Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene and cis-element evolution.

Stephen Richards[1*] Yue Liu[1,13], Brian R. Bettencourt[2], Pavel Hradecky[2], Stan Letovsky[2], Rasmus Nielsen[3], Kevin Thornton[3], Melissa J. Todd[3], Rui Chen[1], Richard P. Meisel[4], Olivier Couronne[5,10], Sujun Hua[6], Mark A. Smith[2], Harmen J. Bussemaker[7], Marinus F. van Batenburg[7,11], Sally L. Howells[1], Steven E. Scherer[1], Erica Sodergren[1], Beverly B. Matthews[2], Madeline A. Crosby[2], Andrew J. Schroeder[2], Daniel Ortiz-Barrientos[8], Catherine M. Rives[1], Michael L. Metzker[1], Donna M. Muzny[1], Graham Scott[1], David Steffen[1], David A. Wheeler[1], Kim C. Worley[1], Paul Havlak[1], K. James Durbin[1], Amy Egan[1], Rachel Gill[1], Jennifer Hume[1], Margaret B. Morgan[1], George Miner[1], Cerissa Hamilton[1], Yanmei Huang[2], Lenee Waldron[1], Daniel Verduzco[1], Kerstin P. Blankenburg[1], Inna Dubchak[5], Mohamed A. F. Noor[8], Wyatt Anderson[12], Kevin P. White[6], Andrew G. Clark[3], Stephen W. Schaeffer[9], William Gelbart[2], George M. Weinstock[1]and Richard A. Gibbs[1]

[1]Human Genome Sequencing Center, Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, [2]FlyBase-Harvard, Department of Molecular and Cellular Biology, Harvard University, Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 021383, [3]Departments of Biological Statistics & Computational Biology and Molecular Biology & Genetics, Cornell University, Ithaca, NY 14853, [4] Interdepartmental Graduate Program in Genetics, The Pennsylvania State University, University Park, PA 16802, [5] Lawrence Berkeley National

Laboratory, Berkeley, CA 94720, [6] *Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520,* [7] *Department of Biological Sciences and Center for Computational Biology and Bioinformatics, Columbia University, New York 10027,* [8] *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, 70803,* [9] *Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 208 Erwin W. Mueller Laboratory, University Park, PA 16802,* [10] *U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598,* [11] *Swammerdam Institute for Life Sciences, University of Amsterdam, The Netherlands,* [12] *Institute of Ecology, University of Georgia, Athens, Georgia 30602,* [13] *Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston TX, W.M. Keck Center for Computational Biology, Houston, TX.*

*** To whom correspondence should be addressed**

Abstract

The genome sequence of a second fruit fly, *D. pseudoobscura*, presents an opportunity for comparative analysis of a primary model organism *D. melanogaster*. The vast majority of Drosophila genes have remained on the same arm, but within each arm gene order has been extensively reshuffled leading to the identification of approximately 1300 syntenic blocks. A repetitive sequence is found in the *D. pseudoobscura* genome at many junctions between adjacent syntenic blocks. Analysis of this novel repetitive element family suggests that recombination between offset elements may have given rise to many paracentric inversions, thereby contributing to the shuffling of gene order in the *D. pseudoobscura* lineage. Based on sequence similarity and synteny, 10,516 putative orthologs have been identified as a core gene set conserved over 35 My since divergence. Genes expressed in the testes had higher amino acid sequence divergence than the genome wide average consistent with the rapid evolution of sex-specific proteins. Cis-regulatory sequences are more conserved than control sequences between the species – but the difference is slight, suggesting that the evolution of cis-regulatory elements is flexible. Overall, a picture of repeat mediated chromosomal rearrangement, and high co-adaptation of both male genes and cis-regulatory sequences emerges as important themes of genome divergence between these species of *Drosophila*.

Introduction

Comparative genome sequencing is an important tool in the ongoing effort to exploit conservation to annotate and analyze genes, cis-regulatory elements and architectural features of genomes. The structure of the genetic code facilitates identification of conserved protein-coding regions (*1*), while approaches such as "phylogenetic footprinting" may aid the identification of functional non-coding elements. A recent study of four Drosophila species (*2*) suggested that the sequence divergence between *D. pseudoobscura* and *D. melanogaster* is appropriate for the identification of cis-regulatory regions. Such a comparison also provides support for gene predictions, allows conserved protein-coding sequences to be identified and is a major rational for the *D. pseudoobscura* genome sequencing project.

Comparative genomic sequencing can also provide insights into the evolutionary mechanisms of genome rearrangment, which is of special interest in these species. Drosophila has been a model system for studying the evolution of chromosomes and gene order(*3*, *4*). Many chromosomal rearrangements have negative fitness consequences in many organisms because of he deleterious effects of segmental aneuploidy resulting from chromosomal segregation (reciprocal translocations and transpositions) or recombination (pericentric inversions) (*5*). In *Drosophila*, however, special features of meiosis avoid the negative fitness effect for one class of rearrangements – paracentric inversions (inversions with both breakpoints on the same chromosome arm). In male meiosis there is no crossing over and hence no recombinant aneuploid dicentric/acentric gametes. In female meiosis, where crossing over does occur, the dicentric/acentric recombinant

chromosomes are directed into polar bodies rather than the functional gamete (*6*). As a result, paracentric inversions are highly polymorphic within populations of most *Drosophila* species (*7*) and some of these inversions become fixed during speciation.

Sturtevant and Dobzhansky discovered a wealth of naturally occurring chromosomal inversion polymorphisms in *D. pseudoobscura*, predominantly on the third and X chromosomes (*8*), through an examination of salivary chromosomes (*9*). Dobzhansky first used paracentric inversion events to reconstruct relationships among *D. pseudoobscura* and *D. persimilis* third chromosomes (*10*). Genes within the *D. pseudoobscura* chromosomal inversions are likely targets of selection as the polymorphic gene arrangements form stable geographic clines (*10*), altitudinal clines in certain populations (*11*), seasonal cycling (*11*) and exhibit high levels of linkage disequilibrium (*12*). The accumulation of sequence differences among inverted chromosomes can lead to increased phenotypic variation and may contribute to the formation of new species (*13, 14*). Random breakage (*15, 16*), transposon-mediated recombination (*17-21*), and fragile breakpoints (*22, 23*) have been suggested as possible mechanisms for generating paracentric inversions in natural populations, but there is little definitive evidence. Our study provides a unique opportunity to explore the origin of these rearrangements by comparing whole descendant chromosomes.

Genome sequence of *D. pseudoobscura*

Using a whole genome shotgun method we produced 2.6 million sequence reads and assembled them into a high quality draft genome sequence. The sequence is comprised of 755 scaffolds with an N50 of 1.0 Mb, covering a total of 139 Mb (Table S1). These scaffolds can be placed into 17 ultra-scaffolds anchored onto the six chromosomal arms or Muller's elements (*24*). These ultra-scaffolds have an N50 of ~ 12 Mb, with Muller's element C and E covered by single ultra-scaffolds, Muller's B comprises 4 ultra-scaffolds and Muller's A and D have 5 and 6 ultra-scaffolds respectively. The chromosome arms of *D. pseudoobscura* are approximately 17% bigger than that of *D. melanogaster* with the exception of Muller's element C which is approximately the same size (Table S2).

Chromosomal evolution

Comparison of the *D. pseudoobscura* and *D. melanogaster* genome sequences identifies conserved linkage blocks and the associated rearrangement breakpoints in the two lineages. Despite strong conservation of sequence blocks within the 5 orthologous chromosome arms, each chromosome arm has experienced extensive internal shuffling much of which can be interpreted as a sequential series of paracentric inversions (Fig. 1). The paracentric rearrangements have produced a mean conserved linkage block size of 10.6 genes, or 83 kb in length, (see Fig S1 for size distribution). No large inter-arm translocations were observed (with one possible exception), consistent with previous small-scale analyses (*25*). *D. pseudoobscura* scaffold 7059_2327 had a mixture of best hits from genes located at the base of 2L and 2R in *D. melanogaster*. This may reflect a

class of pericentric inversions, whose breaks are so proximal on each arm that recombination does not overlap the inversion, allowing them to be tolerated without loss of fitness. Such a similar pericentric inversion has been observed within the melanogaster species subgroup (*26*), and it is possible that the *D. melanogaster* gene distribution between proximal 2L and 2R is not ancestral.

