Research Notes

Characterization of novel repetitive element *Leviathan* **in** *Drosophila pseudoobscura***.**

Garfield, David A., and Mohamed A.F. Noor.   Biology Department, Duke University, Durham, NC, USA; noor@duke.edu.

Introduction

Repetitive sequences, such as transposable elements, may often contribute to the formation of chromosomal inversions or translocations via ectopic exchange (Charlesworth *et al.*, 1994; Coghlan *et al.*, 2005; Noor and Chang, 2006). The pattern of a strong association between rearrangement breakpoints and repeat sequences has been detected in plants, fungi, and various animals (Coghlan *et al.*, 2005). This association has been demonstrated in many *Drosophila* species (Lim, 1988; Caceres *et al.*, 1999; Evgen'ev *et al.*, 2000) and was particularly striking in the comparison of the published *D. melanogaster* and *D. pseudoobscura* genome sequences, in which at least 60% of all break-point junctions in *D. pseudoobscura* showed significant similarity (by BLAST) to at least one other breakpoint within the genome (Richards *et al.*, 2005). However, this pattern was not apparent at the breakpoints distinguishing *D. melanogaster* from its closer relative *D. yakuba* (Ranz *et al.*, 2007), suggesting a possible burst in transpositional activity in the lineage leading to *D. pseudoobscura*.

Previous analysis of the *D. pseudoobscura* genome sequence assembly identified two particular "breakpoint motifs" that together were associated with nearly half of the almost 1000 inversion breakpoints separating it from *D. melanogaster* as well as both breakpoints of an intraspecific polymorphic inversion (Richards *et al.*, 2005). This strong association is particularly impressive given that these species diverged close to 55 million years ago (Tamura *et al.*, 2004) yet this sequence has no significant similarity to any known *Drosophila* transposable elements (Richards *et al.*, 2005). As such, it seems to have arisen within the *D. pseudoobscura* lineage yet still played a major role in genome rearrangement.

Recently, we localized the inversion breakpoints between *D. pseudoobscura* and its sibling species *D. persimilis* (Ortiz-Barrientos *et al.*, 2006; Machado *et al.*, 2007; Noor *et al.*, 2007) which diverged only within the last 500,000-1,000,000 years (Aquadro *et al.*, 1991; Hey and Nielsen, 2004; Leman *et al.*, 2005). At the breakpoint regions of the XR-chromosome arm inversion separating these species, we identified a tandemly repeated sequence (Machado *et al.*, 2007), and like the previously described "breakpoint motif," we find that this 319-bp motif also appears to be both abundant across the *D. pseudoobscura* genome and absent from genomes of other distantly related *Drosophila* species. As such, it, too, may have been an architect for recent genome structural changes in *D. pseudoobscura* and perhaps led to an increase in the divergence between the closely related *D. persimilis* and *D. pseudoobscura* (*e.g.*, Noor *et al.*, 2007).

In this study, we took three computational steps to characterize this motif, named "Leviathan," in addition to experimentally looking for evidence of transcriptional activity. First, because different molecular mechanisms of replication leave different signatures in the genome (Burt and Trivers, 2006), we examined the abundance and chromosomal distribution of Leviathan across the *D. pseudoobscura* genome, and examined the history of its spread through the genome using sequence divergence between repeat elements. We also tested the "master gene" model (Deininger *et*

al., 1992), wherein one or a few elements within a genome act as "templates" for other copies in the genome, which cannot themselves be used to produce further copies, regarding the spread of this element. The number of templates for a repetitive element provides not only a clue for the mechanisms by which the element replicates (Burt and Trivers, 2006) but also an estimate of the selective costs of replication (Charlesworth and Charlesworth, 1983) and allows for a more thorough characterization of a repetitive element. Finally, we compared the distribution of Leviathan elements between *D. pseudoobscura* and *D. persimilis*, looking for differences in divergence across different parts of the genome as well as for differences in the evolutionary dynamics of Leviathan in one lineage as compared to the other.

