The 1.688 Repetitive DNA of *Drosophila*: Concerted Evolution at Different Genomic Scales and Association with Genes

Gustavo C. S. Kuhn,*[,1,2] Heinrich Küttler,[3] Orlando Moreira-Filho,[1] and John S. Heslop-Harrison[4]

[1]Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, São Paulo, Brazil

[2]Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[3]Mathematisches Institut, Ludwig-Maximilians-Universität, München, Germany

[4]Department of Biology, University of Leicester, Leicester, United Kingdom

***Corresponding author:** E-mail: gcskuhn@ufmg.br.

Associate editor: Norihiro Okada

Abstract

Concerted evolution leading to homogenization of tandemly repeated DNA arrays is widespread and important for genome evolution. We investigated the range and nature of the process at chromosomal and array levels using the 1.688 tandem repeats of *Drosophila melanogaster* where large arrays are present in the heterochromatin of chromosomes 2, 3, and X, and short arrays are found in the euchromatin of the same chromosomes. Analysis of 326 euchromatic and heterochromatic repeats from 52 arrays showed that the homogenization of 1.688 repeats occurred differentially for distinct genomic regions, from euchromatin to heterochromatin and from local arrays to chromosomes. We further found that most euchromatic arrays are either close to, or are within introns of, genes. The short size of euchromatic arrays (one to five repeats) could be selectively constrained by their role as gene regulators, a situation similar to the so-called "tuning knobs."

Key words: concerted evolution, molecular drive, recombination, satellite DNA, repetitive DNA, *Drosophila melanogaster*.

Arrays of tandemly repeated DNA families are found in most eukaryotes. In many families, sequences in one species show more similarity than members of the same family from a closely related species (see Kuhn et al. 2008 for examples in *Drosophila*), although different repeat families show contrasting behaviour, sometimes with little species specificity in sequence (Kuhn et al. 2010). The phenomenon leading to amplified homogeneous tandem repeats is known as concerted evolution (Zimmer et al. 1980; Dover 1982).

Theoretical models suggest that nonreciprocal mechanisms of recombination, such as unequal crossing over and gene conversion, are important agents that promote intraspecific homogenization (leading to concerted evolution) of arrays for specific repeat variants (Smith 1976; Nagylaki and Petes 1982; Dover 1982). However, most studies have relied on the analysis of mixed (and often small) samples of repeats cloned randomly from different chromosomes and arrays. With the availability of mapped bacterial artificial chromosome (BAC) sequences, we can now examine concerted evolution at the chromosomal or the repeat-array level (short-read shotgun sequences cannot be used because of the impossibility of assembly of the tandem repeats).

The 1.688 satellite DNA, with approximately 15,000 c. 359 bp long repeats, is among the most abundant tandemly repeated DNA in *Drosophila melanogaster* (Brutlag 1980; Lohe and Roberts 1988). Arrays of repeats are located throughout the left heterochromatic arm of chromosome X (Hsieh and Brutlag 1979; Lohe et al. 1993), with other arrays at discrete heterochromatic locations on chromosomes 2 and 3 (Losada and Villasante 1996; Abad et al. 2000). Several arrays are also located within the euchromatin of chromosome X (Waring and Pollack 1987; DiBartolomeis et al. 1992), whereas three euchromatic arrays were inferred to reside on chromosome 2 (Losada and Villasante 1996) and one on chromosome 3 (Koryakov et al. 1999). The high copy number and distribution across the genome make the 1.688 repetitive DNA an excellent model to study concerted evolution at different genomic scales. Here, we analyzed 326 repeat units from 52 arrays on the three chromosomes (fig. 1; supplementary table 1, Supplementary Material online).

Within each array, repeat sequence varied on average by ~6% on chromosome 2, ~5% on chromosome 3, and ~11% on chromosome X. Repeats from different arrays from a single chromosome varied by as much as ~36% on chromosome 2, ~31% on chromosome 3, and ~27% on chromosome X. Such patterns of variability suggest that homogenization and concerted evolution occur at the array level. To investigate the evolutionary relationships between repeats from the same and different arrays under a phylogenetical and statistical framework, we first constructed maximum likelihood (ML) trees separately for each chromosome (fig. 2). The tree topologies showed strong clustering of repeats from the same array. For chromosome 2, where repeats from both arms were available, arrays from the right euchromatic arm clustered on one branch. All trees showed heterochromatic repeats clustering away from repeats in euchromatic arrays.

