Transposable Elements: Insertion Pattern and Impact on Gene Expression Evolution in Hominids

Maria Warnefors,1 Vini Pereira,2 and Adam Eyre-Walker*,1
1Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton, United Kingdom
2Apple Barn, Whimpwell Street, Happisburgh, Norfolk, United Kingdom

*Corresponding author: E-mail: a.c.eyre-walker@sussex.ac.uk.
Associate editor: James McInerney

Abstract

Transposable elements (TEs) can affect the regulation of nearby genes through several mechanisms. Here, we examine to what extent recent TE insertions have contributed to the evolution of gene expression in hominids. We compare expression levels of human and chimpanzee orthologs and detect a weak increase in expression divergence (ED) for genes with species-specific TE insertions compared with unaffected genes. However, we show that genes with TE insertions predating the human–chimpanzee split also exhibit a similar increase in ED and therefore conclude that the increase is not due to the transcriptional influence of the TEs. These results are further confirmed by lineage-specific analysis of ED, using rhesus macaque as an outgroup: Human–chimpanzee ortholog pairs, where one ortholog has suffered TE insertion but not the other, do not show increased ED along the lineage where the insertion occurred, relative to the other lineage. We also show that genes with recent TE insertions tend to produce more alternative transcripts but find no evidence that the TEs themselves promote transcript diversity. Finally, we observe that TEs are enriched upstream relative to downstream of genes and show that this is due to insertional bias, rather than selection, because this bias is only observed in genes expressed in the germ line. This provides an alternative neutral explanation for the accumulation of TEs in upstream sequences.

Key words: transposable elements, expression divergence, alternative splicing, insertion bias, Homo sapiens, Pan troglodytes.

Introduction

Almost half of the human genome is made up of transposable elements (TEs) (Lander et al. 2001). These DNA sequences are able to insert into a new genomic location through the process of transposition. Although most such insertions are likely to be subsequently lost due to selection or genetic drift, our lineage has still accumulated more than 7500 TE copies since the split from chimpanzees (Mills et al. 2006), with three TE families accounting for more than 95% of these transposition events: the Long Interspersed Element 1 (L1), the Alu element, which belongs to the Short Interspersed Elements (SINEs), and the SVA element (SINE-R, variable number of tandem repeats and Alu).

TEs have commonly been viewed as selfish parasites, whose persistence in the genome is best explained by their success as replicating units, rather than any benefit they might bestow on the host (Doolittle and Sapienza 1980; Orgel and Crick 1980). Indeed, the presence of TEs can severely impair genome function, either by direct disruption of functional sequences (Kazazian et al. 1988) or by promoting ectopic homologous recombination, which can lead to potentially harmful duplications, deletions, and genome rearrangements (Hedges and Deininger 2007).

On the other hand, some TE-derived sequences are among the most conserved elements of the human genome (Kamal et al. 2006; Lowe et al. 2007), suggesting that some TEs are functional. In particular, they seem to play a role in transcriptional regulation by providing genes with promoters and enhancers (Jordan et al. 2003; van de Lagemaat et al. 2003; Bejerano et al. 2006; Bourque et al. 2008). Several human genes are transcribed from a promoter situated within the L1 element (Nigumann et al. 2002) and transcripts originating within Alus have also been reported (Faulkner et al. 2009). The evolutionary potential of TE-derived cis-regulatory sequences was recently demonstrated in rice, where recent TE insertions have led to upregulation of gene expression and the creation of new regulatory networks (Naito et al. 2009).

Other mechanisms may also contribute to the transcriptional impact of TEs, such as reduced elongation efficiency or premature polyadenylation following intronic L1 insertion (Han et al. 2004). Furthermore, mammalian TE activity is under epigenetic control, through short interfering RNAs (Yang and Kazazian 2006), histone modifications (Martens et al. 2005), and DNA methylation (Walsh et al. 1998). In Arabidopsis thaliana, a side effect of epigenetic silencing has been to reduce expression of neighboring cellular genes (Hollister and Gaut 2009).

