LETTER

The Contribution of LTR Retrotransposon Sequences to Gene Evolution in Mus musculus

Jeremy D. DeBarry, Eric W. Ganko, Eugene M. McCarthy, and John F. McDonald
Department of Genetics, University of Georgia

Approximately 1.5% of mouse genes (Mus musculus) contain long terminal repeat retrotransposon sequences (LRS). Consistent with earlier findings in Caenorhabditis elegans, Drosophila melanogaster, and Homo sapiens, LRS are more likely to be associated with newly evolved genes. Evidence is presented that LRS are often recruited as novel exons or as spliced additions to existing exons. These novel gene configurations may be expressed initially as alternative transcripts providing an opportunity for the evolution of new gene function.

Background

Once considered parasitic sequences of little or no functional significance (Doolittle and Sapienza 1980; Orgel and Crick 1980; Charleston, Sniegowski, and Stephan 1994), transposable elements (TE) are now widely recognized as significant contributors to gene evolution (McDonald 1993; Brosius 1999; Kidwell and Lisch 2001; Bowen et al. 2003; Makalowski 2003; Kazazian 2004). It has recently been reported that retrotransposon sequences contribute to ~4% of protein-coding regions (Nekrutenko and Li 2001), ~27% of untranslated regions (van de Lagemaat et al. 2003), and ~25% of promoter regions (Jordan et al. 2003) of human genes. We recently reported that long terminal repeat retrotransposon sequences (LRS) (defined as full-length elements or fragments of full-length elements) are present within the regulatory region and/or the transcription boundaries of 0.6% of Caenorhabditis elegans genes (Ganko et al. 2003) and 1.8% of Drosophila genes. (Ganko et al. 2006). Here we report on the contribution of LRS to transcribed regions of genes in the sequenced genome of the mouse (Mus musculus).

LRS Are Components of Many Mouse Genes

Recently, 21 families of long terminal repeat (LTR) retrotransposons have been identified in the mouse genome, including 13 not previously described (McCarthy and McDonald 2004). Genes with at least one fully sequenced mRNA from the mouse Unigene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) were Blasted against consensus sequences representing each of these 21 families of mouse LTR retrotransposons. Of the 18,374 Unigenes examined, 11,341 had a homolog assigned by the Homologene data set (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene). Homologenes are genes associated with functions that are generally conserved among even phylogenetically diverse groups of species (Wheeler et al. 2003). Of the 18,374 Unigenes examined, 11,341 had a homolog assigned by the Homologene data set, 88 (~0.8%) of these contained an LRS. Thus, LRS are associated with homologenes about half as frequently as they are with all mouse genes.

Many LRS Are Located in Mouse Exons

Our results indicate LRS are located within the coding regions of many mouse genes. Such associations are believed to have arisen directly by insertion of an LRS into an existing exon or indirectly by exon recruitment of an LRS from an adjacent intron or untranslated leader region (ULR) (Nekrutenko and Li 2001). To estimate the number of LRS located within the coding regions of mouse genes, a database of mouse exonic sequences was obtained from Ensembl (http://www.ensembl.org/Multi/martview). If the predicted transcripts or proteins display significant similarity to species-specific Swiss-Prot, RefSeq, or TrEMBL entries, Ensembl classifies them as “known” otherwise they are classified as “novel” genes. At the time of this analysis (September 2004), the Ensembl database was composed of 25,307 mouse genes. Of these, 20,166 were subclassified as known and 5,141 as novel genes. Many of the novel genes in the Ensembl database are unannotated. Thus, as a precaution against overestimating the contribution of LRS to mouse gene evolution, we limited analysis to well-annotated genes (Ensembl known genes). The 20,166 annotated (known) genes in the Ensembl data set contain 186,823 exons. A total of 239 of these genes are associated with LRS located in 263 exons. Ten of these exons have two independently inserted LRS, and one has three (i.e., 275 associations). We found 22 of these 275 associations (22/275 or 8.0%) were composed of ≥95% LRS. Exons...
Table 1  
LRS-Associated Mouse Genes Encoding Multiple Transcripts Rarely Contain LRS in All Alternative Transcripts

<table>
<thead>
<tr>
<th>Number of Alternative Transcripts/Gene</th>
<th>Number of Alternative Transcripts with LRS</th>
<th>Number of LRS-Associated Genes</th>
<th>Percentage of Genes with LRS in All Alternative Transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>37</td>
<td>24.5 (12/49)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>12</td>
<td>12.5 (2/16)</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>0 (0/3)</td>
</tr>
<tr>
<td>5—8</td>
<td>0</td>
<td>0</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0 (0/1)</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0 (0/1)</td>
</tr>
</tbody>
</table>

Note.—A database of mouse alternate transcript information that included the exons present in each transcript and a count of the total number of transcripts for each gene were assembled using Ensembl data. We queried each LRS-associated exon against this database to determine the number of alternative transcripts containing an LRS (number of alternative transcripts with LRS). The number of genes with LRS-associated relative to the number of LRS-associated genes was also calculated. The results indicate that very few LRS-associated genes that encode multiple transcripts contain LRS in all transcripts. Only genes known to encode multiple transcripts were included.

composed almost exclusively of LRS are considered to have been recruited as novel exons from LRS located in adjacent introns or ULRs (e.g., Nekrutenko and Li 2001). Of the remaining LRS associated with exons, 55 (55/275 or 20.0%) were located at the 5‘ or 3‘ exon boundaries. LRS located at exon boundaries are considered to have been added to preexisting exons due to the presence of appropriate splice acceptor or donor sites (e.g., Nekrutenko and Li 2001). The remaining 198 (72.0%) LRS associated with exons, including the 11 exons containing more than one LRS, are located within the body of the exon. The origins of these LRS are more difficult to reconstruct, but some are likely to represent insertions into preexisting exons.