Letter

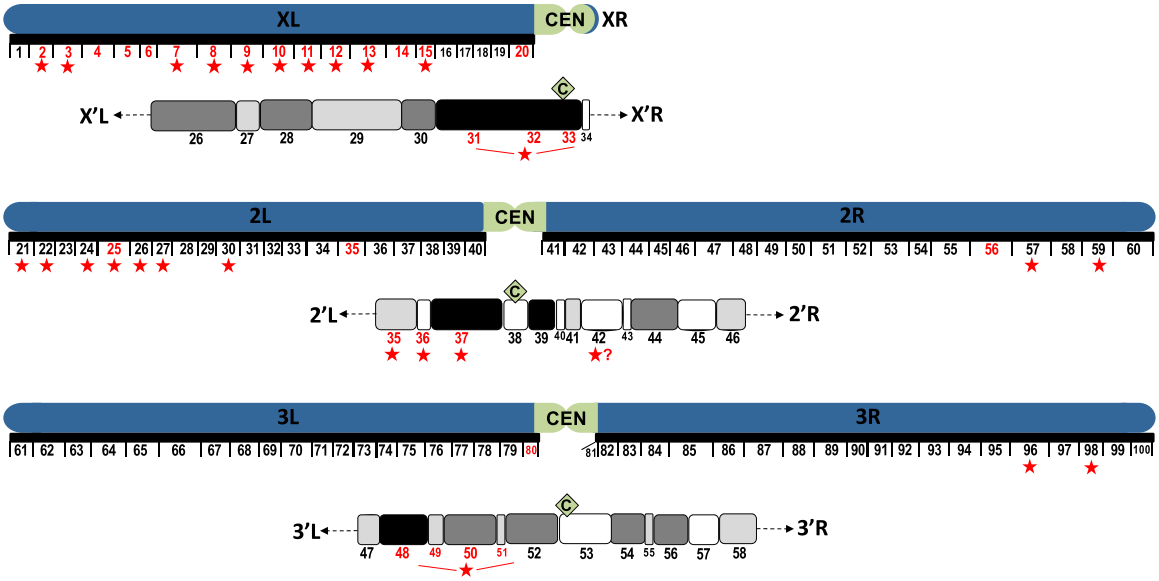

Fig. 1. Cytological map of chromosomes X, 2, and 3 of *Drosophila melanogaster* showing cytological divisions for euchromatin (X, 2, and 3) and heterochromatin (X′, 2′,and 3′) (adapted from Gatti and Pimpinelli 1992; Lohe et al. 1993; www.flybase.org). Divisions in gray (print version) or red (online version) mark the locations of 1.688 repeats described in the literature; stars show 1.688 arrays studied in figures 2 and 3.

An ML tree with all 326 repeats from our sample (fig. 3) revealed that euchromatic repeats from the same chromosome tended to cluster on branches that are closer to each other compared with the ones containing repeats from nonhomologous chromosomes. A few repeats from different chromosomes were interspersed on the same branch, suggesting exchanges between nonhomologous chromosomes through time. The ML tree also showed most heterochromatin repeats clustering on one branch with chromosome specificity, although chromosome 3 repeats were clustered on two branches separated by repeats from chromosome X. Heterochromatic repeats from the same chromosome were generally more related than their euchromatic counterparts. This pattern can be explained by the different organization of 1.688 repeats on these two chromatin domains: In heterochromatin, they form long and continuous arrays, but in euchromatin, they are shorter (see below) and interrupted by long nonhomologous sequences. Interruptions may limit recombination between arrays and consequently the homogenization process (Brutlag 1980; Dover 1982).

Our analysis of the 1.688 arrays, inferred by nucleotide variability and clustering patterns of repeats on ML trees provides strong evidence that homogenization of repeats in arrays occurs differentially and independently for distinct genomic regions, as observed for the alpha satellite in humans (Schueler et al. 2005; Rudd et al. 2006).