With this in mind, it is tempting to ask how the evolution of human gene expression has been affected by TE activity. Expression divergence (ED) is a measure of the difference in gene expression levels between two species. Two previous studies have suggested a relationship between TE insertions and ED. Firstly, there is a correlation between the number of Alu insertions and ED as measured between human and mouse, although the direction of the correlation depends on the statistic used to measure ED.
(Urrutia et al. 2008). Secondly, a positive correlation between ED and the number of lineage-specific SINEs and long terminal repeat (LTR) elements has been found in rodents, where, although the amount of variance explained was modest, the average effect of TEs was considerable and appeared to have contributed around 20% of the total ED between mouse and rat (Pereira, Enard, and Eyre-Walker 2009).

Here, we investigate to what extent TE activity has contributed to hominid evolution by analyzing quantitative changes in gene expression and transcript diversity between human and chimpanzee.

Materials and Methods

We used two data sets to study the evolution of gene expression. In the first, microarray expression data for brain, heart, kidney, liver, and testis were available from six humans and five chimpanzees (Khaitovich et al. 2005). These experiments were conducted using the Affymetrix U133plus2 array, which was designed for human sequences, but contains a number of probes that match chimpanzee sequences equally well. This array has been shown to perform well in comparison with other arrays, including the newer exon arrays (Robinson and Speed 2007). The raw data were masked using the protocol developed by Toleno et al. (2009), in which probes were removed unless they had a perfect single match in both the human and the chimpanzee genome. Furthermore, only probe sets that contained at least six such probes were used for further analysis, as probe sets represented by fewer probes tend to give unreliable results (Toleno et al. 2009). Expression values were calculated using the robust multichip analysis (RMA) function in the BioConductor affy package (Irizarry, Bolstad, et al. 2003; Irizarry, Hobbs, et al. 2003; Gentleman et al. 2004). (Processed data were kindly provided by Joe Hacia of the University of Southern California.)

For each gene, we calculated ED between human and chimpanzee as the Euclidean distance between the average log-transformed expression values for each tissue. If a gene was assigned multiple probe sets, a single probe set was chosen at random to represent that gene, in order to avoid bias in the estimation of ED (see Results). Gene coordinates were downloaded from the UCSC Genome Bioinformatics site (http://genome.ucsc.edu), using genome build hg18 for human and panTro2 for chimpanzee. For genes with alternative transcripts, a single transcript was chosen at random among those that matched the probe set representing that gene.

To allow lineage-specific analysis of ED, we analyzed a second data set, which included data from rhesus macaque as an outgroup species. Somel et al. (2009) measured gene expression levels in the prefrontal cortex of 39 humans, 14 chimpanzees and 9 rhesus macaques, using the Affymetrix U133plus2 platform as in the first data set. The raw data were masked using files made available by the authors to include only probes that had a single perfect hit in the genomes of all three species and to require each gene to be represented by at least eight such probes. Log-transformed expression values were calculated using the RMA function in BioConductor (Irizarry, Bolstad, et al. 2003; Irizarry, Hobbs, et al. 2003; Gentleman et al. 2004). We calculated ED as the Euclidean distance between the average expression levels for the relevant species and normalized the values by dividing by the mean ED value for that species pair. To determine whether the individuals in the data set had reached puberty or not, we used life history data from the AnAge database (de Magalhaes and Costa 2009).

Recently inserted TEs in the human and chimpanzee genomes had previously been identified by Mills et al. (2006). We converted the data to current genome coordinates, using the UCSC liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Due to rearrangements in the updated genome assemblies, conversion failed for 7 human and 440 chimpanzee entries. These were excluded from the set. We then scored each of the genes for which we had expression data, according to the presence or absence of a recent TE insertion within the following seven regions: 0–2, 2–10, and 10–20 kb upstream and downstream of the transcript and within the introns.

To identify TEs present in both human and chimpanzee, but not in the rhesus macaque (genome release rheMac2), we used the human–chimpanzee and human–macaque net alignments displayed in the UCSC Genome Browser (http://genome.ucsc.edu). We identified all gaps in the human–macaque alignment that did not match a human–chimpanzee gap and then compared these with TEs in the RepeatMasker track. To allow for slight annotation errors, we isolated all RepeatMasker entries where the coordinates matched a gap in the rhesus macaque sequence, ±20 bp.