Methods

Initial characterization of Leviathan:

We performed initial searches for sequences homologous to those flanking the fixed inversions between the two species using BLAST (Altschul *et al.*, 1990) to search the genome of each species (Gilbert, 2007) as well as the NCBI databases. Full genome sequences for *Drosophila pseudoobscura* (comparative analysis freeze 1, genome sequence version 2.0.1) and *D. persimilis* (also CAF1) were downloaded from Baylor College of Medicine (BCM-HGSC, 2006) and FlyBase (Grumbling and Strelets, 2006), respectively. We searched for the presence of Leviathan within annotated CDS and introns using annotations downloaded from FlyBase as well as our own annotation of the *D. persimilis* genome (Noor *et al.*, 2007). We processed several contigs containing multiple BLAST hits using the program Dotter (Sonnhammer and Durbin, 1995) with default parameters.

Tandem Repeat Finder:

To delineate arrays of repetitive sequence, and to assess the distribution of the repetitive sequence across the genome, we ran the Tandem Repeat Finder Program (aka TRF: Benson, 1999) on both the *D. pseudoobscura* and *D. persimilis* sequences using the suggested default settings: matching weight = 2, mismatching penalty = 7, indel penalty = 7, match probability = 80, indel probability = 10, minimum alignment score = 50, maximum period size = 500. We assembled the results into a single, searchable database using custom scripts written in Python.

TRF uses a probabilistic model of tandem repeats to identify candidate tandem repeats in a sequence and uses statistics similar to those employed by BLAST to evaluate the significance of those candidates. Because TRF employs a probabilistic model, running the program with different parameters can yield different results. To test the robustness of the results from the TRF analysis, we repeated the TRF analysis on a subset of the contigs containing Leviathan sequence, changing, in turn, each parameter by 20% of its original value and then rerunning the program. We also ran the program using sets of parameters found by other labs to produce reliable, reproducible results in searches for tandem repeats in other species.

We manually inspected the repeat regions generated by TRF to ensure that individual arrays were not in fact connected by sequence too degenerate to be picked up by the TRF algorithm. To ensure that arrays of repetitive sequences were counted only once in our analysis of the distribution of repeats within the genomes (below), we merged the database entries corresponding to connected arrays.

Estimating proportion of Repeat Elements in each genome:

The proportion of a genome made up a particular repetitive element was estimated in a two-step process. First, the repeat database constructed above was screened using BLAST, and the total length of any repeat arrays with a significant match (e-value < -25) to the query sequence was recorded. Second, a BLAST database was constructed from the genome after masking all of the tandem arrays identified by TRF. BLAST was used to identify all significant matches to the query sequence (e-value < -25). Because BLAST can truncate good matches separated by regions of dissimilarity, BLAST may underestimate the total length of an actual match. Therefore, each match identified by BLAST was extended by 100bp on either side and the extended sequence and the query were realigned using a custom implementation of the Smith-Waterman algorithm in Python. The lengths of the resulting alignments were added to the lengths of the arrays identified in the first step. To estimate the proportion of each genome consisting of a particular repetitive element, the total lengths found above was divided by the total number of sequenced bases per chromosome in each species.

Phylogenetic trees:

We simultaneously estimated phylogenetic trees and the associated model parameters using genetic algorithms as implemented in GARLI (Zwickl, 2006). We estimated the significance of our results by generating 1000 bootstrap replicates and then constructing a majority-rule consensus tree from the resulting trees using PAUP* (Swofford, 2003).

Test of the Master Gene Hypothesis:

We assessed whether spread of this element was consistent with the master gene hypothesis through the method of Johnson and Brookfield (Johnson and Brookfield, 2006) using the sequence identified by RT-PCR as the putative master gene. Briefly, we compiled a list of the consensus repeat sequence from each repeat region in the *D. pseudoobscura* genome as determined by TRF and aligned them using ClustalW (Chenna *et al.*, 2003) followed by manual corrections of the alignment. Each sequence from this compilation was then compared with each other sequence along with an out-group sequence (the most divergent sequence from the database) and the putative master gene to generate $\binom{n}{2}$ clusters of four sequences.

Under the master gene hypothesis, each observed repeat sequence should share a common ancestor with the putative master gene more recently than with the outgroup. Following Johnson and Brookfield, we assigned each variable site in each cluster as either supporting this hypothesis or better supporting, under the criterion of parsimony, an alternative evolutionary scenario. Supporting sites are sites at which the putative master gene shares a variable site with one or both observed sequences. Alternative sites are those in which the putative master gene shares a variable site with the outgroup or none of the other sequences. We assigned an X,Y coordinate to each set of four sequences based on the number of informative sites compatible with the master-gene hypothesis (Y) and the number of sites that were not (X). We then performed a linear regression on the resulting points.