Repeats belonging to the 1.688 family are also present in the genome of other species from the *melanogaster* subgroup but as a few hundreds of copies (Lohe and Roberts 1988). Strachan et al. (1985) described 1.688 consensus sequences from *D. melanogaster* and five other species from the *melanogaster* group (*D. teissieri*, *D. mauritania*, *D. simulans*, *D. orena*, and *D. yakuba*) and showed that they all evolved by concerted evolution. The range of variability between 1.688 repeats from *D. melanogaster* and the other species is between 23% and 34%, on average. The 1.688 consensus repeats from the other five species (where the chromosomal location and position in their arrays-of-origin is unknown) do not represent a diverged outgroup or sister clade but rather were placed on the branches of the ML tree containing euchromatic repeats from *D. melanogaster* (fig. 3). Therefore, it is possible that they were also derived from euchromatic arrays.

Previous studies of five arrays from chromosome X revealed that they are made of a small number of tandem repeats (two to four) located adjacent to genes (Waring and Pollack 1987; DiBartolomeis et al. 1992). We found that short arrays with under six repeats are the most common for euchromatic arrays (supplementary fig. 1, Supplementary Material online), contrasting with the thousands of repeats residing in heterochromatic regions. There are several examples in *Drosophila* and other organisms showing that the density of repetitive DNA elements is lower in genomic regions associated with intermediate to high recombination rates (Gatti and Pimpinelli 1992; Lohe et al. 1993; Kuhn et al. 2008). This phenomenon is more likely a consequence of selection against deleterious effects of chromosomal rearrangements caused by ectopic recombination between repeats (Charlesworth et al. 1986; Dolgin and Charlesworth 2008). In other words, high recombination rates and purifying selection is expected to eliminate or limit the propagation of tandem repeats. If so, why the genome of *D. melanogaster* is replete with euchromatic 1.688 arrays with an almost constant low number of repeats? To address this question, we investigated the association of 1.688 with genes and found that most arrays on chromosomes 2, 3, and X are either very close to genes or show transcriptional

