejecting the above null assumption.

Comparison of flanking motifs of bidirectional *versus* unidirectional genes

We wanted to know if there was any difference in the motif composition of upstream sequences of gene pairs that retained bidirectional arrangement between species compare to gene pairs that were bidirectional in one species but unidirectional in the other. The *D. melanogaster* bidirectional gene pairs were used for comparison of each species to know which gene pairs retained bidirectional arrangement and which gene pairs were rearranged to unidirectional positions. As the common upstream sequence of bidirectional gene pairs is limited to 1 kb length, we analyzed 1 kb upstream motif sequences of rearranged unidirectional genes in each species. The motif incidences of all the 6,216 identified motifs were determined by using Clover tool [28]. Two-tailed probability was inferred from Spearman rank order correlation from the motif incidences of genes that retained bidirectional structure *versus* those rearranged to unidirectional structure. Besides, we also determined the number of upstream motif sequences (1 kb) of all coding genes in each genome. Based on the total number of upstream motifs of all genes (coding) and the proportion of genes that are bidirectional in the genome, the number of motifs expected to be associated with bidirectional genes were calculated for each genome. The expected number of motifs was compared with the observed number of motifs of the bidirectional promoters among the 23 species. The genome-wide data was then used to calculate upstream motif counts of bidirectional *versus* unidirectional genes in each species. Based on the number of bidirectional and unidirectional genes and their upstream motifs in the genome, the specific motifs that were more prevalent (at least 2-fold) in bidirectional promoters compared to unidirectional promoters were also identified.

Analysis of gene expression and gene ontology

We analyzed microarray data of multiple species to study expression patterns of bidirectional gene pairs. We selected three specific expression datasets of *D. melanogaster* (GEO# GDS3639) [21], *A. gambiae* (GEO# GDS3844) [22] and *A. mellifera* (GEO# GSE52291) [23] that were generated by genome-wide microarray experiments. An analysis of gene expression of all the 23 insect species was not within the scope of this study.

In *D. melanogaster*, the ratio of male-average and female-average expression levels were determined for each gene across different strains of *D. melanogaster*. They were used to identify individual genes that had either similar expression (fold change < 0.1) or biased expression (greater than 2-fold differences in expression) between males and females. These categorization was made both for bidirectional as well as unidirectional genes. Then their

frequency was compared by Chi-square contingency test to determine statistical significance of association of gene expression variation and bidirectional arrangement of genes in *D. melanogaster*. The variation of gene expression of *A. gambiae* and *A. mellifera* was determined based on standard deviation of expression levels among the samples. The standard deviation values 0.98 and 0.42 were used for *A. gambiae* [22] and *A. mellifera* data [23], respectively, to define the low or high variation of expression changes among the samples. These specific values were chosen based on the mean value of standard deviation of expression changes of individual genes among the samples. After generating the count data of genes associated with low or high expression changes among the samples, the Chi-square contingency test was performed based to determine statistical significance of association between bidirectional localization of genes and variation in transcription of those species.

The correlation of gene expression among bidirectional gene pairs was performed by calculating Pearson correlation coefficient of expression changes between males and females in *D. melanogaster* using the same expression data (GEO# GDS3639). The bidirectional gene pairs associated with either positive correlation (correlation coefficient > 0.9) or negative correlation (correlation coefficient < -0.9) were further analyzed for association with gene aggregation. The bidirectional gene pairs associated with either positive or negative correlation expression patterns were categorized based on presence or absence of gene aggregation around those genes. Then Chi square contingency test was performed based on count statistics of those gene pairs to determine statistical significance of association between gene aggregation and gene expression.