To assess the statistical significance of the regression, we compared our result to results generated from four sets of 1000 simulated datasets generated by a model corresponding to the

master gene hypothesis using HyPhy (Kosakovsky-Pond *et al.*, 2005). The parameters describing each of the four sets were chosen to mimic those used by Johnson and Brookfield (Johnson and Brookfield, 2006) and span a range of potential distributions each consistent with the master gene hypothesis.

In the first set (A), we assumed a clock-like rate of evolution on all sequences, including the master gene, such that all branch lengths were equal. In the second set (B), branch lengths were taken from the actual dataset. In the third set (C), we assumed a clock-like rate of evolution, but held the master gene constant, simulating strong constraint on the master gene sequence. In the fourth dataset (D), we repeated the first analysis except we used a gamma distribution fitted to the original data to model the distribution of rates within a sequence and to model the possibility of multiple substitutions at some sites.

Empirical confirmation of Leviathan transcription via RT-PCR:

Using a consensus repeat sequence identified by Tandem Repeat Finder, we designed primers to amplify Leviathan from genomic DNA and cDNA, the latter of which was to confirm its current expression. Genomic DNA from the Flagstaff 1993 strain of *D. pseudoobscura* was isolated using the method of Gloor and Engels (Gloor and Engels, 1992). Total RNA was isolated from 25 seven-day posteclosion adult females of the same strain using the Qiagen RNeasy kit. The RNA was reverse transcribed in a reaction containing 5 mM $MgCl_2$, 50 mM KCl, 10mM Tris-HCl, 1mM dNTPs, 20 units RNasin, 20 units reverse transcriptase, and 2.5 µM of primer LeviathanR (5'-TCACGATTTTCGCAAAAAATCATGATGGTTACATC -3'). PCRs were then performed on the cDNA generated above, the genomic DNA, the total RNA, and a negative control using primers LeviathanR and LeviathanF (5'- GCAAATCACAATCTTCGGGAGGC -3'). PCR products were visualized on a TBE agarose gel, and the cDNA amplification product was sequenced using ABI BigDye Terminator version 3.1.

BLAST was used to make sure that the resulting sequence was not found in any annotated coding sequence.

Results

Description of Leviathan:

We identified numerous significant BLAST (Altschul *et al.*, 1990) matches to the repetitive sequence ("Leviathan") identified from the breakpoint region of the fixed inversion difference between *D. pseudoobscura* and *D. persimilis* on the right arm of chromosome X (XR). These matches were found on all four major chromosomes in the two species, both in intergenic regions and within the introns of annotated genes. No significant hits were found in the available genome sequences of other *Drosophila* (Gilbert, 2007) or in NCBI databases, nor were we able to identify similar sequence in Repbase (Jurka *et al.*, 2005) or the RepeatMasker (Smit *et al.*, 1996-2004) server. Manual investigation of regions of the *D. persimilis* and *D. pseudoobscura* genome sequences with multiple hits showed that the individual elements identified by BLAST were often parts of long swaths of repeated units up to more than nine kilobases in length.

To better visualize any patterns within the arrays of repetitive sequence, we ran several contigs containing repeat arrays through the program Dotter (Sonnhammer and Durbin, 1995). The results show a clear periodicity within the repeat regions of ~319bp. These 319bp repeats are highly palindromic, contain motifs reminiscent of the TATA box, transcription factor binding sites, and polyadenylation.

Using Tandem Repeat Finder (TRF) (Benson, 1999, and see Methods), we identified several dozen regions containing repetitive sequence with high similarity to the sequence flanking the inversion on chromosome XR and to each other. These repeat regions are found throughout the genomes of both *D. pseudoobscura* and *D. persimilis*, making up 0.1% to 0.15% of the assembled genome of the two species and nearly 4% of all repetitive sequence, and consist of between 2 and 16 repetitions of Leviathan, often surrounded by and interspersed with degenerate Leviathan sequence (see Methods). We were unable to locate any singleton copies of the element. For contrast we calculated that the two repetitive elements associated with chromosomal rearrangements identified by Richards *et al.* (2005) makes up a total of 0.2% of the assembled genome of *D. pseudoobscura* and 8.7% of all repetitive sequence.