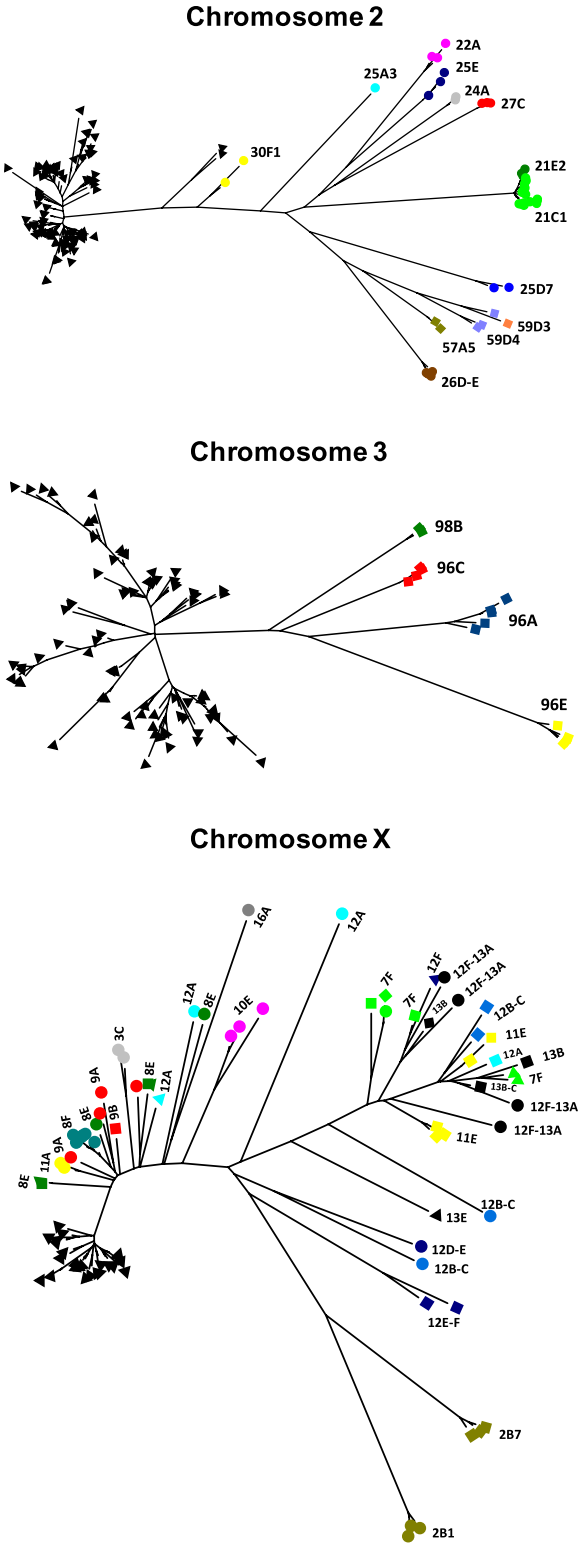

Fig. 2. ML trees of 1.688 repeats from chromosomes 2, 3, and X under the TIM1+G, TPM2uf+G and TVM+I+G models, respectively. Heterochromatic repeats (dark triangles) are mostly grouped and lie away from euchromatic repeats (colours and other symbols). Chromosomes 2 and 3: circles and squares represent repeats from the left and right arm, respectively. Chromosome X: different symbols with same colour represent repeats from different arrays but same cytological division. Numbers refer to locations of arrays (see fig. 1). The colour version of this figure is available online.

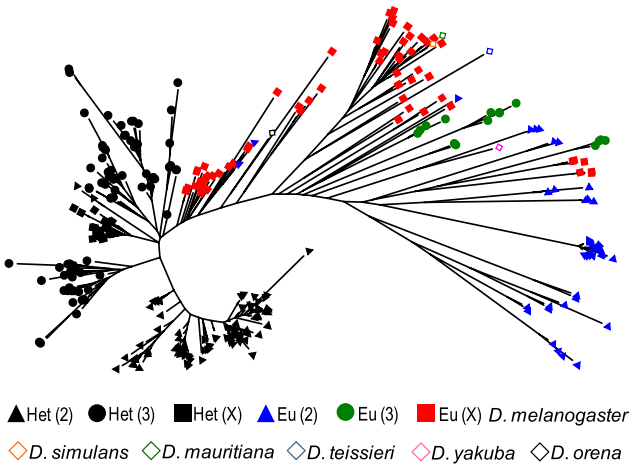

Fig. 3. ML tree of all 1.688 repeats analyzed under the TVM+I+G model. Het, heterochromatin and Eu, euchromatin. The colour version of this figure is available online.

activity (fig. 4; supplementary table 2, Supplementary Material online). However, in all transcribed cases, sequence analysis showed the arrays resided in intronic regions.

Several examples show that tandem arrays affect the expression of genes in a dose-dependent manner (e.g., Maiorano et al. 1997; Rockman and Wray 2002; Vinces et al. 2009). Such a property is likely to be a consequence of their interaction with transcription factors (reviewed by Tomilin 2008; Plohl et al. 2008). Moreover, transcripts derived from tandem repeats have been suggested to play a role in gene regulation (reviewed by Pezer et al. 2011) and Usakin et al. (2007) have recently showed abundant small RNAs transcribed by 1.688 repeats in *D. melanogaster*. We suggest that the limited number of 1.688 tandem repeats on each euchromatic array could be selectively constrained by their possible role as gene regulators. Such a situation would be similar to the so-called "tuning knobs" described for micro- and mini-satellite repeats (King et al. 1997). Thus, divergence of repeats by mutation and local concerted evolution could play an important role for diversification of arrays and regulation of specific genes, as well as leading to evolutionary divergence between species.