Single gene transpositions between Muller elements were observed, and in some cases a lack of introns in one ortholog indicates that these arose through retrotransposition events. Analysis of 27 well-defined retrotransposition events showed that 11 were from the *D. melanogaster* X chromosome to a *D. pseudoobscura* autosome (probability < 0.01), suggesting that gene movement away from the X chromosome is favored, consistent with observations made by Betran *et al.* (*27*). Thus far, transcripts from 7 of the 11 *D. melanogaster* derived from the X to autosome transpositions have only been found in testis-derived EST libraries. This is consistent with the hypothesis that the selective pressure for autosomal testis-specific gene expression, to retain these gene functions during male gametogenesis may be the evolutionary force underlying the biased pattern of retrotransposition (*27*)

Chromosomal rearrangements

Transposable elements are obvious candidates as causal agents for rearrangement of *Drosophila* chromosomes (*17-21, 28-35*) through recombination between offset copies of an element in reverse orientation. The junctions between adjacent syntenic blocks represent rearrangement breakpoints that have occurred in both lineages since the two species diverged. Most of the rearrangement breakpoints represent interspecific

inversions long ago fixed in one or the other lineage, but a few are intraspecific polymorphic inversions that happened to be in the sequenced strains of *D. pseudoobscura* and *D. melanogaster*.

If a bracketing pair of reverse-orientated transposable elements had caused a specific inversion, junction fragments between the relevant syntenic blocks might retain common sequences that will then degrade over time. Divergent copies of a properly located repetitive element were sought through a computational analysis of one pair of *D. pseudoobscura* polymorphic inversion breakpoints and 921 fixed breakpoints.

Identification of Intraspecific Inversion Breakpoints.

The *D. pseudoobscura* genomic sequence was determined from a strain homozygous for the Arrowhead series of four polymorphic inversion events on Muller's C (*36, 37*). Syntenic block analysis was used to identify the Arrowhead inversion breakpoints, which were then confirmed by PCR amplification across the breakpoints (Fig. 2).

The proximal Arrowhead inversion breakpoint occurred between two syntenic blocks that were not contiguous in *D. melanogaster* indicating that an additional rearrangment breakpoint has occurred at this position. The distal breakpoint, on the other hand, occurred within a 307 kb *D. melanogaster* conserved linkage block that split into 67 and 240 kb blocks. The two breakpoints define a 6.0 Mb inverted region of Muller's C in *D. pseudoobscura* that is predicted to have 775 putative orthologs, any of which

could be selective targets that alter inversion frequencies in southwestern United States populations (*38*).

The junctions between syntenic blocks for the proximal and distal Arrowhead breakpoints, as defined by their flanking syntenic blocks, are each approximately 3 kb in length. Comparison of the junction revealed shared, short conserved sequence motifs sequences (which will be referred to as *breakpoint motifs*) that varied in length from 128 to 450 bp (Figure S2). The breakpoint motifs are in reverse orientation relative to each other, consistent with them representing the vestiges of a pair of transposable elements that recombined to produce the arrowhead inversion. The breakage event between elements was staggered, at opposite ends of a 128 bp matching sequence (Fig. S2). The sequences show no significant similarity to any known *Drosophila* sequences and we have been unable to detect coding function for either a transposase or a reverse transcriptase near the conserved breakpoint motifs.

Analysis of Interspecific Breakpoints.

Junctions between syntenic blocks from the six Muller elements were extracted from *D. pseudoobscura* and *D. melanogaster* genomic sequence (Table S7 and Fig. S3). Junctions without inferred gaps had an average length of 5.6 kb, and tend to be A/T rich sequences with a mean A+T content of 60 % (Table S3).

The breakpoint motif found at the two Arrowhead breakpoints of *D. pseudoobscura* is also found at moderate frequencies at other synteny breakpoints. Over 60 % of the breakpoint sequences had sequence similarity to at least one other breakpoint

within its chromosomal arm (Table S4), this similarity is entirely due to the breakpoint motif. Each chromosomal arm had at least one breakpoint that matched over 40% of breakpoint sequences, supporting the idea that the breakpoint motif constitutes a single repetitive element family that has numerous degenerate copies in the *D. pseudoobscura* genome. Breakpoints on Muller elements C and E have higher mean inter-breakpoint match fractions than Muller elements A, B, and D with a non-parametric Kruskal-Wallis test. The distribution of match fraction for breakpoints on the five major Muller elements is shown in Fig. S3.

Fig S4 shows the degree of similarity among breakpoint sequences compared to the two arrowhead breakpoint sequences in a Percent Identity Plots (*39*) (PIPs). PIPs show that each interspecific breakpoint sequence matches similar regions in the proximal or distal Standard to Arrowhead breakpoints without respect to the chromosomal arm that the breakpoints were drawn. The length of sequence that matches among interspecific breakpoints is 25 to 601 bp for a 95% confidence interval.

Distribution of the breakpoint motif. The repeat sequence is found at other locations in the genome, but the frequencies are much reduced (Table 1). The breakpoint motif is found at the highest frequencies at junctions between syntenic blocks (33.8 - 42.6%), at moderate frequencies in noncoding sequences (10.3 - 15.3 %), and at minimal frequencies in coding regions (0.4 to 0.8%). These observed frequency differences are significantly different from each other with chi-square heterogeneity tests.

We examined the relationship of the breakpoint motif to paracentric inversions more closely on Muller element C (Fig. 3). Of 80 breakpoint motif containing junctions between syntenic blocks on Muller element C, 18 are apparently the result of simple 2 break rearrangements occurring between two ancestral syntenic blocks. Hypothesized inversion events in these nine cases, we find that the rearrangements unite adjacent conserved linkage groups in both species. Fig. 3 also shows that the orientations of repeats tend to alternate more frequently than expected at random based on a runs test (*40*) (t_s=2.20, P<0.05). The other 62 breakpoints cannot be explained as simple 2-break events, perhaps because they have been involved in multiple rearrangements.

Other transposable elements were found in the junctions between syntenic blocks, but the junctions were not enriched for these known transposable elements. For example, some *D. pseudoobscura* breakpoints had sequences similar to the *mini-me* element (*41*), which uses reverse transcriptase for retrotransposition (*42*). The *mini-me* element is found at a lower frequency at breakpoints than the breakpoint motif (3.4% vs. 38.9 %) and is not found at significantly different frequencies between breakpoints and noncoding regions with chi-square heterogeneity tests.

A phylogeny of the breakpoint motifs, which are 85% identical on average, (Fig. S5) is star-like suggesting that the breakpoint motif has rapidly radiated throughout the *D. pseudoobscura* genome. Breakpoint motifs fail to form monophyletic clusters by chromosome or region of origin, rejecting the idea that these elements are unique to a particular chromosome or have diversified based on their chromosome of origin. Also, breakpoint motifs from the same local genomic region are not more similar than

sequences separated by longer distances. In fact, the two motifs that were the most similar in this subset of sequences are from different chromosomes.