Expression state in the germ line was assigned according to eGenetics/SANBI EST data (Kelso et al. 2003), as incorporated in Ensembl release 56 (www.ensembl.org), by considering genes active if they were associated with the Cell Type term “germ cell”.

Results

We set out to investigate if recent TE insertions in the human or chimpanzee lineage have led to increased ED in nearby genes. Lineage-specific TEs had previously been identified (Mills et al. 2006) by identifying indels in the human–chimpanzee genome alignment and matching these to TEs in RepBase version 10.02 (Jurka 2000). Thus, the set of new TE insertions may also contain a small number of ancient TEs that were precisely deleted in one species. Genes were classified according to the presence of a recently inserted TE within 0–2 kb, 2–10 kb, and 10–20 kb either upstream or downstream of the transcribed sequence or within the introns. No exonic TEs were found.

Microarray expression data for both species were available for 8995 genes and five tissues (Khaitovich et al. 2005; Toleno et al. 2009). We calculated ED as the Euclidean distance between the log-transformed tissue-specific
Expression values for each species. We decided against another commonly used alternative definition of ED, based on the correlation coefficient, as it tends to overestimate ED for genes with conserved uniform expression (Pereira, Waxman, and Eyre-Walker 2009).

Calculations of ED were complicated by the fact that some genes were represented by more than one probe set in the microarray data. Although the platform used to generate the data was not designed to address alternative splicing, some probe sets have still been created to target different transcripts of the same gene. If different numbers of probe sets are used to generate the ED values and if the probability of retaining a TE is related to whether the affected gene undergoes alternative processing, this could introduce a bias into the analysis. Indeed, we found that human genes to which we had mapped at least one recently inserted TE had on average 2.7 annotated Ensembl transcripts, whereas genes without insertions had 2.3 transcripts (P = 2 × 10^{-16}, Mann–Whitney U test). The corresponding values for chimpanzee were an average of 2.0 transcripts for genes with TEs and 1.8 transcripts for genes without (P = 5 × 10^{-10}). To avoid bias in our estimates of ED, we therefore decided to let each gene be represented by a single probe set chosen at random.

To evaluate the effect of TE insertions on ED, we considered each TE family and each of our seven gene regions in turn. Although we found a marginally significant increase in median ED for genes with L1 insertions within 0–2 kb upstream and genes with SVA insertions within 0–2 kb downstream (Mann–Whitney U test, P = 0.030 and P = 0.032, respectively), these results are not significant after correcting for multiple tests. We therefore combined the data from each TE family (fig. 1). In spite of a general tendency towards an increase in median ED, none of the regions gave significant results when considered separately. However, if we combine these P values, using the Z transformation method (Whitlock 2005), the result is significant (P = 0.024), and even more so if we exclude the regions 10–20 kb upstream and downstream (P = 0.0027).

It therefore seems that genes with new TEs have higher ED. It is, however, not possible to infer the direction of causality based on these results, as they could be explained either by an increased ED as an effect of TE insertion or by a tendency for genes with higher ED to accumulate TEs. To test between these alternatives, we identified TE insertions that occurred before the human–chimpanzee split, but after the split from rhesus macaque, we reasoned that these fairly recent insertions should affect human and chimpanzee equally and therefore not contribute to ED between the two species. We found that genes with shared TE insertions did display a significantly higher level of ED, if we combined TEs within regions and probabilities as above (combined P value = 0.00003), indicating that TEs tend to integrate and/or be retained in genes that for some other reason are more likely to change their expression level (fig. 2).