The functions of genes associated with bidirectional promoters were predicted by gene ontology (GO) analysis. The gene ontology terms associated with all coding genes of the 23 insect species were obtained from *EnsembleMetazoa* gene database. These lists were then used to identify GO terms associated with bidirectional genes in the individual species. Among all the GO terms that were associated with the bidirectional genes, the rank orders of association frequency was calculated to identify the top three ranking GO terms. Additionally, statistical inferences on GO terms enrichment were made by hypergeometric tests. For this analysis, the frequency of each GO term associated with either the bidirectional genes only or all the protein coding genes were calculated for each genome. Then the total number of different GO terms associated with either the bidirectional genes only or all the protein coding genes were determined for individual genomes. These count data were then analyzed by hypergeometric tests to determine statistical significance of overrepresentation of individual GO terms among bidirectional genes in each species.

Statistical analysis

All the statistical analyses were performed using *R* [34]. The threshold of *p*-value was 0.05 for all tests unless stated otherwise.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by funding (RO1 AI079125-A1) from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, USA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors are thankful to Joanne Cunningham for reading the manuscript, and also two anonymous reviewers whose comments were useful in improving the manuscript.

References

1. Adachi N, Lieber MR. Bidirectional gene organization. *Cell*. 2002; 109:807–809. [PubMed: 12110178]
2. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. *Genome Res*. 2004; 14:62–66. [PubMed: 14707170]
3. Koyanagi KO, Hagiwara M, Itoh T, Gojobori T, Imanashi T. Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene*. 2005; 353:169–176. [PubMed: 15944140]
4. Xu C, Chen J, Shen B. The preservation of bidirectional promoter architecture in eukaryotes: what is the driving force? *BMC Syst Biol*. 2012; 6(Suppl 1):S21.
5. Shu J, Jelinek J, Chang H, Shen L, Qin T, Chen W, Oki Y, Issa JJ. Silencing of bidirectional promoters by DNA methylation in tumorigenesis. *Cancer Res*. 2006; 66:5077. [PubMed: 16707430]
6. Orekhova AS, Rubtsov PM. Bidirectional promoters in the transcription of mammalian genomes. *Biochemistry (Mosc)*. 2013; 78:335–341. [PubMed: 23590436]
7. Wei W, Pelechano V, Järvelin AI, Steinmetz LM. Functional consequences of bidirectional promoters. *Trends Genet*. 2011; 27:267–276. [PubMed: 21601935]
8. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. *Annu Rev Biochem*. 2003; 72:449–479. [PubMed: 12651739]
9. Engstrom PG, Suzuki H, Ninomiya N, Akalin A, Sessa L, et al. Complex loci in human and mouse genomes. *PLoS Genet*. 2006; 2:e47. [PubMed: 16683030]
10. Yang S, Sleight SC, Sauro HM. Rationally designed bidirectional promoter improves the evolutionary stability of synthetic genetic circuits. *Nucleic Acids Res*. 2013; 41:e33. [PubMed: 23093602]
11. Gotea V, Petrykowska HM, Elnitski L. Bidirectional promoters as important drivers for the emergence of species-specific transcripts. *PLoS One*. 2013; 8:e57323. [PubMed: 23460838]
12. Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol*. 2002; 3:RESEARCH0087. [PubMed: 12537576]
13. Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, et al. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res*. 2011; 21:182–192. [PubMed: 21177961]
14. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*. 2012; 13:233–245. [PubMed: 22392219]
15. Sinha S, Ling X, Whitfield CW, Zhai C, Robinson GE. Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees. *Proc Natl Acad Sci U S A*. 2006; 103:16352–16357. [PubMed: 17065326]
16. Sieglaff DH, Dunn WA, Xie XS, Megy K, Marinotti O, James AA. Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes. *Proc Natl Acad Sci U S A*. 2009; 106:3053–3058. [PubMed: 19211788]
17. Kim J, Cunningham R, James B, Wyder S, Gibson JD, Niehuis O, et al. Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances. *PLoS Comput Biol*. 2010; 6:e1000652. [PubMed: 20126523]
18. Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, et al. Creating a buzz about insect genomes. *Science*. 2011; 331:1386. [PubMed: 21415334]
19. i5K Consortium. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered*. 2013; 104:595–600. [PubMed: 23940263]