While we were unable to locate singleton copies of the element, TRF still identified the same arrays of ~319bp repeats under a variety of different parameters (see Methods). The robustness of these results, along with the figures produced by Dotter, suggest that, while we cannot rule out longer sequences, 319bp is likely the fundamental size of a single repeat unit.

Empirical confirmation of expression:

Because tandem arrays of Leviathan sequence are often surrounded by degenerate Leviathan sequence, it is difficult to determine exactly where an array begins or ends. Thus, we were unable to determine the exact frame of Leviathan. Nonetheless, we were able to successfully amplify a 184-bp portion of Leviathan sequence from *D. pseudoobscura* cDNA using primers designed from the consensus repeat sequences identified by TRF. Amplification from genomic DNA was also successful, but no amplification from RNA or our negative control was obtained, demonstrating that our amplification from cDNA reflects transcription of the Leviathan sequence and not DNA contamination. There was no single perfect match to the sequence we amplified (GenBank accession EU081847) in either genome sequence, perhaps reflecting the difficulties of incorporating highly repetitive regions into genomic assemblies or a potential heterochromatic origin of the isolated transcript.

As the sequence of the RT-PCR product does not match that of genes predicted to contain the repetitive element (data not shown), our results may constitute evidence for an actively transcribed mobile element. We cannot, however, rule out the possibility that the sequence identified is the result of the transcription of Leviathan sequence proximal to a coding or non-coding gene or transcribed as a part of mechanisms inducing heterochromatin (Grewal and Moazed, 2003, and see Discussion).

Genomic distribution and mechanism of spread:

The distribution of Leviathan can be analyzed by looking both at the distribution of individual 319bp elements and at the sequences arrays (consisting of swaths of 319bp elements of nearly identical sequence) in which they are found. We hypothesize (see Discussion) that the two distributions give different views of how Leviathan spreads in the genome. Individual elements within an array likely owe their existence to mechanisms such as unequal crossing over, whereas disjoint repeat arrays originate with a transpositional or other replicative event. Biases in either distribution thus shed light on the different molecular mechanisms by which Leviathan spreads.

We compared the distribution of arrays within and between the two genome sequence assemblies using the filtered results of TRF (see Methods). We saw no evidence for a non-random distribution of number of Leviathan sequence arrays within the genome of either species after correcting for the differences in lengths between the different chromosomal segments, (χ^2 test, NS).

We did, however, find that the proportion of total Leviathan sequence varied widely between chromosomes within each species, with an excess of Leviathan sequence in the XR and, to a lesser extent, 4th chromosomes (χ^2, p < 2.2e-16, Table 1). We found no evidence for differences in the distribution of either total Leviathan sequence or the number of Leviathan sequence arrays between the genomes of the two species (Kolmogorov-Smirnov Test, NS).

Table 1. The Proportion of Leviathan Sequence per Chromosome Per Species. The total size of each chromosome for each species is given as well as the total length of Leviathan sequence in each chromosome and the total proportion of each chromosome made up of Leviathan sequence. Proportions that differ significantly from the others (see text) are marked with asterisks.

Chromosome	D. pseudoobscura			D. persimilis		
	Total	Repeat	Proportion	Total	Repeat	Proportion
Ch2	30,751,037bp	10,528bp	0.034%	31,203,717bp	5545bp	0.0178%
Ch3	19,758,590bp	3906bp	0.0198%	19,649,734bp	3159bp	0.0161%
Ch4	27,194,902bp	91,392bp	0.337%*	27,622,741bp	69,377bp	0.251%*
ChXL	24,728,506bp	21,967bp	0.0889%	25,734,256bp	23,870bp	0.0928%
ChXR	24,578,333bp	147,472bp	0.60%*	24,376,180bp	77,779bp	0.319%*

Individual repeat arrays had nearly identical repeat sequences, but repeat sequences differed slightly more between arrays. We estimated phylogenetic trees using the consensus repeat sequences identified by TRF in each genome to see if related sequences were clustered in the genome of either species. Our trees do not show any evidence for clustering within or between genomes besides that within arrays, though the presence of extensive polytomies suggests that there may not be sufficient variation between the sequences to detect clustering were it to exist.