Thus, at least part of the increase in ED for genes with species-specific TE insertions can be explained as a background effect, which also affects genes with shared TEs. Nevertheless, it is possible that TEs induce an additional increase in ED. To investigate this, we calculated the relative effect of TEs on ED as the ratio between the average ED values for genes with species-specific TEs and genes without such TEs, divided by the ratio between the average ED values for genes with shared TEs and genes without such TEs. If the relative effect is above one, it indicates that the presence of species-specific TEs acts to increase ED over and above the general tendency for TEs to integrate into genes with high ED. However, we find that the relative effect is not significantly above one for any of the seven regions under consideration (fig. 3). The highest relative effect is observed for genes with TEs within 0–2 kb downstream of the transcript, but the 95% confidence interval obtained by bootstrapping is (0.97, 1.48) for this single value, and thus, the result is not significant. As we cannot detect any increase in ED due to new TE insertions, beyond what can be explained by a general tendency for genes with higher ED to retain TEs, we conclude that TE activity has not contributed to the genome-wide evolution of gene expression levels in humans and chimpanzees.

Although we find no evidence that new TE insertions increase ED in the analysis above, it is possible that this is due to a lack of power. We therefore sought to test whether TEs affect ED using a complementary approach.
For genes with a new TE insertion in humans, we compared the ED between human and macaque with the ED between chimpanzee and macaque. We also performed the corresponding analysis for genes with a TE in chimpanzees. If TEs affect ED, we predict that genes with a human-specific TE insertion will show higher ED between human and macaque than between chimpanzee and macaque, with the converse being the case for genes with a chimpanzee-specific insertion. To perform the analysis, we only considered genes that had one or more insertions in one species, but none in the other. We analyzed microarray data for 3747 genes in the prefrontal cortex of 39 humans, 14 chimpanzees, and 9 rhesus macaques (Somel et al. 2009). The presence of an outgroup in this data set allowed us to assess changes in ED on a lineage-specific basis. To do so, we calculated human–macaque and chimpanzee–macaque ED as the Euclidean distance between the means of the log-transformed expression values for each species. Because human and chimpanzee share a common history, these ED values represent the sum of a species-specific component as well as a shared component that accounts for all ED between rhesus macaque and the human–chimpanzee ancestor. Any difference between the human–macaque and chimpanzee–macaque ED values can therefore be directly attributed to human-specific or chimpanzee-specific events.

On average, chimpanzee–macaque ED is higher than human–macaque ED in this data set. Consequently, if we test for an increase in ED for the lineage with TE insertions, the test would be too conservative for human-specific TEs and too liberal for chimpanzee-specific TEs. To allow for an unbiased test, we normalized all ED values by dividing the ED for each gene by the mean ED for that species pair. Note the fact that ED between chimpanzee and macaque is higher than that between human and macaque does not necessarily imply accelerated evolution along the chimpanzee lineage. Rather, it might be best explained by the higher variance among chimpanzee individuals in this data set, especially considering previous work indicating that ED in the brain is higher along the human lineage (Khaitovich et al. 2005).

Consistent with our previous analysis, we find no evidence, in any of the regions examined, that a lineage-specific TE tends to increase ED in that species relative to ED in the other species (fig. 4). This is true even if we combine probabilities across introns and flanking regions ($P = 0.32$ for human-specific TEs and $P = 0.13$ for chimpanzee-specific TEs; Mann–Whitney U test and Z transformation). Because the samples used to generate the expression data were taken from individuals of varying ages (Somel et al. 2009), we repeated the analysis separately for samples from prepubertal and postpubertal individuals, in order to reduce age-related variation. Again, the results were not significant (combined probabilities, prepubertal individuals: $P = 0.42$ for human TEs and $P = 0.17$ for chimpanzee TEs and postpubertal individuals: $P = 0.84$ for human TEs and $P = 0.13$ for chimpanzee TEs), providing further support for the hypothesis that recent TE insertions have not acted to increase ED between humans and chimpanzees.