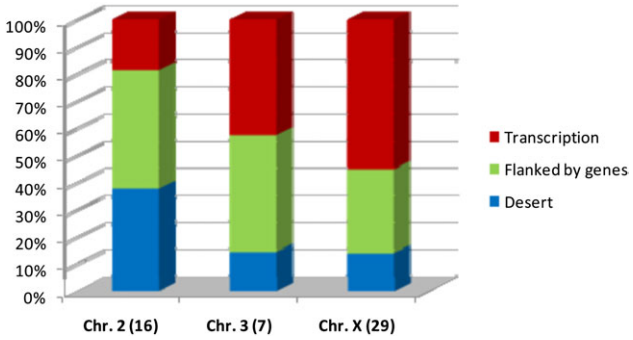

Fig. 4. Proportion of 1.688 arrays that falls into the three genomic landscape classes defined in the present work. Numbers within brackets correspond to the number of analyzed arrays from each chromosome.

Methodology

Basic local alignment search tool (BLAST) searches with repeats representing the four main 1.688 variants from chromosomes X (Hsieh and Brutlag 1979), 2 (Abad et al. 2000), and 3 (Losada and Villasante 1996) were used to identify large sequenced clones (mostly BACs) of *D. melanogaster* with known (by in situ hybridizations) or presumable cytological location featuring at least one "full-length" 1.688 repeat, *e* values lower than 10^{-5} and minimum homology of ~70% over a segment covering at least ~80% of the query sequence. The 1.688 arrays within each clone were investigated, one by one, using dotplots (Dotlet program) of the clone sequence against itself or against selected 1.688 consensus sequences, using 15 bp windows and >~65% homology. The cytological division of each clone was checked by BLASTs of repeats and flanking sequences from each array against the genome of *D. melanogaster* (release dmel r.28_FB2010_05), using the presence of annotated genes surrounding the arrays as supporting evidence (supplementary table 2, Supplementary Material online). In parallel, we used the same 1.688 consensus repeats to BLAST the *D. melanogaster* assembled heterochromatin (release r5.29_FB2010_06). We used the "U-File" to retrieve a sample of heterochromatic repeats from the X chromosome. Additional 1.688 repeats were obtained from GenBank based on published data. All selected 1.688 sequences are listed in supplementary table 1 (Supplementary Material online) and can be provided under request.

Nucleotide variability was calculated by the proportion of different nucleotide sites using the MEGA software (Tamura et al. 2007). Nucleotide sites containing indels were analyzed according to the "pairwise-deletion" option in MEGA. ML trees were estimated with the PhyML 3.0 software (Guindon and Gascuel 2003) using best-fit models inferred by the jModelTest 0.1.1 software (Posada 2008). Branch support was evaluated by bootstrap analysis (100 replicates). Only full-length 1.688 repeats were used for the construction of ML trees. The genomic landscape of 1.688 arrays was inferred by the "Genomeview" browser against the latest assembled genome of *D. melanogaster* (www.flybase.org), using 10 kb windows. The results were divided in three groups: 1) "desert" when no annotated genes were present, 2) "flanked by genes" when a gene was found on at least one side of the array, and 3) "transcription" when the DNA sequence matches RNA transcripts from genes. The "Genome Decorated-Fasta region" browser (www.flybase.org) was used to determine whether transcribed arrays correspond to intronic or exonic regions of the genes.

Supplementary Material

Supplementary tables 1 and 2 and figure 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Fundação de Apoio a Pesquisa do Estado de São Paulo (FAPESP) (Grant 2009/08738-3).

References

Abad JP, Agudo M, Molina I, Losada A, Ripoll P, Villasante A. 2000. Pericentromeric regions containing 1.688 satellite DNA sequences show anti-kinetochore antibody staining in prometaphase chromosomes of *Drosophila melanogaster*. *Mol Gen Genet*. 264:371–377.

Brutlag D. 1980. Molecular arrangement and evolution of heterochromatic DNA. *Annu Rev Genet*. 14:121–144.

Charlesworth B, Langley CH, Stephan W. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics* 112:947–962.

DiBartolomeis SM, Tartof KD, Jackson FR. 1992. A superfamily of *Drosophila* satellite related (SR) DNA repeats restricted to the X chromosome euchromatin. *Nucleic Acids Res*. 20:1113–1116.

Dolgin ES, Charlesworth B. 2008. The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* 178:2169–2177.