Conservation of genes between *D. pseudoobscura* and *D. melanogaster*

To gain a general picture of the level of conservation of genes between *D. pseudoobscura* and *D. melanogaster*, we examined both the nucleotide and amino acid sequences of the orthologous genes. Using the filtered global BLASTZ alignment and the *D. melanogaster* 3.1 gene model annotations, we were able to investigate the conservation of gene features between *D. pseudoobscura* and *D. melanogaster*. Fig. 4 shows the degree of sequence conservation in promoter regions, upstream regions, untranslated regions (UTR's), coding regions, introns and other parts of the gene, averaged over a large number of orthologous *D. pseudoobscura - D. melanogaster* gene pairs. The average identity of coding sequence at the nucleotide level is approximately 70% for the first and second bp of the codon, and 49% for the wobble base. Intron sequences are approximately 40% identical, UTRs 45 –50%, and protein binding sites from the literature 63%. Genome-wide aligned sequences are 46% identical. We also examined sequence conservation at the protein level. Fig. S6 depicts the percent amino acid identity of aligned orthologous protein sequences as a frequency histogram for alignments for different percentage identity. The vast majority of protein sequences show greater than 70% amino acid identity, with a peak around 85%. Proteins with ESTs derived from testis-specific libraries had a mean amino acid identity of 60%

Male-specific proteins are less conserved than others.

We searched for *D. melanogaster* genes for which no ortholog could be found in the entire *D. pseudoobscura* sequence set including unassembled sequence reads. We focused on cases where the syntenic neighbors of the *D. melanogaster* orthologs of the missing *D. pseudoobscura* gene were present. We found 75 such genes, 20 of which contained no introns, suggesting they might be the result of a retrotransposition event. It is impossible to ascertain the origins of this class of genes without additional data, but of the 20 intronless *D. melanogaster* genes not found in *D. pseudoobscura*, 11 were male specific, based on representations in testis-derived EST libraries (chi-square value = 59.7, df =1, $p < 0.00001$).

Further comparison of *D. melanogaster* ESTs from testis libraries to the *D. pseudoobscura* draft sequence found that testis-derived ESTs identify *D. melanogaster* genes that are far less likely to have putative orthologs in the *D. pseudoobscura* draft sequence than *D. melanogaster* genes with ESTs from other libraries. We could find no TBLASTN hit in the *D. pseudoobscura* sequence for 20% (1445 of 7300) of testis derived ESTs, compared to 13% (773 of 6000) of non-testis ESTs ($\chi^2 = 113$, $p < 0.00001$). Further, in 761 cases where putative orthologs genes with testis-derived ESTs could be identified, the mean identity was ~15% more divergent than for other orthologs ($p < e\text{-}75$, Fig. S6).

Evolutionary analysis of divergence of orthologous gene pairs

The ratio of divergence at amino-acid replacement (nonsynonymous) sites (d_N) to synonymous sites (d_S) can be informative of the long-term evolutionary dynamics of a gene. We calculated maximum likelihood estimates of the synonymous and nonsynonymous divergence rates between *D. pseudoobscura* and *D. melanogaster*. The median number of synonymous substitutions per synonymous site was 1.79, and the median number of nonsynonymous substitutions per nonsynonymous site was 0.14, with a skewed distribution around both values (Fig. S7). Estimates of median d_S for XL, XR and the autosomes were 1.82, 1.75, and 1.81, indicating that all have a mean with multiple hits per site. As the high level of synonymous divergence between *D. pseudoobscura* vs. *D. melanogaster* gene sequences resulted in low power and low reliability to detect positive selection ($d_N/d_S > 1$) using the d_N/d_S ratio, an alternative test was required (*43*). We fitted substitution models that split the nonsynonymous substitution rate into two bins (radical vs. conservative), each with its own rate parameter. A rate ratio of radical to conservative amino acid substitutions > 1 implies accelerated rate of radical changes. 44 genes were identified as having accelerated rates of radical substitution (Tables S5 and S6).

Conservation of known regulatory elements

To investigate conservation of cis-regulatory elements (CREs), we collected a set of experimentally characterized regulatory sites curated by the FlyBase project from published papers. We restricted our attention to sites of length less than 50 bp that seemed likely to correspond to individual CREs. Our collection comprised 142 sites

over 30 genes, characterized using a variety of experimental methods, ranging from in vitro binding assays to detection of a mutational phenotype (*44*). About 65% of these sites were upstream of their respective gene, with a modal position of 2 kb from the putative transcription start site. We compared the level of conservation of these elements to two classes of control sites: random intergenic control (RIC) sites, and nearby sites (Table 2). The comparisons reveal a small but statistically significant increase of conservation in CREs relative to both types of controls, consistent with the expectation that binding sites are under stabilizing selective pressure. This difference is most pronounced when CREs are compared to random intergenic controls, both because RICs are aligned less often than CREs or nearby sites, and because they are less conserved when aligned. We expect sites near CREs to consist of a mixture of relatively unconstrained sequence and partial overlaps of neighboring CREs, known and unknown. Therefore, we predict an intermediate level of conservation for these. The RICs may also contain CREs by chance, but presumably they are fewer in number. Fig. 5 shows the distribution of % identities above any given threshold for aligned sites. We used the Kolmogorov-Smirnov test to evaluate the significance of the maximum vertical distance between the curves. The CREs have relatively fewer instances in the 50-70% range and relatively more instances in the 80-90% range over both controls, which may correspond to regulatory elements under stabilizing selection. On a per site basis, however, the mean excess conservation of CREs relative to both control sets is very small, on the order of 5 - 6 %. As the average CRE was 17.4 bases long, this corresponds to 1 bp of additional identity in a CRE vs. a random site. Such a slight difference in conservation seems to

offer scant hope of discovering CREs of this type through their pairwise conservation in these species. One caveat is that our percent identity values are based on our reference alignment; perhaps a more sophisticated alignment strategy tuned to small patches of conservation would yield different results. Also, alignments of more than two species may reveal conserved structures that a pairwise alignment cannot, as in Kellis *et al.* (*45*). Finally, the *D. melanogaster* - *D. pseudoobscura* evolutionary distance may not be optimal for revealing CRE conservation.

Discussion

Evolutionary Model of Genomic Rearrangement.

One striking feature of conservation between *D. melanogaster* and *D. pseudoobscura* is the overwhelming conservation of gene location on chromosome arm. This contrasts with the *Anopheles gambiae* - *D. melanogaster* comparison, where there is a tendency for arm conservation, but with a considerable frequency of violations. Thus, although the basic mechanism favoring paracentric rearrangements appears to be a dipteran-wide phenomenon, over longer evolutionary time (250-300 Mya since the divergence of *Anopheles* and *Drosophila*, compared with 35 Mya since the divergence of *D. melanogaster* and *D. pseudoobscura*) there is clearly a breakdown of arm by arm integrity (*46*). Perhaps scaffold 7059_2327 with its mixture of proximally-located genes from *D. melanogaster* 2L and 2R arms is a hint at one mechanism that can, over long evolutionary time, lead to extensive reshuffling of genes between arms.

Several pieces of evidence are consistent with the breakpoint motif being causal in the generation of chromosomal rearrangements in the *D. pseudoobscura* lineage. The breakpoint motifs in the Arrowhead inversion are in reverse orientation, consistent with a mechanism where ectopic exchange generates an inversion event (Fig. 6). The conserved sequence motif is virtually absent from intron and coding sequences. This suggests that strong purifying selection has acted to prevent the accumulation of this sequence within introns. If the conserved motif serves as the target for rearrangements, then inversions that use elements within a gene would cause loss-of-function mutations that would be quickly removed from populations (*47*). Repeated sequences have also been detected at conserved linkage breakpoints among trypanosome species (*48*).

One problem with the high frequency of the breakpoint motif is that ectopic exchange between elements in the same orientation would lead to deletion mutations. Two factors minimize the impact of the repeat element as a cause of deleterious mutations. First, the repeats tend to accumulate nucleotide substitutions rapidly, reducing the risk of pairing. Second, the breakpoint motifs tend to degrade in length fairly rapidly. The size of the repeat motif varies significantly, suggesting that these elements tend to accumulate deletions that decrease the average size of the elements. These data are consistent with the "dead-on-arrival" elements of *D. virilis* that preferentially delete sequence (*49, 50*). As a consequence, few intact elements are capable of ectopic exchange. Molecular evolutionary studies of homologous breakpoint motifs will be necessary to test the element degradation hypothesis. The conclusion that the conserved sequence element

causes paracentric inversions should be tempered as other possible explanations for the coincidence of the breakpoint repeat and inversion breakpoint may exist.