During our analysis of species-specific TEs, we observed that upstream insertions were more frequent than downstream insertions. In total, we identified 561 genes with at least one new TE within 20 kb upstream of the transcription start site in either human or chimpanzee and 496 genes with at least one new TE downstream of the transcribed region. The difference is just significant ($P = 0.049$, two-tailed binomial test), and upstream insertions are also more common if we only consider TEs within...
10 or 2 kb upstream or downstream of genes, although the overrepresentation is not significant ($P = 0.075$ and $P = 0.47$, respectively). This enrichment of upstream insertions is surprising because we might expect that TEs inserted upstream would be more likely to disrupt transcriptional regulatory elements and therefore tend to be selected against, although it has previously been noted that TE insertions in the 3' flanking region of rodent genes tend to show bigger effects on ED than those in the 5' region (Pereira, Enard, and Eyre-Walker 2009). Another explanation is that TEs are preferentially inserted upstream of genes, as is the case for P elements in Drosophila melanogaster (Spradling et al. 1995), where it is presumed to be linked to the altered chromatin structure around the transcription start site of active genes (Kelley et al. 1987; Voelker et al. 1990). If the same were true for hominin TEs, then we would expect an enrichment of upstream TE insertions for genes that are expressed in the germ line, but not for other genes. Based on expression data downloaded from Ensembl (see Materials and Methods), we categorized all genes as active or inactive in the germ line and compared the number of upstream and downstream insertions for active and inactive genes. When we considered all recent TE insertions together, we found that the inactive genes have approximately the same number of upstream and downstream insertions, whereas active genes have significantly more upstream insertions ($P = 0.003$, $\chi^2$ test). The pattern is contributed mainly by Alu and, to some extent, SVA elements, whereas L1 elements appear unaffected (table 1).

Although species-specific TEs have not affected ED between human and chimpanzee, they may still have had an influence on other aspects of gene expression evolution, such as transcript diversity. As described above, we established that genes with recent TE insertions have a significantly higher number of annotated transcripts than genes without such insertions. Because both Alu and L1 elements can be involved in processes such as alternative promoter usage (Nigumann et al. 2002; Faulkner et al. 2009) and alternative splicing (Makalowski et al. 1994; Sorek et al. 2002; Belancio et al. 2006; Lev-Maor et al. 2008), which act to increase transcript diversity, we speculated that the TE insertions themselves might in part explain why the affected genes tended to produce more transcripts. The differences in annotation quality between the human and the chimpanzee transcriptomes make a direct comparison of transcript numbers difficult. Instead, we reasoned that if TEs increase transcript diversity, then human genes should have more transcripts on average if they contained a human-specific TE, than if their chimpanzee ortholog contained a chimpanzee-specific TE. Conversely, we would expect chimpanzee genes with chimpanzee-specific TEs

**Table 1.** Number of Recent Upstream and Downstream TE Insertions in Genes that Are Active or Inactive in the Germ Line.

<table>
<thead>
<tr>
<th>TE Type</th>
<th>Active Upstream</th>
<th>Active Downstream</th>
<th>Inactive Upstream</th>
<th>Inactive Downstream</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>255</td>
<td>181</td>
<td>306</td>
<td>315</td>
<td>0.003</td>
</tr>
<tr>
<td>Alu</td>
<td>169</td>
<td>129</td>
<td>206</td>
<td>215</td>
<td>0.04</td>
</tr>
<tr>
<td>L1</td>
<td>19</td>
<td>19</td>
<td>38</td>
<td>32</td>
<td>Not significant</td>
</tr>
<tr>
<td>SVA</td>
<td>51</td>
<td>33</td>
<td>53</td>
<td>55</td>
<td>Not significant</td>
</tr>
</tbody>
</table>
to produce more transcripts than chimpanzee genes where the human equivalent had undergone TE insertion.

We calculated the number of transcripts in the Ensembl database (release 54) for human and chimpanzee genes that contained recently inserted Alu, L1, or SVA insertions (fig. 5). Before correction for multiple tests, there was only one significant result; human genes with a new SVA insertion have significantly more transcripts than human genes with a new SVA insertion in chimpanzees. However, this result is not significant after correction for multiple tests, and we do not see a similar pattern for chimpanzee genes. Of course, it should be noted that the lack of observed effect of TEs on transcript diversity could be due to insufficient annotation of alternative isoforms.

**Discussion**

TEs have previously been proposed as important contributors to the evolution of gene regulation (Britten and Davidson 1971; Feschotte 2008). In contrast to this, our results show that recent TE activity has not had a detectable effect on ED between human and chimpanzee, suggesting that although TEs may contribute occasionally to gene ED in hominids, they are not a major source of regulatory change.