We also estimated a tree using sequences from both genomes to evaluate the extent to which Leviathan may be introgressing between the two genomes. The resulting tree does not support clustering by species, though there is some clustering of elements on the XR, indicating a greater divergence of elements on the XR from the arrays on other chromosomes. The lack of clustering by species, coupled with the similarities in the distributions of Leviathan within each of the two species, suggests that interspecies gene flow may be sufficient to homogenize these genomes with respect to this element.

Master gene test:

Finally, we assessed whether spread of this element was consistent with the master gene hypothesis through the regression method of Johnson and Brookfield (Johnson and Brookfield, 2006; see also Materials and Methods). We found a strong negative relationship in our regression, which is evidence *against* the master gene hypothesis (most sites do not support the master gene hypothesis). Because the points used in the regression are not independent of one another, we cannot directly assign a probability value to our results. To do this, we compared our result to results from simulated data using three different models consistent with the master gene hypothesis (see Methods). Three datasets from set A were found to have a more negative relationship than that observed in the actual data (p = 0.001). In none of the other simulations was a single more negative slope observed (p < 0.001). Similar results were obtained when other Leviathan sequences were treated as the putative master gene (results not shown).

Discussion

Our results suggest that Leviathan is an actively transcribed element in the genomes of *Drosophila pseudoobscura* and *D. persimilis* and that its extensive presence in the genomes of these two species may be due both to transcription via an RNA intermediate and an unknown mechanism (perhaps unequal crossing over) that increases the size of a Leviathan element array once it has been inserted in the genome. We cannot definitively rule out that Leviathan is transcribed because of its proximity to a coding or non-coding gene, but its presence on multiple chromosomes of these species, but not related species, suggests recent transcriptional, and later retro-transcriptional, activity.

The similarity of the distributions of Leviathan within the two genomes and patterns of variation seen within Leviathan sequences suggest that Leviathan may introgress between the two genomes. *D. persimilis* and *D. pseudoobscura* are known to hybridize in the wild, and substantial introgression has been detected at some loci (Hey and Nielsen, 2004). Alternatively, the observed pattern may be due to the recent divergence of the two species. In either case, Leviathan is likely to exert the similar evolutionary pressures in both lineages.

Characterization of Leviathan and its spread:

The size and composition of the 319bp repeat constrains the possible classifications of Leviathan. 319bp is likely too short to encode a functional protein, suggesting that Leviathan is a satellite, MITE, or parasitic fragment of a DNA transposon. Because the sequence contains several pol-III termination sequences in both orientations, Leviathan is unlikely to be a SINE. Further characterization must await future genome assemblies in which the terminal ends of a Leviathan array can be better identified and analyzed.

Recent work on the role of tandem repeats in RNAi-based mechanisms of heterochromatin formation (Grewal and Moazed, 2003; Martienssen, 2003; Lippman *et al.*, 2004) raises the intriguing possibility that Leviathan is a lineage-specific component of epigenetic control of gene expression. While we have no evidence to support this claim and consider it unlikely given that Leviathan is not found in related, sequenced genomes, it would explain the presence of an RNA transcript and is intriguing enough to warrant investigation in future studies.

One of the most striking features of Leviathan is the size of the tandem arrays in which it is found and the long stretches of degenerate sequence often seen within arrays. It is likely that while Leviathan acts as some sort of mobile element, much of the total Leviathan sequence in these species owes its existence to ectopic exchange or unequal crossing over rather than transposition *per se*. Such a pattern has been observed in previous studies of repetitive sequence (Cox and Mirkin, 1997), and supports the contention that Leviathan is responsible for the generation of at least one of the fixed chromosomal inversions between these two species. Our finding that the XR contains a greater proportion of Leviathan elements than other chromosomes and that these arrays are often more divergent than arrays on other chromosomes (and thus likely older), further supports this contention.