20. The 1KITE (1K Insect Transcriptome Evolution) project. <http://www.1kite.org/>
21. Baker DA, Meadows LA, Wang J, Dow JA, et al. Variable sexually dimorphic gene expression in laboratory strains of *Drosophila melanogaster*. BMC Genomics. 2007; 8:454. [PubMed: 18070343]
22. Goltsev Y, Rezende GL, Vranizan K, Lanzaro G, et al. Developmental and evolutionary basis for drought tolerance of the *Anopheles gambiae* embryo. Dev Biol. 2009; 330:462–470. [PubMed: 19298808]
23. Cameron RC, Duncan EJ, Dearden PK. Biased gene expression in early honeybee larval development. BMC Genomics. 2013; 14:903. [PubMed: 24350621]
24. Hou Y, Lin S. Distinct gene number- genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for Dinoflagellate Genomes. PLoS ONE. 2009; 4:e6978. [PubMed: 19750009]
25. Gregory, TR. Genome size evolution in animals. In: Gregory, TR., editor. The Evolution of the Genome. San Diego: Elsevier; 2005. p. 3–87.
26. Michalak P. Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. 2008; 91:243–248. [PubMed: 18082363]
27. Ueda HR, Hayashi S, Matsuyama S, Yomo T, Hashimoto S, Kay SA, et al. Universality and flexibility in gene expression from bacteria to human. Proc Natl Acad Sci U S A. 2004; 101:3765–3769. [PubMed: 14999098]
28. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 2004; 32:1372–1381. [PubMed: 14988425]
29. Yang L, Yu J. A comparative analysis of divergently-paired genes (DPGs) among *Drosophila* and vertebrate genomes. BMC Evol Biol. 2009; 9:55. [PubMed: 19284596]
30. Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. Nucleic Acids Res. 2011; 39 (Database issue):D283–8. [PubMed: 20972218]
31. Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, et al. The *Drosophila melanogaster* genetic reference panel. Nature. 2012; 482:173–178. [PubMed: 22318601]
32. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol. 2007; 24:1596–1599. [PubMed: 17488738]
33. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25:1451–1452. [PubMed: 19346325]
34. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2013. URL <http://www.R-project.org/>

Highlights

- Bidirectional promoters are widespread in insect genomes
- Motifs associated with insect bidirectional promoters are identified
- Bidirectional promoters act as transcription hotspots in insect genomes
- Bidirectionality corresponds to selective pressure on genomes
- Bidirectional promoters influences insect transcriptomes

Figure 1. Bidirectional gene pairs in different insect genomes. Top: Different bidirectional pairs between coding (C) and non-coding gene (NC) genes. Bottom: The stacked column graph shows the density (per Mb of genome) of these different gene pairs (I, II and III) for individual species. The stack columns are color coded to represent respective density values of type I, II and III pairs.

Figure 2.

Over-representation of specific transcription factors binding sites in the bidirectional promoters of insects. The graph shows number of known binding motifs (y-axis) of different transcription factors (x-axis) identified in the bidirectional promoters among the 23 insect species (shown in different colors of columns). The asterisk indicates the transcription factors whose binding motifs are enriched in the bidirectional promoters across species.

Figure 3.

Relative variation in the number of bidirectional promoters based on motif sequences closest to the transcription start sites of gene pairs. Top: The four patterns of motif composition of bidirectional promoters are shown. They differ in the closest motifs of the transcription start sites of bidirectional gene pairs. The closest motifs are either known transcription factor binding motifs (TF binding motif) or novel motifs or combinations of both. Bottom: The column graph shows the number (y-axis) of bidirectional promoters with differential motif composition (I, II, III and IV types) in different insects (x-axis).

Figure 4.

Relationship of bidirectionality with evolutionary pressure on insect species. Left panel: The bidirectionality index of 22 insect species, measured relative to *D. melanogaster* (see methods), were compared to mean ratio of non-synonymous to synonymous changes (dN/dS) as the relative measure of selective pressure on the genomes (relative to *D. melanogaster*, see text). The relative rate changes of bidirectionality and mean dN/dS are shown in y-axis for each species listed in x-axis. Right panel: Distribution curve between bidirectionality and average dN/dS among the 22 species. The polynomial trend of the distribution curve is shown along with the equation.