The addition of LRS to preexisting exons may be expected to disrupt gene function and be eliminated by natural selection. One possible hypothesis to explain the maintenance of relatively high number of LRS associated with exons is that they are tolerated by natural selection at loci encoding multiple alternative transcripts. Under such a scenario, an inserted sequence would, due to the presence of appropriate splice acceptor/donor sites, be associated with generation of one or more novel alternative transcripts while the native transcript maintains the original gene function. Over evolutionary time, a novel transcript containing the LRS may evolve to encode a function favored by natural selection and thus be selectively maintained in conjunction with, or in place of, the original transcript. Under this hypothesis, alternative transcripts generated by TE insertions may provide an opportunity for the evolution of new gene functions in a manner similar to what has been proposed for gene duplications (Ohno 1970). Consistent with this view, we found that those mouse genes confirmed to encode alternative transcripts rarely contained LRS in all transcripts (table 1).

LRS Are Preferentially Associated with Genes Encoding Metabolic Functions

Functional information from the Gene Ontology (GO) Consortium (http://www.geneontology.org/) was used to investigate possible functional trends among genes associated with LRS. GO networks are composed of three main functional classifications: molecular function, cellular component, and biological process. A number of subclasses are listed under each of these classifications. Based on the observed frequency of all experimentally verified genes in the Unigene database that group under each of the GO classifications, we computed the number of LRS-associated genes expected to group under each GO classification. This expected number was compared with the observed number to identify significant differences within the subclasses of each main classification. Only subclasses within the “biological process” classification demonstrated a significant difference between observed and expected numbers of associations (chi square = 30.05, df = 6, \( P < 0.025 \)).

The results presented in table 2 show mouse genes grouped under the biological process classification that encode physiological functions associated with LRS more frequently (251 observed, 0.68 success probability, 307 trials, \( P < 0.025 \)), while genes encoding cellular processes are associated with LRS less frequently (39 observed, 0.24 success probability, 307 trials, \( P < 0.025 \)) than expected. Significant deviations from expected numbers were also observed in two additional subclasses of physiological function. Genes associated with LRS that encode cell growth and maintenance functions are less frequent, while LRS-associated genes encoding metabolic functions were more frequent than expected.

Table 2  
Biological Process GO Terms for Mouse Genes Associated with LRS

<table>
<thead>
<tr>
<th>Biological Process</th>
<th>Expected</th>
<th>Observed</th>
<th>Probability of Success</th>
<th>Trials</th>
<th>( P ) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0007582 Physiological Process</td>
<td>209</td>
<td>251</td>
<td>0.680</td>
<td>307</td>
<td>( 3 \times 10^{-2} )</td>
</tr>
<tr>
<td>GO:00008151 Cell growth and/or maintenance</td>
<td>55</td>
<td>24</td>
<td>0.217</td>
<td>251</td>
<td>( 2 \times 10^{-6} )</td>
</tr>
<tr>
<td>GO:0008152 Metabolism</td>
<td>159</td>
<td>190</td>
<td>0.632</td>
<td>251</td>
<td>( 1 \times 10^{-4} )</td>
</tr>
<tr>
<td>GO:0009987 Cellular Process</td>
<td>75</td>
<td>39</td>
<td>0.243</td>
<td>307</td>
<td>( 2 \times 10^{-6} )</td>
</tr>
</tbody>
</table>

Note.—Shown are the numbers of subordinate descriptors from two biological process terms where there was a significant difference between expected and observed values for mouse genes associated with an LRS. The expected value is based on the ratio of descriptor GO terms for all mouse genes in the biological process ontology. Binomial tests (with Bonferroni multicomparison correction) were used to determine observed descriptors that were significantly different than expectations for a given number of trials (total GO terms in a descriptor subclass, Ganko et al. 2006).
Summary and Conclusions

Consistent with earlier studies of Homo sapiens (Nekrutenko and Li 2001), Drosophila (Ganko et al. 2006), and C. elegans (Ganko et al. 2003) genomes, we found LRS are contained within the coding regions of a significant proportion of mouse genes. Also consistent with earlier findings (Waterston et al. 2002; van de Lagemaat et al. 2003), our results indicate LRS may be preferentially associated with more recently evolved genes. The mechanisms by which LRS have been incorporated into genes over evolutionary time are likely to be varied and complex. However, our results suggest the recruitment of LRS as novel or spliced additions to existing exons is likely a primary mechanism. We propose that these novel gene configurations may be expressed initially as alternative transcripts, providing an opportunity for the evolution of new gene functions.

Literature Cited


Pierre Capy, Associate Editor

Accepted December 12, 2005