One can speculate about why breakpoint repeat elements are found only in the *D. pseudoobscura* lineage. Perhaps a new repetitive DNA element has been introduced in the obscura group lineage. *D. subobscura* is a close relative of *D. pseudoobscura* and five of the six chromosomal elements are segregating for paracentric inversions in European populations (*17*). It will be interesting to know if the repeat motif is present at the breakpoints of *D. subobscura* rearrangements.

Fixed inversion differences between the species may play a significant role in the formation of new species because inversions prevent the spread of incompatibility genes between different chromosomal backgrounds (*13, 14*). By reducing rates of crossover, chromosomal inversions act as a barrier to gene flow, allowing Dobzhansky-Muller incompatibility genes to be fixed in different gene arrangement backgrounds greatly enhancing the possibility of speciation (*13, 14*). In *Drosophila* hybrid male sterility genes appear to be involved in the process of speciation. In fact, we find that *D. pseudoobscura* genes with testis expression show a significant decrease in identity with their *D. melanogaster* orthologs. Genes supported by expression sequences from testes cDNA libraries accumulate sequence variation faster than the average. It will be interesting to determine if genes within inverted regions, and particularly those with male

specific expression are associate with the sterility of male hybrids of *D. pseudoobscura* and *D. persimilis*.

Conservation of Known Cis Regulatory regions

D. pseudoobscura was chosen as the second fly species to be sequenced in part because it appeared to have the right sequence divergence from *D. melanogaster* to locate cis-regulatory sequences (*2*). We were somewhat surprised at the lower level of conservation of known cis-regulatory regions. An examination of many *Drosophila* genes using the VISTA Genome Browser (http://pipeline.lbl.gov/pseudo) will identify several conserved non-coding regions. Bergman *et al*. (*2*) used clusters of these conserved non-coding sequences to identify enhancer sequences in the *apterous* gene. However when known regulatory regions are examined, the conservation signal is not striking. Others have come to a similar conclusion using different alignment methods (*51*). Alignment of *C. elegans* and *C. briggsae* has also suggested that many conserved non-coding regions will not be due to cis-regulatory sequences, increasing the noise in the conservation signal of these elements (*52*). Alignments of additional species of intermediate divergence may improve the detection of known regulatory elements as in Kellis *et al*. (*45*), assuming the elements are conserved.

The lack of a clear conservation of cis-regulatory sequences suggests that neutral models of sequence divergence in regulatory regions may be naïve. Ludwig et al. (*53*) observed the *D. pseudoobscura eve* stripe 2 enhancer was functional in *D. melanogaster* despite

significant differences between the regulatory protein binding sites. In contrast, chimeric *eve* stripe 2 promoters had improper expression patterns suggesting that stabilizing selection acting on the enhancer (*54*) where "…selection can maintain functional conservation of gene expression for long periods of evolutionary time despite binding site turnover." The *D. pseudoobscura* transcription factor proteins are 17% diverged from their *D. melanogaster* orthologs (Fig S6.), different enough to allow variation of binding specificity. Evidence of cis-regulatory binding site conservation is encouraging, however it is clear the *D. pseudoobscura – D. melanogaster* sequence comparisons will not identify binding sites alone. Instead the approach of phylogenetic shadowing (*55*) using a multiple alignment with species of intermediate divergence shows more promise, due to the reduced chance of binding site turnover between more recently diverged species.

References

1. A. Nekrutenko, K. D. Makova, W. H. Li, *Genome Res* **12**, 198 (Jan, 2002).
2. C. M. Bergman *et al.*, *Genome Biology* **3** (30th December 2002, 2002).
3. A. H. Sturtevant, C. C. Tan, *Journal of Genetics* **34**, 415 (1937).
4. A. H. Sturtevant, E. Novitski, *Genetics* **26**, 517 (1941).
5. C. P. Swanson, T. Merz, W. J. Young, *Cytogenetics: The chromosome in division, inheritance and evolution* (Prentice-Hall, Inc., Englewood, NJ, 1981), pp.
6. A. H. Sturtevant, G. W. Beadle, *Genetics* **21**, 544 (1936).
7. D. Sperlich, P. Pfriem, in *The Genetics and Biology of Drosophila* M. Ashburner, H. L. Carson, J. N. Thomson, Eds. (Academic Press, New York, NY, 1986), vol. 3e, pp. 257 - 309.
8. A. H. Sturtevant, T. Dobzhansky, *PNAS* **22**, 448 (1936).
9. T. S. Painter, *Genetics* **19**, 175 (1934).
10. T. Dobzhansky, C. Epling, in *Carnegie Institution of Washington Publication 554*. (Washington, DC, 1944) pp. 47 - 144.
11. T. Dobzhansky, *Genetics* **33**, 158 (1948).
12. S. W. Schaeffer *et al.*, *Proc Natl Acad Sci U S A* **100**, 8319 (Jul 8, 2003).
13. A. Navarro, N. H. Barton, *Evolution Int J Org Evolution* **57**, 447 (Mar, 2003).

14. M. A. Noor, K. L. Grams, L. A. Bertucci, J. Reiland, *Proc Natl Acad Sci U S A* **98**, 12084 (Oct 9, 2001).
15. S. Ohno, *Nature* **244**, 259 (1973).
16. J. H. Nadeau, B. A. Taylor, *Proc Natl Acad Sci U S A* **81**, 814 (Feb, 1984).
17. C. B. Krimbas, in *Drosophila Inversion Polymorphism* C. B. Krimbas, J. R. Powell, Eds. (CRC Press, Boca Raton, 1992) pp. 127-220.
18. K. D. Mathiopoulos *et al.*, *Parassitologia* **41**, 119 (Sep, 1999).
19. M. Caceres, J. M. Ranz, A. Barbadilla, M. Long, A. Ruiz, *Science* **285**, 415 (Jul 16, 1999).
20. F. Casals, M. Caceres, A. Ruiz, *Mol Biol Evol* **20**, 674 (May, 2003).
21. M. B. Evgen'ev *et al.*, *Proc Natl Acad Sci U S A* **97**, 11337 (Oct 10, 2000).
22. E. Novitski, *Genetics* **31**, 508 (1946).
23. P. Pevzner, G. Tesler, *Proceedings of the National Academy of Sciences USA* **100**, 7672 (Jun 24, 2003).
24. H. J. Muller, in *The New Systematics* J. Huxley, Ed. (Clarendon Press, Oxford, 1940) pp. 185 - 268.
25. J. M. Ranz, F. Casals, A. Ruiz, *Genome Res* **11**, 230 (2001).
26. F. Lemeunier, M. A. Ashburner, *Proc R Soc Lond B Biol Sci* **193**, 275 (May 18, 1976).
27. E. Betran, K. Thornton, M. Long, *Genome Res* **12**, 1854 (Dec, 2002).
28. F. Sheen, J. K. Lim, M. J. Simmons, *Genetics* **133**, 315 (Feb, 1993).
29. M. Collins, G. M. Rubin, *Nature* **308**, 323 (Mar 22-28, 1984).
30. T. W. Lyttle, D. S. Haymer, *Genetica* **86**, 113 (1992).
31. S. S. Potter, *Mol Gen Genet* **188**, 107 (1982).
32. J. K. Lim, *Proc Natl Acad Sci U S A* **85**, 9153 (Dec, 1988).
33. V. Ladeveze, S. Aulard, N. Chaminade, G. Periquet, F. Lemeunier, *Proc R Soc Lond B Biol Sci* **265**, 1157 (Jul 7, 1998).
34. R. K. Blackman, R. Grimaila, M. M. Koehler, W. M. Gelbart, *Cell* **49**, 497 (May 22, 1987).
35. W. R. Engels, C. R. Preston, *Genetics* **107**, 657 (Aug, 1984).
36. J. R. Powell, in *Drosophila Inversion Polymorphism* C. B. Krimbas, J. R. Powell, Eds. (CRC Press, Ann Arbor, MI, 1992) pp. 73-126.
37. A. Popadic, W. W. Anderson, *Proceedings of the National Academy of Sciences USA* **91**, 6819 (1994).
38. S. W. Schaeffer *et al.*, *Proceedings of the National Academy of Sciences USA* **100**, 8319 (2003).
39. S. Schwartz *et al.*, *Genome Res* **10**, 577 (Apr, 2000).
40. R. R. Sokal, F. J. Rohlf, *Biometry* (W. H. Freeman and Co., New York, ed. 2, 1981), pp. 859.
41. J. Wilder, H. Hollocher, *Molecular Biology and Evolution* **18**, 384 (2001).
42. Stephen W. Schaeffer, unpublished data
43. A. C. Methods, *Methods AC* **1**, 1 (2003).
44. Stan, *Blank* **1**, 1 (2003).

45. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, *Nature* **423**, 241 (May 15, 2003).
46. E. M. Zdobnov *et al.*, *Science* **298**, 149 (Oct 4, 2002).
47. B. Charlesworth, A. Lapid, D. Canada, *Genet Res* **60**, 115 (Oct, 1992).
48. E. Ghedin *et al.*, *Mol Biochem Parasitol* **134**, 183 (Apr, 2004).
49. D. A. Petrov, E. R. Lozovskaya, D. L. Hartl, *Nature* **384**, 346 (Nov 28, 1996).
50. D. A. Petrov, D. L. Hartl, *Mol Biol Evol* **15**, 293 (Mar, 1998).
51. E. G. Emberly, N. Rajewsky, E. D. Siggia, *BMC Bioinformatics* **4**, 57 (Nov 20, 2003).
52. L. D. Stein *et al.*, *PLoS Biol* **1**, E45 (Nov, 2003).
53. M. Z. Ludwig, N. H. Patel, M. Kreitman, *Development* **125**, 949 (1998).
54. M. Z. Ludwig, C. Bergman, N. H. Patel, M. Kreitman, *Nature* **403**, 564 (Feb 3, 2000).
55. D. Boffelli *et al.*, *Science* **299**, 1391 (Feb 28, 2003).
56. W. R. Rice, *Evolution* **43**, 223 (1989).

Table 1. Breakpoint sequence motif frequencies in three classes of sequence in six Muller's elements in *D. pseudoobscura*.

	Breakpoints		Noncoding			Coding		
Element	n^*	(% ± SD)†	*n*	(% ± SD)	χ^2	*n*	(% ± SD)	χ^2
A	210	33.8 ± 3.3	1698	15.3 ± 0.9	45.5‡	1851	0.8 ± 0.2	513.6‡
B	135	43.0 ± 4.3	2031	12.9 ± 0.7	90.3‡	2124	0.8 ± 0.2	703.0‡
C	205	39.0 ± 3.4	2082	11.4 ± 0.7	119.4‡	2276	0.7 ± 0.2	733.4‡
D	141	42.6 ± 4.2	2068	14.3 ± 0.8	78.3‡	2159	0.6 ± 0.2	758.0‡
E	223	38.1 ± 3.3	2636	10.3 ± 0.6	146.1‡	2923	0.4 ± 0.1	985.8‡
F	7	57.1 ± 18.7	76	44.7 ± 5.7	0.4	63	17.5 ± 4.8	5.9‡

- The total number of sequences within each category. † The percentage of sequences within each category that matched the conserved sequence motif ± standard deviation. The three categories are: Breakpoints, sequences at the boundary of two conserved linkage groups; Noncoding, sequences that are not breakpoints or coding; and Coding, sequences of protein coding genes including introns. ‡, Probability of the χ^2 value for the heterogeneity test with one degree of freedom is ≤ 0.05 after apply a Bonferroni correction for multiple comparisons (*56*). A χ^2 heterogeneity test is used to determine if the frequency of the breakpoint motif is significantly different between either the noncoding or coding regions.

Table 2 Comparison of conservation of cis-regulatory elements (CREs) to two types of control sites. Twenty RICs were generated for each CRE by randomly choosing sites of the same length as the CRE, on the same chromosome and strand, and rejecting any that overlapped a known gene. Ten nearby control sites were generated for each CRE by adding positive and negative (i.e., 3' and 5') offsets of 50, 100, 150, 200, and 250 bp to the coordinates of each true CRE. Percentage identities for all CRE and control sites were computed relative to reference alignment, on both a per site and per base basis. Unaligned bases, mismatches, and D. melanogaster insertions contributed zeros to % identity results; D. pseudoobscura insertions were ignored. The distributions of % identity values were clearly not normal, so we avoided using tests such as the t-test which assumes normality. We compared the per site and per base mean % identities of each group using a re-sampling test, where the p-value of the observed difference was estimated as the frequency (over a million trials) in which a value as large or larger than the observed CRE mean was observed in an equal-sized sample of control sites. Similarly, the p-value of the difference between the two control sets was estimated using a randomisation test (over a million trials) in which the sets mixed and then repartitioned into corresponding mock control sets. We compared the distributions using the Kolmogorov-Smirnov test, which measures the likelihood that samples came from the same continuous distribution.

Table 2

	Group 1 vs. Group 2	CRE vs. Nearby	CRE vs. Random Intergenic	Nearby vs. Random Intergenic
PER-SITE ANALYSIS	Group 1 mean per-site % identity	51.3%	51.3%	47.8%
	Group 2 mean per-site % identity	47.8%	42.9%	42.9%
	Difference of means (group 1 - group 2)	3.6%	8.4%	4.9%
	Difference of means resampling p-value	0.05	0.003	1E-5
	Distribution comparison KS p-value	0.026	0.0016	2E-6
PER-BASE ANALYSIS	Group 1 mean per-base % identity	47.8%	47.8%	46.3%
	Group 2 mean per-base % identity	46.3%	42.4%	42.4%
	Difference of means (group 1 - group 2)	1.5%	5.4%	3.9%
	Difference of means resampling p-value	0.024	0.05	5.8E-4

Fig. 1. The syntenic relationship between *D. pseudoobscura* and *D. melanogaster*. Synteny dot-plots showing the shuffled syntenic relationships between *D. pseudoobscura* and *D. melanogaster* for the five chromosome arms. In each case the *D. melanogaster* chromosome is shown on the X axis and *D. pseudoobscura* chromosome on the Y axis. Note that lines within the graph are all of the same thickness, but are of varying length. Chromosomes have been color coded to allow identification of inter-chromosomal synteny blocks. Muller element F is not shown due to the lack of sequence anchoring data on this chromosome.

Fig. 2. Mapping intraspecific inversion breakpoints. The Arrowhead inversion converted the Standard gene arrangement into the Arrowhead gene arrangement. The distal breakpoint of the Standard to Arrowhead arrangement was shown to map near the *vestigial* locus based on *in situ* localization to salivary chromosomes (*38*). A conserved linkage breakpoint was detected 25 kb distal to the *vestigial* gene (Fig. 2A). PCR confirmed that the conserved linkage break distal to the *vestigial* locus is the distal ST to AR breakpoint because the derived primer pair (dAR, b and d in Fig. 2B) amplifies Arrowhead genomic DNA, but fails to amplify Standard genomic DNA (Fig. 2C). The proximal ST to AR breakpoint was identified as the conserved linkage break that reunites the 58D-E region of *D. melanogaster* (Fig. 2A). PCR confirmed that the conserved linkage break proximal to the *D. pseudoobscura* homologue of the *D. melanogaster* 58E region is the proximal ST to AR breakpoint because the derived primer pair (dAR, a and c in Fig. 2B) amplifies Arrowhead DNA and not Standard DNA (Fig.2C). Further support was obtained by doing the reciprocal PCR experiments where the ancestral primer

pairs (pST a and b; dST c and d in Fig. 2B) only amplified DNA from Standard and not Arrowhead strains. Sequence analysis of the PCR products from the Standard and Arrowhead backgrounds verifies that PCR amplified the appropriate sequences.