Figure 5.

Relationship of bidirectionality with genome size and gene density. The genome size (left panel) and gene density (right panel) of 22 insect species were compared with their corresponding values of bidirectionality. The bidirectionality as well as the fold changes in genome size and gene density of these species were measured relative to *Drosophila melanogaster* (see text). The logarithmic trend lines (with equation) are shown.

Figure 6.

Orthology of bidirectional genes among the 12 *Drosophila* species. **A)** The column graph shows, for each species, the number of gene pairs that either have orthologous copies in all the 12 *Drosophila* species (yellow column) or lack orthology in one or more other *Drosophila* species (light dark color). **B)** Line graph shows the number of gene pairs (y-axis) associated with different orthology patterns (shown in x-axis) among the 12 *Drosophila* species (represented by different line colors). The bidirectional gene pairs (I through IV) on the right show orthology patterns of the partner genes. The solid arrow represents the gene that show orthology in all the 12 species, and blank arrow represents the gene that lacks orthology in one or more other *Drosophila* species.

Figure 7.

Number of motifs of bidirectional promoters of *D. melanogaster* gene pairs that either retained the bidirectional structure or rearranged to unidirectional structure in different insects. The column graph shows the variation in the number of motifs (y-axis) among the 22 species (different colors of columns) for the two groups (retained or rearranged genes). All orthology comparisons are made relative to *D. melanogaster* bidirectional genes.

Figure 8.

Correlated expression of bidirectional gene pairs between males and females of *D. melanogaster*. The expression patterns of gene pairs are shown separately for the two partner genes: gene 1 (left) and gene 2 (right). The cumulative line graphs show different gene pairs (gene names not shown) that have correlated expression changes between males and females. The y-axis represents the fold changes of gene expressions (cumulative values) between males and females. The top panel shows positive correlation and the bottom panel shows negative correlation of expressions of bidirectional gene pairs between males and females.

Figure 9.

Similar association of bidirectionality with gene expression variation in different insects. The graphs show number of genes, either bidirectional or unidirectional, that are associated with differential (low or high) variation of gene expressions in **A**) males and females of different strains of *D. melanogaster*, **B**) different embryonic developmental stages of *A. gambiae*, and **C**) the queen- and worker- destined larval stages of *A. mellifera*. The Chi-square value (2×2 contingency tests) and the p -value of test statistics are shown in each case.

Table 1

Number of bidirectional gene pairs identified in 23 insect species.

Species	Total protein coding genes	Bidirectional gene pairs	% genes which are bidirectional
Aaeg	15998	234	2.925366
Acep	18062	347	3.842321
Adar	11430	1130	19.77253
Agam	12810	765	11.94379
Amel	10694	592	11.07163
Apis	33816	1101	6.51171
Bmor	14623	145	1.983177
Cqui	18955	649	6.847797
Dana	15069	1427	18.93954
Dere	15044	1510	20.07445
Dgri	14982	1303	17.39421
Dmel	13937	1788	25.65832
Dmoj	14594	1273	17.44553
Dper	16874	1232	14.60235
Dpse	15876	1254	15.79743
Dsec	16467	1502	18.24255
Dsim	15413	1288	16.71316
Dvir	14491	1337	18.45283
Dwil	15512	1256	16.19391
Dyak	16077	1474	18.33675
Nvit	17085	997	11.67106
Phum	10773	1194	22.16653
Tcas	16524	1384	16.75139

Note: The total number of protein coding genes and the proportion (% values) associated with bidirectional arrangements are also shown.

Table 2

Comparison of number of bidirectional genes with total number of genes present in different chromosomes of five insect species. The *p*-value represents statistical significance of overrepresentation of bidirectional genes in chromosome X.

