Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 334–343
GENOMICS
www.elsevier.com/locate/ygeno

Parent genes of retrotransposition-generated gene duplicates in *Drosophila melanogaster* have distinct expression profiles

Morgan G.I. Langille [1], Denise V. Clark *

Department of Biology, University of New Brunswick, Fredericton, Canada NB E3B 6E1

Received 20 November 2006; accepted 5 June 2007
Available online 12 July 2007

Abstract

Genes arising by retrotransposition are always different from their parent genes from the outset. In addition, the cDNA must insert into a region that allows expression or it will become a processed pseudogene. We sought to determine whether this class of gene duplication differs from other gene duplications based on functional criteria. Using amino acid sequences from *Drosophila melanogaster*, we identified retroduplicated gene pairs at various levels of sequence identity. Analysis of gene ontology annotations showed some enrichment of retroduplications in the cellular physiological processes class. Retroduplications show a higher level of nucleotide substitution than other gene duplications, suggesting a higher rate of divergence. Remarkably, analysis of microarray data for gene expression during embryogenesis showed that parent genes are more highly expressed relative to their retroduplicated copies, tandem duplications, and all genes. Furthermore, an expressed sequence tag library representation shows a broader distribution for parent genes than for all other genes and, as found previously by others, retroduplicated gene transcripts are found most abundantly in testes. Therefore, in examining retroduplicated gene pairs, we have found that parent genes of retroduplications are also a distinctive class in terms of transcript expression levels and distribution.

Keywords: Gene duplication; Retroelements; *Drosophila melanogaster*

Gene duplication is considered a major contributor to genome evolution and consequent organismal diversification. The core of the idea was presented by Ohno in 1970 [1]: once established, a newly duplicated gene can be inactivated by mutation or acquire a new function without reducing fitness. The most common path is inactivation, but the rarer path of acquiring a new function could then lead to diversification. Ohno's ideas have since been refined to include other models of diversification. More recent concepts include the idea that both copies can change. For example, the subfunctionalization model predicts that mutations in gene regulatory regions can occur in both genes so that their expression patterns become complementary [2,3]. For each gene, these are partial loss-of-function mutations, so that the complementarity of their expression patterns forces both genes to be maintained. The idea of subfunctionalization has also been applied to amino acid sequence, and computational methods for measuring the distribution of divergence between paralogs have been developed to test this model [4,5].

Gene duplication can occur on the whole-genome scale, on blocks of genes, or on single genes. Single gene duplication can occur by unequal crossing over to produce tandem duplications. Tandem duplications may diverge, but they can also maintain sequence similarity through gene conversion. If the gene is duplicated in its entirety, then both copies are initially identical and functional. Single gene duplication can also occur through retrotransposition, whereby reverse transcription of the mRNA from a parental gene converts it into a cDNA, which is then inserted into chromosomal DNA, forming an intronless paralog [6]. In contrast to tandem duplication, the retrotransposed gene may not carry sequences sufficient for its transcription. To be expressed, the cDNA precursor must be inserted into a transcribed region, have an internal promoter sequence, or acquire transcriptional activity through mutation. Otherwise, the new gene duplication is destined to become a pseudogene.

* Corresponding author. Fax: +1 (506) 453 3583.
E-mail address: clarkd@unb.ca (D.V. Clark).
[1] Current address: Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, Canada BC V5A 1S6.

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.06.001

The availability of complete eukaryotic genome sequences has generated opportunities for exploring gene duplication in a systematic way. Lynch and Conery [7] reported the first genome-wide analysis of gene duplications using three completely sequenced and three partially sequenced genomes. Analysis of nucleotide substitutions for 462 duplications in *Drosophila melanogaster* showed that duplications arise at a rate of about 31 per million years and have a half-life of 2.9 million years. Rubin et al. [8] calculated that 5536 of 13,601 genes arose by gene duplication in *D. melanogaster*. In contrast to Lynch and Conery [7], Rubin et al. identified a larger set of duplications because they included multigene families in their dataset and clustered sequences that matched with a higher BLAST *E* value (10^{-6} vs 10^{-10}).

Other whole-genome studies of duplications showed that the yeast *Saccharomyces cerevisiae* has undergone whole-genome duplication with subsequent loss and diversification of duplicates [9]. This latter mode of gene duplication seems to also account for a portion of the gene duplications in the genome of the nematode *Caenorhabditis elegans,* but was not detectable for the *D. melanogaster* genome in which tandem duplication of single genes was more often observed [10]. The *Drosophila* genome has several types of retrotransposons [11] and, since active elements associated with retrovirus-like particles can exist [12,13], gene duplications can also arise through retrotransposition.