Fig. 3 Rearrangement of the *D. pseudoobscura* genome at interspecific breakpoints that have repeat motifs. The thick horizontal lines represent the chromosomal maps of *D. melanogaster* and *D. pseudoobscura* Muller's element C. Vertical lines drawn either down (*D. melanogaster*) or up (*D. pseudoobscura*) indicate conserved linkage groups that contain one or more genes. Diagonal lines connect homologous linkage groups in the two species. The locations and orientations of 80 conserved linkage breakpoints that contain the repeated motif are indicated with the open and filled triangles. The plot shows that 18 of the 80 repeat elements may have served as templates for ectopic exchange in nine single inversion events (gray lines) , e.g., inversion 1 (black dashed lines). In these cases, the adjacent conserved linkage groups that flank both breakpoints are symmetrically exchanged. The plot also shows the likely result when a pair of adjacent conserved linkage groups were involved in multiple inversion events as is the case for the Standard to Arrowhead inversion (solid black lines).

Fig. 4. **A:** Averaged conservation of different segments of a "prototypical gene". Conservation statistics were computed over thousands of aligned pairs of regions of various types, aligned at different reference points. At each position we compute the fraction of aligned pairs which have identical bases at that position (green+purple tiers), have mismatched bases (red), melanogaster bases aligned to

deleted bases in pseudoobscura (yellow), or are unaligned in our synteny-filtered BLASTZ alignment (blue). The purple tier shows the fraction of bases that would be expected to match by chance given the base composition at that position in both species. The vertical panels correspond to different segments of a prototypical gene, indicated by the cartoon below. The segments are, (A) 140 Protein binding sites of 50 bp or less from literature, (B) compressed sampling of 5' proximal region every 50 bp from 50 to 500, (C) 50 bp proximal to transcription start site (TS), aligned at TS, (D) genomic span of 5' UTR, aligned at TS, (E) 5' UTR span aligned at protein start site (PS), (F) 5' end of protein coding region aligned at PS, (G) 3' end of coding exons aligned at donor site, (H) intron aligned at donor site, (I) introns aligned at acceptor, (J) 5' end of internal coding exons aligned at acceptor site, (K) 3' end of protein coding region aligned at protein end site (PE), (L) 3' UTR span aligned at PE, (M) 3' UTR span aligned at transcript end, (N) 50 bp of 3' proximal region aligned at transcript end, (O) compressed sampling of 3' proximal region every 50 bp from 50 to 500, (P) genome wide average. **B**: Distirbution of d_N/d_S for the *melanogaster-pseudoobscura* comparison of 9,184 inferred orthologous protein-coding genes. **C**: Distributions of α, the ratio of rates of substitution that are radical to those that are conservative, based on 9184 alignments of orthologous protein-coding genes in *D. pseudoobscura and D. melanogaster*. Radical changes influence charge, polarity, or polarity & volume to a greater degree than do conservative changes. A substitution model was fitted by maximum likelihood to estimate these rate parameters.

Fig. 5. Percent of sites greater than a given percent identity threshold, excluding sites that could not be aligned. The Kolmogorov-Smirnov test assesses the likelihood of the maximum difference between two of these curves under the assumption that both samples came from the same distribution. In particular, the bump in the CRE curve in the 60–90% range reflects a shift of the CRE's towards higher percent identity relative to both controls.

Fig. 6. Mechanism for chromosomal inversion with a repeated sequence motif. (**A**) Transposable element-mediated rearrangements. A hypothetical chromosome is shown with genes A through N and two repeated sequence motifs (open and black arrows) in a reverse orientation (top). Repeated motifs are shown pairing during meiosis with a recombination event occurring in the middle of the paired motifs (middle). Resolution of the recombination event between the repeated sequence motifs leading to the inversion of the central gene region (bottom). (**B**) Alternative mechanism where rearrangements occur in sequences that are hotspots for double strand DNA breaks. Repetitive elements insert after the chromosomal rearrangements occur.

Figure 1. Syntenic relationships between *Drosophila pseudoobscura* and *melanogaster*