To definitively categorize and fully characterize Leviathan, we would need to identify the frame as well as the insertion sites to look for characteristic signals such as inverted repeats, poly-A tails, etc. (see review in Kazazian, 2004). We attempted to identify such signals at the ends of the repeat arrays in this study. However, despite extensive search, we were unable to identify single units and, in most cases, were unable to delineate these array ends due both to the degeneracy of the primary sequence at the ends and incomplete coverage of the genomes. In cases where we identified plausible flanking regions to the element insertion, we failed to identify any consistent patterns. In at least one case, we did observe the same sequence, but in opposite orientations, at both ends of the

insertion. This sequence did not BLAST to any other mapped regions of the genome (chromosome 4, group 2, position 850593). However, other forces besides transposition (or retro-transposition) likely affect the spread and sequence of this element (*e.g.*, unequal crossing over). As a result, the signatures of the origin of most Leviathan arrays may be obscured.

We found no evidence to support the Master Gene Hypothesis in our explicit test, nor do our phylogenetic trees support the comb-like topology expected under the master gene hypothesis (Johnson and Brookfield, 2006). This suggests that Leviathan replicates in a manner similar to LTR retroelements or endogenous retroviruses rather than a LINE or Alu elements (Shen *et al.*, 1991; Kass *et al.*, 1995), but no definitive statement is possible without detailed laboratory-bases characterization. Our results are, however, consistent with Leviathan having multiple origins of replication. Previous studies (reviewed in Burt and Trivers, 2006) suggest that repetitive elements with multiple origins of replication generally have negative fitness consequences for their hosts (though see Torkamanzehi *et al.* [1992] for a possible exception). Leviathan is likely no exception, showing that potentially important drivers of evolutionary change and speciation can have negative fitness consequences.

Approach and prospects:

To validate our technique, we looked for the presence of Leviathan in a recently assembled collection of computationally predicted and experimentally verified transposable elements from all twelve sequenced species of *Drosophila* (Quesneville *et al.*, 2005; H. Quesneville, personal communication). This collection identified Leviathan-like sequence on several unassigned contigs in *D. pseudoobscura*, but not in any sequence assigned to a chromosome. This is likely due to the removal of tandem repeat sequence, and other repetitive sequence arrays, in this analysis that, while necessary to avoid false-positives in large scale-surveys, would remove unusual repetitive sequences like Leviathan. Our finding demonstrates the value of computational analyses of individual repetitive elements in conjunction with large-scale searches that, by their very nature, may screen out unusual sequences like Leviathan. We have shown that basic computational tools can be used effectively in identifying new repetitive elements, and other repetitive sequences, when combined with and when validated by empirical methods like RT-PCR. Phylogenetic techniques can be especially powerful for examining how repeat sequences spread in the genome when used in conjunction with methods, such as TRF, for identifying repetitive sequence *ab initio*. In applying TRF to the whole genomes of both *D. persimilis* and *D. pseudoobscura,* we identified a second, 170bp repeat that appears to be unique to *D. persimilis* and is distributed non-randomly in the genome (χ^2 test, p < 0.0001). We did not follow up further with this element, but this result does support a promising future for computational searches for many more interesting repetitive elements.

Conclusions

We identified a novel repetitive element, Leviathan, which is found at the XR inversion breakpoints, appears to be unique to the *D. pseudoobscura* and *D. persimilis* genomes, and may have contributed to the divergence of these species from each other. Leviathan appears to increase in abundance in the genome via both unequal crossing over, which increases the size of arrays of nearly identical Leviathan sequences, and through transposition likely via an RNA intermediate, which leads to the creation of new arrays in new genomic locations. We characterized this element using novel computational tests in additional to traditional laboratory techniques, thus demonstrating the value of

adding computational genomic characterization of repetitive elements to traditional laboratory studies to better understand both the molecular and evolutionary dynamics of repetitive sequence.