Species	Chr 2 - BDGs	Chr 2 - total	Chr 3 - BDGs	Chr 3 - total	Chr 4 - BDGs	Chr 4 - total	Chr X - BDGs	Chr X - total	<i>p</i> -value*
Agam	490	6509	260	4741	0	0	9	1069	0.0001
Dmel	730	5455	801	6105	4	80	251	2172	0.023
Dpse	331	3474	239	2740	232	2862	424	5450	0.631
Dsim	504	5508	619	6510	0	67	126	1845	0.617
Dyak	540	5583	667	6616	2	82	212	2401	0.454

Table 3

Frequency of gene aggregations around bidirectional *versus* unidirectional gene pairs in insect genomes.

Species	n	N	k	K
Aaeg	16	234	215	990
Acep	83	347	982	2549
Adar	363	1130	2041	3268
Agam	191	765	1325	2237
Amel	118	592	955	1648
Apis	189	1101	1367	3293
Bmor	13	145	116	669
Cqui	120	649	1150	2510
Dana	461	1427	2662	4223
Dere	487	1510	2732	4466
Dgri	403	1303	2239	3546
Dmel	676	1788	2933	4802
Dmoj	397	1273	2166	3421
Dper	409	1232	2515	4423
Dpse	383	1254	2359	3933
Dsec	532	1502	2986	5062
Dsim	482	1288	2570	4462
Dvir	413	1337	2218	3507
Dwil	371	1256	2091	3309
Dyak	520	1474	2802	4663
Nvit	271	997	1823	3203
Phum	274	1194	1673	2320
Tcas	464	1384	2446	4124

n= number of aggregation around bidirectional gene pairs, N= number of total number of bidirectional gene pairs, k= number of aggregation around unidirectional gene pairs, K= total number of unidirectional gene pairs.

Table 4

Number of 1:1 orthologs of *D. melanogaster* coding genes in other insects and number of these orthologs that are bidirectional gene pairs (BDGP) between species. Bidirectionality is then expressed as the ratio of BDGP orthologs to the total the number of orthologs between the two species.

Species	1:1 orthologs	BDGP orthologs	Bidirectionality
Aaeg	5862	17	0.00290
Acep	5760	23	0.00399
Adar	6680	188	0.02814
Agam	6104	164	0.02687
Amel	6074	89	0.01465
Apis	4960	34	0.00685
Bmor	5560	11	0.00198
Cqui	5772	59	0.01022
Dana	11496	1648	0.14335
Dere	12286	2173	0.17687
Dgri	9907	1135	0.11457
Dmoj	10517	1246	0.11847
Dper	10376	1096	0.10563
Dpse	10940	1254	0.11463
Dsec	11870	1979	0.16672
Dsim	11322	1691	0.14936
Dvir	10647	1287	0.12088
Dwil	10606	1126	0.10617
Dyak	11842	2037	0.17201
Nvit	5551	115	0.02072
Phum	5986	205	0.03425
Tcas	6215	215	0.03459

Table 5Species-specific bidirectional gene pairs (BDGPs) in *Drosophila*.

Drosophila species	Species-specific BDGPs	Percentage of BDGPs in the genome
Dana	86	6.03
Dere	42	2.78
Dgri	59	4.53
Dmel	139	7.77
Dmoj	55	4.32
Dper	13	1.06
Dpse	29	2.31
Dsec	6	0.4
Dsim	8	0.62
Dvir	53	3.96
Dwil	150	11.94
Dyak	15	1.02

Table 6

Mutations in two bidirectional promoters of *D. melanogaster*.

Gene pair	S	Eta	h	pi	k
FBgn0004656_FBgn0004657	5	5	4	0.0046	0.97
FBgn0029736_FBgn0029737	1	1	2	0.00047	0.133

Note: The two promoters are associated with FBgn0004656_FBgn0004657 (whose orthologs in other *Drosophila* species are rearranged) and FBgn0029736_FBgn0029737 (whose orthologs maintain bidirectional localization in other *Drosophila* species). The polymorphisms were analyzed among 29 inbred lines from DGRP data. The heading abbreviations of the table are: S = segregating mutations, Eta = Total number of mutations, h = Number of haplotypes, pi = Nucleotide diversity, and k = Average number of nucleotide differences among the sequences.