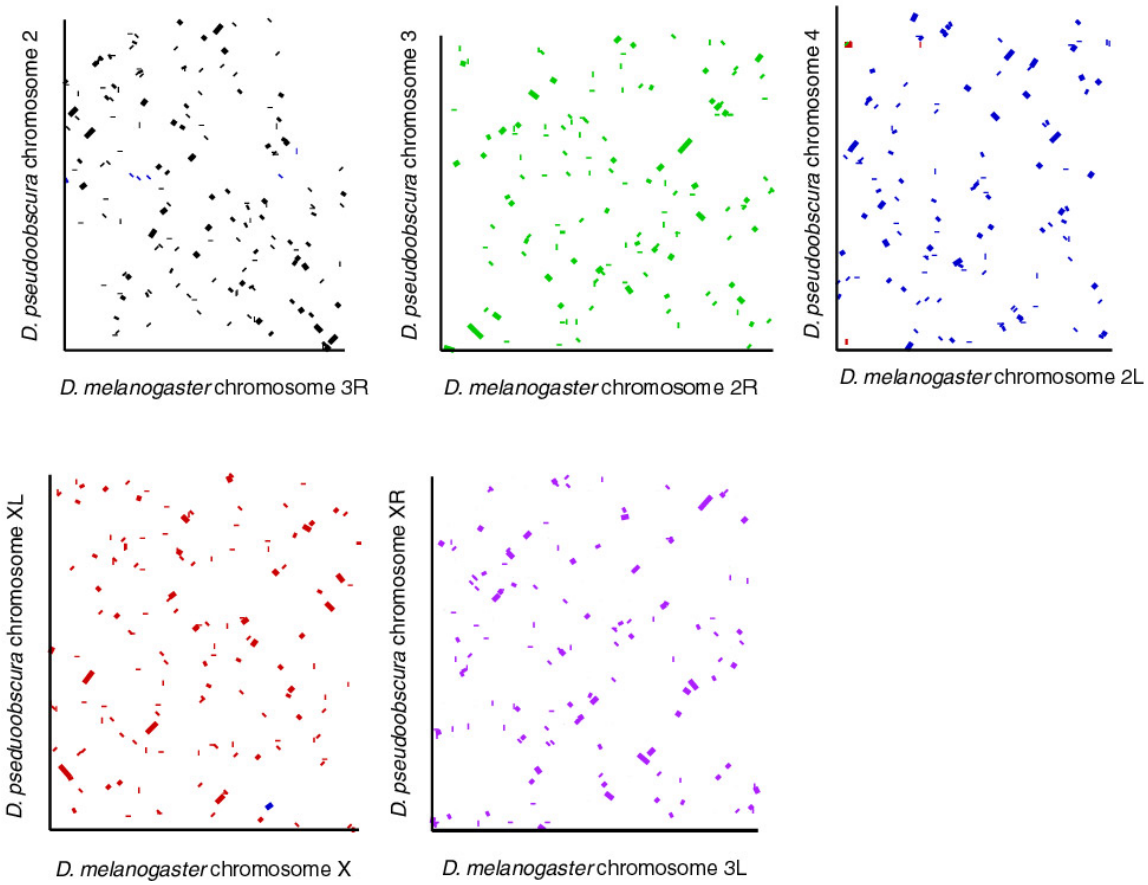


Figure 2

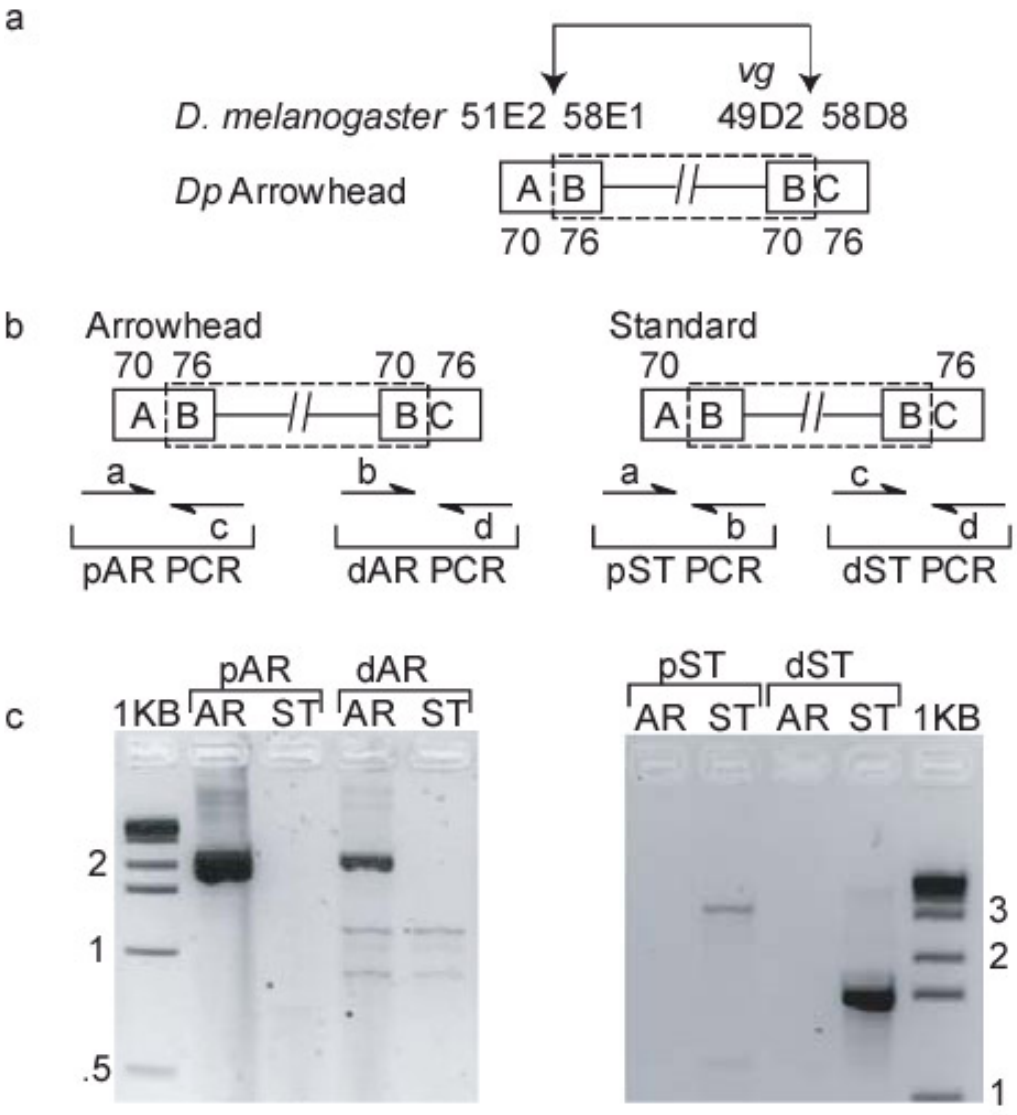

a
vg
D. melanogaster 51E2 58E1 49D2 58D8
Dp Arrowhead
A B B C
70 76 70 76
b
Arrowhead
70 76 70 76
A B B C
a c b d
pAR PCR
dAR PCR
Standard
70 76
A B B C
a b c d
pST PCR
dST PCR
c
pAR dAR
1KB AR ST AR ST
2
1
.5
pST dST
AR ST AR ST 1KB
3
2
1

Figure 3

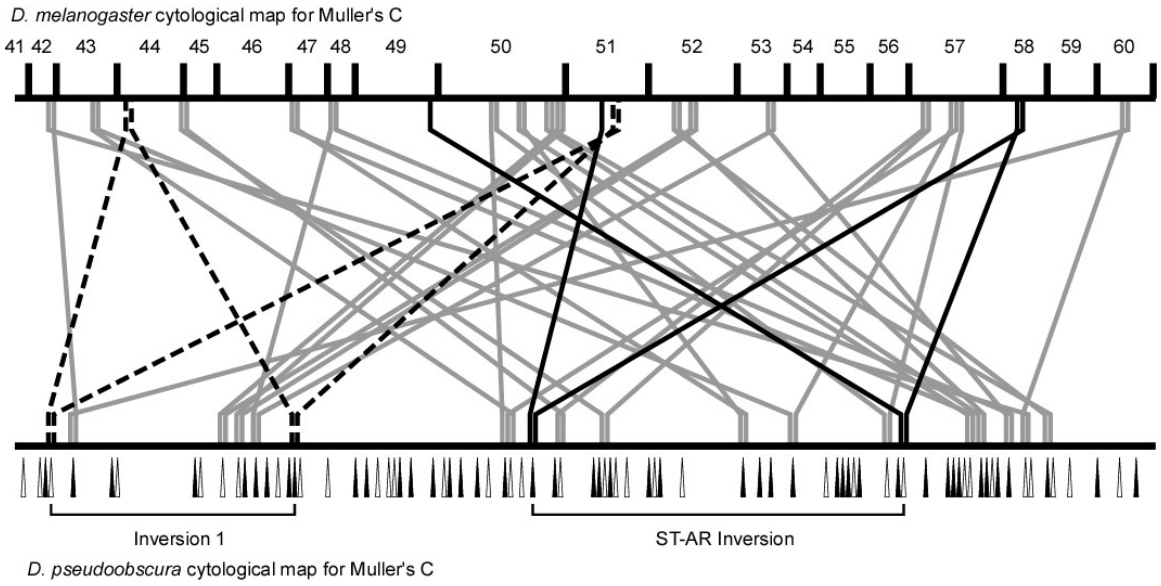

D. melanogaster cytological map for Muller's C
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
Inversion 1
ST-AR Inversion
D. pseudoobscura cytological map for Muller's C