References: Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman 1990, J. Mol. Biol. 215: 403-410; Aquadro, C.F., A.L. Weaver, S.W. Schaeffer, and W.W. Anderson 1991, Proc. Natl. Acad. Sci. USA 88: 305-309;   BCM-HGSC. (URL: http://www.hgsc.bcm.tmc.edu/ projects/drosophila/, Houston, TX, 2006); Benson, G., 1999, Nucleic Acids Res 27: 573-580; Burt, A., and R. Trivers 2006, *Genes in Conflict: The Biology of Selfish Genetic Elements*. Belknap Press of Harvard University Press, Cambridge, Mass.; Caceres, M., J.M. Ranz, A. Barbadilla, M. Long, and A. Ruiz 1999, Science 285: 415-418; Charlesworth, B., and D. Charlesworth 1983, Genet. Res. Camb. 42: 1-27; Charlesworth, B., P. Sniegowski, and W. Stephan 1994, Nature 371: 215-220; Chenna, R., H. Sugawara, T. Koike, R. Lopez, T.J. Gibson *et al.*, 2003, Nucleic Acids Res 31: 3497-3500; Coghlan, A., E.E. Eichler, S.G. Oliver, A.H. Paterson, and L. Stein 2005, Trends Genet. 21: 673-682; Cox, R., and S.M. Mirkin 1997, Proc. Natl. Acad. Sci. USA 94: 5237-5242; Deininger, P.L., M.A. Batzer, C.A. Hutchison, 3rd, and M.H. Edgell 1992, Trends Genet. 8: 307-311; Evgen'ev, M.B., H. Zelentsova, H. Poluectova, G.T. Lyozin, V. Veleikodvorskaja *et al.*, 2000, Proc. Natl. Acad. Sci. USA 97: 11337-11342; Gilbert, D.G., 2007, Nucleic Acids Res. 35: D480-D485; Gloor, G.B., and W.R. Engels 1992, Dros. Inf. Serv. 71: 148-149; Grewal, S.I., and D. Moazed 2003, Science 301: 798-802; Grumbling, G., and V. Strelets 2006, Nucleic Acids Res. 34: D484-D488; Hey, J., and R. Nielsen 2004, Genetics 167: 747-760; Johnson, L.J., and J.F. Brookfield 2006, Mol. Biol. Evol. 23: 235-9; Jurka, J., V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany *et al.*, 2005, Cytogenet. Genome Res. 110: 462-467; Kass, D.H., M.A. Batzer, and P.L. Deininger 1995, Mol. Cell. Biol. 15: 19-25; Kazazian, H.H., Jr., 2004, Science 303: 1626-1632; Kosakovsky-Pond, S.L., S.D. Frost, and S.V. Muse 2005, Bioinformatics 21: 676-679; Leman, S.C., Y. Chen, J.E. Stajich, M.A.F. Noor, and M.K. Uyenoyama 2005, Genetics 171: 1419-1436; Lim, J.K., 1988, Proc. Natl. Acad. Sci. USA 85: 9153-9157; Lippman, Z., A.V. Gendrel, M. Black, M.W. Vaughn, N. Dedhia *et al.*, 2004, Nature 430: 471-476; Machado, C.A., T.S. Haselkorn, and M.A.F. Noor 2007, Genetics 175: 1289-1306; Martienssen, R.A., 2003, Nat. Genet. 35: 213-214; Noor, M.A.F., and A.S. Chang 2006, Curr. Biol. 16: R890-R892; Noor, M.A.F., D.A. Garfield, S.W. Schaeffer, and C.A. Machado 2007, Genetics, in press; Ortiz-Barrientos, D., A.S. Chang, and M.A.F. Noor 2006, Genet. Res. Camb. 87: 23-31; Quesneville, H., C.M. Bergman, O. Andrieu, D. Autard, D. Nouaud *et al.*, 2005, PLoS Comput. Biol. 1: 166-175; Ranz, J.M., D. Maurin, Y.S. Chan, M. von Grotthuss, L.W. Hillier *et al.*, 2007, PLoS Biol. 5: e152; Richards, S., Y. Liu, B.R. Bettencourt, P. Hradecky, S. Letovsky *et al.*, 2005, Genome Res. 15: 1-18; Shen, M.R., M.A. Batzer, and P.L. Deininger 1991, J. Mol. Evol. 33: 311-320;   Smit, A.F.A., R. Hubley, and P. Green. (URL: http://www.repeatmasker.org, 1996-2004); Sonnhammer, E.L., and R. Durbin 1995, Gene 167: GC1-10; Swofford, D.L. 2003, Sinauer Associates, Sunderland, MA.; Tamura, K., S. Subramanian, and S. Kumar 2004, Mol. Biol. Evol. 21: 36-44; Torkamanzehi, A., C. Moran, and F.W. Nicholas 1992, Genetics 131: 73-78; Zwickl, D., 2006, in Integrative Biology, University of Texas, Austin, TX.