Dover GA. 1982. Molecular drive: A cohesive mode of species evolution. *Nature* 199:111–117.

Gatti M, Pimpinelli S. 1992. Functional elements in *Drosophila melanogaster* heterochromatin. *Annu Rev Genet*. 26:239–275.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Hsieh T, Brutlag D. 1979. Sequence and sequence variation within the 1.688 g/cm3 satellite DNA of *Drosophila melanogaster*. *J Mol Biol*. 135:465–481.

King DG, Soller M, Kashi Y. 1997. Evolutionary tuning knobs. *Endeavour*. 21:36–40.

Koryakov DE, Alekseyenko AA, Zhimulev IF. 1999. Dynamic organization of the beta-heterochromatin in the *Drosophila melanogaster* polytene X chromosome. *Mol Gen Genet*. 260:503–509.

Kuhn GCS, Schwarzacher T, Heslop-Harrison JS. 2010. The non-regular orbit: three satellite DNAs in *Drosophila martensis* (*buzzatii* complex, *repleta* group) followed three different evolutionary pathways. *Mol Gen Genomics*. 284:251–262.

Kuhn GCS, Sene FM, Moreira-Filho O, Schwarzacher T, Heslop-Harrison JS. 2008. Sequence analysis, chromosomal distribution and long-range organization show that rapid turnover of new and old pBuM satellite DNA repeats leads to different patterns of variation in seven species of the *Drosophila buzzatii* cluster. *Chromosome Res*. 16:307–324.

Lohe AR, Hilliker AJ, Roberts PA. 1993. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaster*. *Genetics* 134:1149–1174.

Lohe AR, Roberts PA. 1988. Evolution of satellite DNA sequences in *Drosophila*. In: Verma RS, editor. Heterochromatin: molecular and structural aspects. Cambridge: Cambridge University Press. p. 148–186.

Losada A, Villasante A. 1996. Autosomal location of a new subtype of 1.688 satellite DNA of *Drosophila melanogaster*. *Chromosome Res*. 4:372–483.

Maiorano D, Cece R, Badaracco G. 1997. Satellite DNA from the brine shrimp Artemia affects the expression of a flanking gene in yeast. *Gene* 189:13–18.

Nagylaki T, Petes TD. 1982. Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics* 10:315–377.

Pezer Z, Brajković J, Feliciello I, Ugarković D. 2011. Transcription of satellite DNAs in insects. *Prog Mol Subcell Biol*. 51:161–178.

Plohl M, Luchetti A, Mestrović N, Mantovani B. 2008. Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. *Gene* 409:72–82.

Posada D. 2008. jModelTest: Phylogenetic model averaging. *Mol Biol Evol*. 25:1253–1256.

Rockman MV, Wray GA. 2002. Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol*. 19:1991–2004.

Rudd MK, Wray GA, Willard HF. 2006. The evolutionary dynamics of alpha-satellite. *Genome Res*. 16:88–96.

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, NISC Comparative Sequencing Program, Rocchi M, Willard HF, Green ED. 2005. Progressive proximal expansion of the primate X chromosome centromere. *Proc Natl Acad Sci U S A*. 102: 10563–10568.

Smith GP. 1976. Evolution of repeated DNA sequences by unequal crossover. *Science* 191:528–535.

Strachan T, Webb D, Dover G. 1985. Transition stages of molecular drive in multiple-copy DNA families in *Drosophila*. *EMBO J*. 4:1701–1708.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 24:1596–1599.

Tomilin NV. 2008. Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. *BioEssays* 30: 338–348.

Usakin L, Abad J, Vagin VV, de Pablos B, Villasante A, Gvozdev VA. 2007. Transcription of the 1.688 satellite DNA family is under the control of RNA interference machinery in *Drosophila melanogaster* ovaries. *Genetics* 176:1343–1349.

Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. 2009. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* 324:1213–1216.

Waring GL, Pollack JC. 1987. Cloning and characterization of a dispersed, multicopy, X chromosome sequence in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*. 84: 2843–2847.

Zimmer EA, Martin SL, Beverly SM, Kan YW, Wilson AC. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc Natl Acad Sci U S A*. 77:2158–2162.