The mechanism of gene duplication by retrotransposition has been studied in the yeast *S. cerevisiae* [14]. Here, retrotransposition is mediated by retrotransposon sequences and reverse transcriptase, as evidenced by Ty1 element sequences flanking the duplicated gene, tracts of poly(A) sequences downstream from the coding sequence, and an increase in duplication rate upon induction of a high level of Ty1 reverse transcriptase expression. However, even in these newly generated duplications, the poly(A) sequences are not always found and the flanking Ty1 sequences are not arranged in a way so that integration could occur as it does for wild-type Ty1 elements. Thus, analysis of retrotransposition events in yeast has not provided evidence for a simple, unifying model for the mechanism of retrotransposition mediated by the Ty1 retrotransposon.

Gene duplications arising by retrotransposition were examined in humans with the initial release of the genome sequence, in which 97 functional intronless paralogs were identified [15]. In this group of genes, there is an excess of translation and nuclear regulation proteins and metabolic and regulatory enzymes. In *D. melanogaster*, whole-genome analysis resulted in the characterization of 24 gene duplications that appear to have been generated by retrotransposition [16]. These gene pairs fit the criteria that the two genes are on different chromosomes, they have at least 70% amino acid sequence identity, one member has no introns, and, in a few cases, there are signs of retrotransposition, as poly(A) tracts, for gene pairs in which both are intronless. Analysis of expression data and chromosome linkage showed that there was a significant tendency for genes on the X chromosome to produce new copies on the autosomes, and the new copies examined are mostly expressed in the testes. This observation is consistent with the hypothesis that genes on the X chromosome are escaping the X-chromosome inactivation that is thought to occur during spermatogenesis [16,17].

With the continued expansion of the *Drosophila* genome project, we now have the most comprehensive developmental expression data to date from microarray analysis [18], expressed sequence tag (EST) sequences from a wider range of libraries [19], and systematic functional annotations in the form of gene ontology (GO) descriptions [20]. We have combined these data with an analysis of gene duplications, focusing on possible retrotransposed duplications as a subset. Our analysis shows that the parent genes of retrotransposed genes are distinct in having a consistently higher level of expression.

Results

Identification of retrotransposed genes

Assembly of a set of duplicated gene pairs first involved an all-against-all comparison of *Drosophila* protein sequences using a global alignment algorithm. Cluster analysis was then performed to identify gene pairs and gene families with more than two members. Since it would be difficult to determine the parent/child relationship in gene families with more than two members with similar levels of amino acid identity, these families were excluded from the gene duplication datasets we used for further analysis. The cluster analysis was performed at several levels of amino acid sequence identity to derive datasets of gene duplications at the 50, 60, and 70% levels. The gene duplication datasets were then subdivided into potentially retrotransposed versus all others by two filters. These filters were (1) a minimum intergenic distance of 100,000 bp if the two genes were on the same chromosome arm and (2) one member of the gene pair having no introns in the amino acid coding region. After filtering, there were 67 gene pairs at the 50% amino acid sequence identity cutoff, 39 pairs at 60% identity, and 20 pairs at 70% identity (see Supplemental File 1). These gene pairs include the changes introduced after updating the sequence dataset with FlyBase release 5.1. One pair was removed (*CG32713* and *CG12725*) as it now formed a cluster of three genes. Seven new pairs were identified at the 50% cutoff, but only one of these pairs met the filtering criteria (*CG34132* and *Tim13*). This update did not change any of the conclusions for the data analyses in this paper.

We found that a minimum intergenic distance of 100,000 bp was a natural cutoff for deriving our subset of gene duplications by retrotransposition after inspecting the distribution of intergenic distances for gene pairs with introns versus gene pairs in which one gene has no introns. A plot of intergenic distances in log base pairs between each pair of duplicated genes on the same chromosome shows there is a bimodal distribution (Fig. 1A). The majority of duplicated genes have an intergenic distance of less than 100,000 bp. For those pairs in which both genes contain introns, 74.9% of them have an intergenic distance of less than 100,000 bp (Fig. 1B). In contrast, only 36.4% of duplications with one gene containing an intron and the other gene containing no introns had an intergenic distance of less than 100,000 bp