Our results are consistent with those of Urrutia et al. (2008) but are surprising considering previous results in mouse and rat, in which it was estimated that 20% of all ED was due to the insertion of new SINE and LTR elements (Pereira, Enard, and Eyre-Walker 2009). The discrepancy between hominids and rodents might be due to qualitative differences in TE activity in the two groups. In rodents, the TEs with strongest apparent influence on ED were LTRs and SINEs (Pereira, Enard, and Eyre-Walker 2009); however, new LTR insertions are rare in the human and chimpanzee genomes, and SINEs, although common, are represented mainly by the primate-specific Alu element (Mills et al. 2006). Pereira, Enard, and Eyre-Walker (2009) also attempted to establish causality between ED and new TE insertions by considering the correlation between ED and TE insertions shared by mouse and rat, but these shared TEs were potentially much older than those we have used here and may therefore have been an imperfect control if the pattern of TE insertion had changed over time.

The results presented here are not consistent with a model where TEs affect gene expression by disrupting existing sequences or providing “ready-to-use” regulatory elements. In particular, we find no indications that intronic L1 insertions affect ED, as might have been expected considering that in vitro assays have shown that such insertions can attenuate reporter gene expression by reducing elongation efficiency (Han et al. 2004). On the other hand, although a few candidate cases exist (Schwahn et al. 1998; Yajima et al. 1999), it has yet to be shown that this form of regulation is used in vivo (Han and Boeke 2005).

It has been argued that TEs initially may only show a weak impact on gene expression and that this regulatory function is subsequently refined by selection (Faulkner and Carninci 2009). Possibly, the short time scale of this study might therefore not allow us to gauge the full impact of TEs on gene expression. However, at least for Alu elements, recent insertions do not appear to be under selection (Cordaux et al. 2006). This is not to say that decaying TEs may not provide sequence material in which functional elements can later evolve. There are several examples of human enhancers that have arisen in this way (Britten 1994; Ackerman et al. 2002; Medstrand et al. 2005). Nevertheless, the presence of TE-derived regulatory sequences might best be explained by the abundance of TEs in the genome. Considering that 45% of the human genome has been contributed by TEs (Lander et al. 2001), it stands to reason that these sequences would harbor a fair share of regulatory modules.

It should also be appreciated that while we find no evidence for TEs contributing to differences in gene expression between hominid species, it is still possible that they contribute to variation within a single species. For example, it may be that TEs in general cause mutations of large effect, which rarely are beneficial or neutral and therefore never become fixed between species. Such large effect mutations can contribute substantially to variation in fitness and phenotypes, even if they are very deleterious (Eyre-Walker 2010).

In a recent study, it was shown that human genes are more likely to be expressed at high levels and in broad patterns if their promoters are rich in TEs, which might indicate that TEs are used to modify chromatin structure upstream of the transcription start site (Huda et al. 2009). Our results, showing that TEs preferentially insert...
upstream of genes that are transcribed in the germ line, suggest insertion bias as a possible alternative explanation of these results. The same process might also have contributed to the overall enrichment of SINEs in upstream sequences previously observed by Medstrand et al. (2005). Interestingly, it seems that it is primarily Alu elements and, to some extent, SVA elements that experience insertion bias, whereas L1 elements appear to be unaffected. This is surprising, considering that Alus and SVAs are nonautonomous elements that do not encode proteins necessary for transposition, but instead parasitize the L1 machinery (Dewannieux et al. 2003; Ostertag et al. 2003). Although there are some mechanistic differences between Alu and L1 insertions (Kroutter et al. 2009), it is unclear how this might contribute to the observed bias.

The distribution of TEs in the human genome is nonrandom and correlates with various aspects of gene expression, such as expression levels, transcript diversity, and activity in the germ line. Importantly, as illustrated in this study, a correlation does not necessarily imply causality. When studying the contributions of TEs to gene expression evolution, it is therefore crucial to apply proper controls in order to disentangle any real effects from the background.

Acknowledgments

We are very grateful to Joe Hacia for providing data, to John Brookfield for helpful discussion, and to several reviewers for constructive criticism. The work was funded by the Biotechnology and Biological Sciences Research Council and the University of Sussex.

References