Figure 4

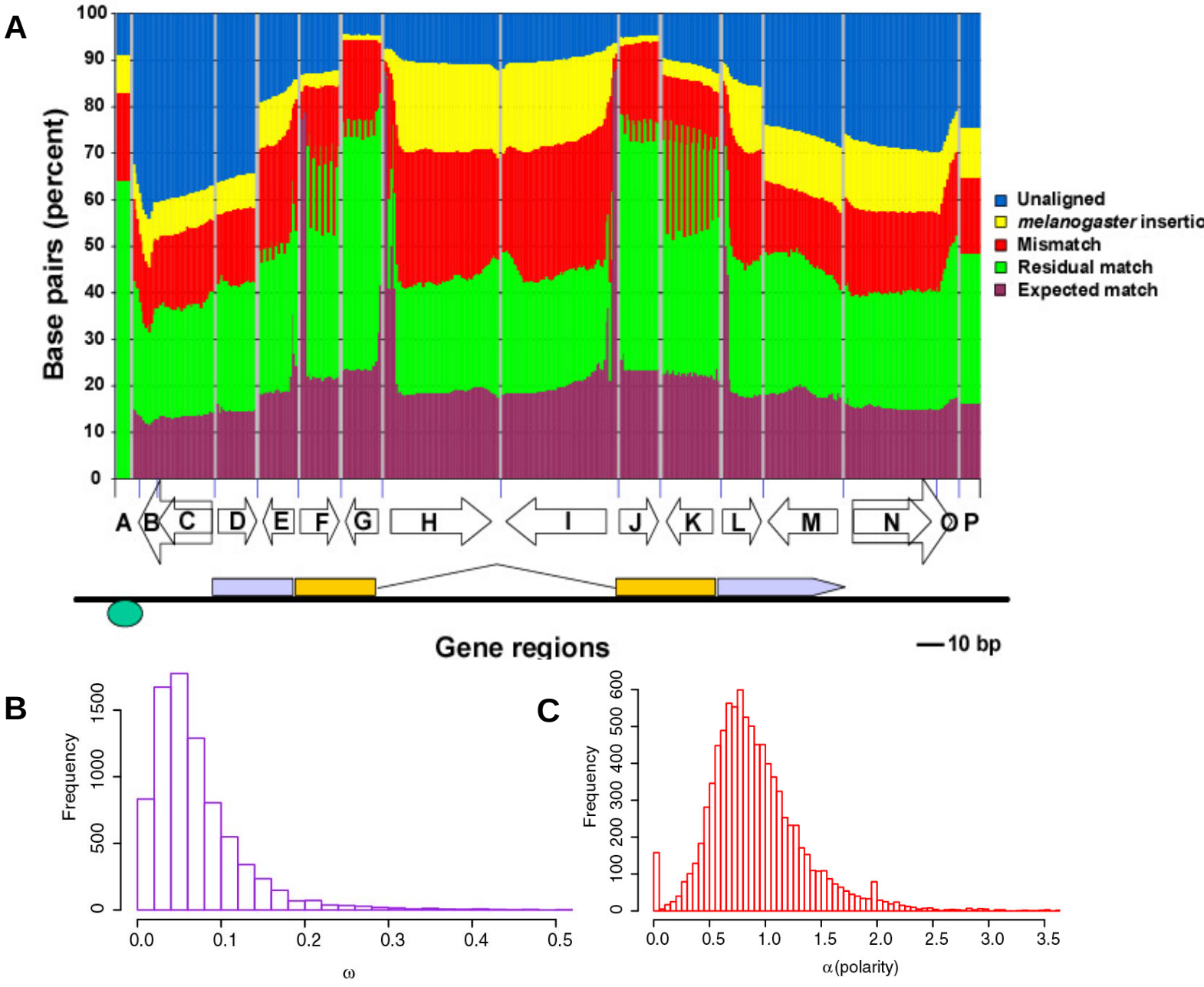

A
Base pairs (percent)
100
90
80
70
60
50
40
30
20
10
0
Unaligned
melanogaster insertio
Mismatch
Residual match
Expected match
A B C D E F G H I J K L M N O P
Gene regions
10 bp
B
Frequency
1500
1000
500
0
0.0
0.1
0.2
0.3
0.4
0.5
ω
C
Frequency
600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
α (polarity)

Figure 5

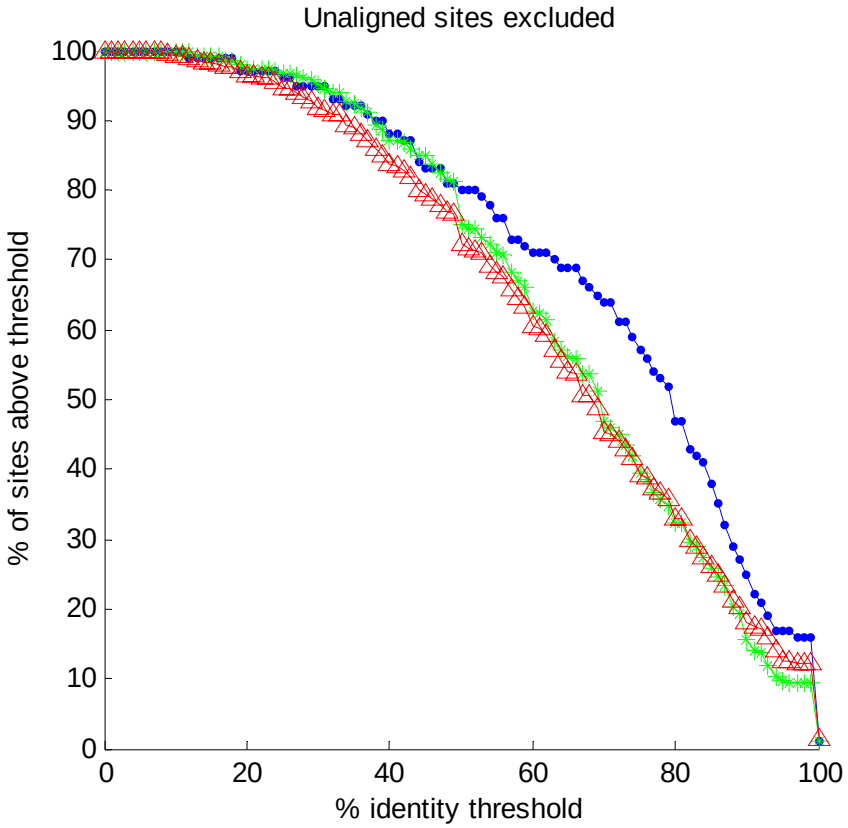
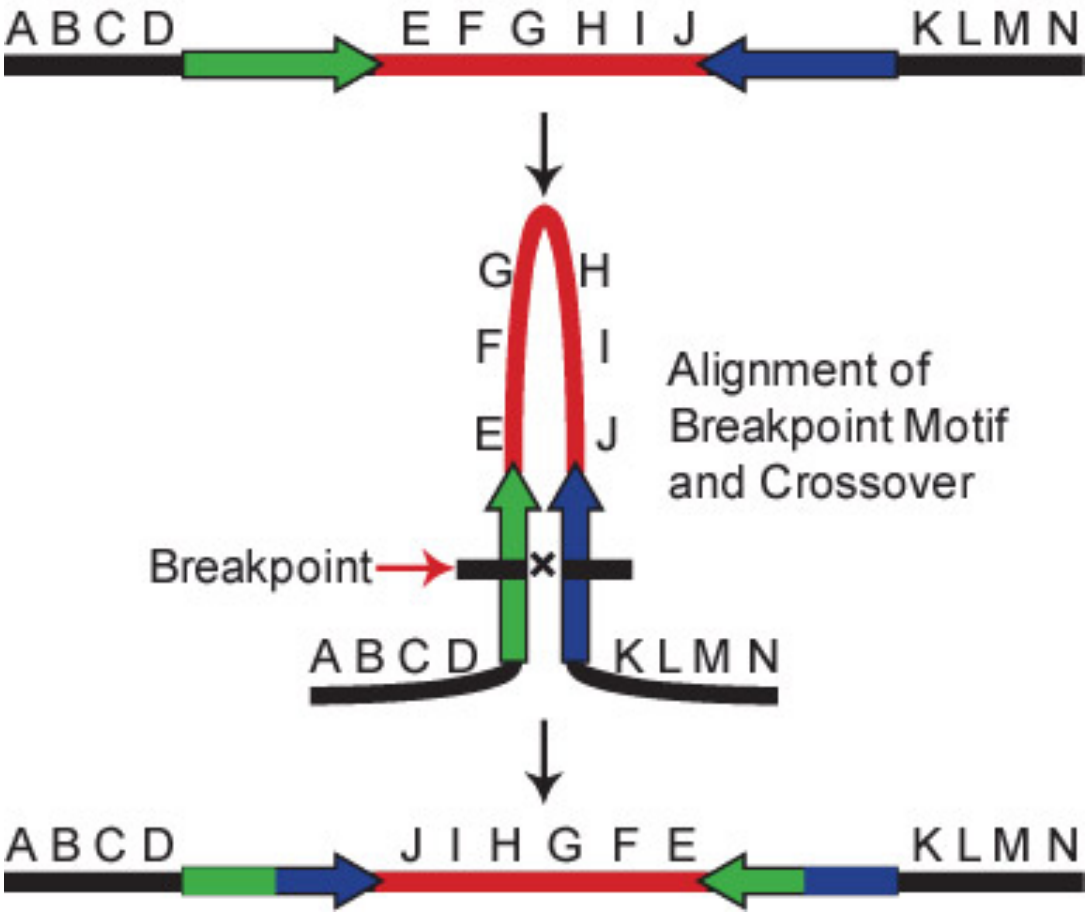

Unaligned sites excluded
% of sites above threshold
% identity threshold
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100

Figure 6


A B C D
E F G H I J
K L M N
G H
F I
E J
Alignment of
Breakpoint Motif
and Crossover
Breakpoint
A B C D
K L M N
A B C D
J I H G F E
K L M N