Effect of the Transposable Element Environment of Human Genes on Gene Length and Expression

Daudi Jjingo[1], Ahsan Huda[1], Madhumati Gundapuneni[1,2], Leonardo Mariño-Ramírez[3,4], and I. King Jordan*[,1,4]

[1]School of Biology, Georgia Institute of Technology

[2]Institute for Systems Biology, Seattle, Washington

[3]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

[4]PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia

*Corresponding author: E-mail: king.jordan@biology.gatech.edu.

Accepted: 22 February 2011

Abstract

Independent lines of investigation have documented effects of both transposable elements (TEs) and gene length (GL) on gene expression. However, TE gene fractions are highly correlated with GL, suggesting that they cannot be considered independently. We evaluated the TE environment of human genes and GL jointly in an attempt to tease apart their relative effects. TE gene fractions and GL were compared with the overall level of gene expression and the breadth of expression across tissues. GL is strongly correlated with overall expression level but weakly correlated with the breadth of expression, confirming the selection hypothesis that attributes the compactness of highly expressed genes to selection for economy of transcription. However, TE gene fractions overall, and for the L1 family in particular, show stronger anticorrelations with expression level than GL, indicating that GL may not be the most important target of selection for transcriptional economy. These results suggest a specific mechanism, removal of TEs, by which highly expressed genes are selectively tuned for efficiency. MIR elements are the only family of TEs with gene fractions that show a positive correlation with tissue-specific expression, suggesting that they may provide regulatory sequences that help to control human gene expression. Consistent with this notion, MIR fractions are relatively enriched close to transcription start sites and associated with coexpression in specific sets of related tissues. Our results confirm the overall relevance of the TE environment to gene expression and point to distinct mechanisms by which different TE families may contribute to gene regulation.

Key words: gene expression, gene regulation, selection hypothesis, genomic design hypothesis, L1, MIR.

Introduction

The relationship between gene architecture and gene expression has been and remains a subject of continuing interest for genome analysis. In a pioneering study, Castillo-Davis et al. (2002) observed that, for human and worm genes, intron length was negatively correlated with the level of expression. In other words, shorter genes were found to be expressed at higher levels and longer genes at lower levels. To explain this trend, the authors formulated the "selection hypothesis" (Castillo-Davis et al. 2002). This hypothesis posits that highly expressed genes are shorter due to selective forces that operate in favor of minimizing the energy and time expended during transcription. Subsequently, the relationship between gene length (GL) and expression level was confirmed by a number of studies, providing support for the selection hypothesis (Eisenberg and Levanon 2003; Urrutia and Hurst 2003; Comeron 2004; Chen et al. 2005; Seoighe et al. 2005; Li et al. 2007).

In 2004, Vinogradov (2004) also observed that compact genes were more highly expressed, but he offered a different explanation for this trend. Vinogradov proposed the "genomic design" hypothesis, which postulates that the shorter length of highly expressed genes is better explained by the fact that these genes also tend to be broadly expressed across numerous tissues and thus have simpler regulation, and require fewer regulatory sequence elements, than

genes expressed in a more narrow tissue-specific fashion. In other words, the relative paucity of regulatory elements in broadly expressed genes explains their shorter average length. The genomic design hypothesis rests on the notion that the apparent correlation between GL and the level of expression actually reflects a relationship between GL and the breadth of expression, that is, the number of tissues in which a gene is expressed.

The selection hypothesis and the genomic design hypothesis make distinct testable predictions regarding the relationship between GL and gene expression. The selection hypothesis predicts the strongest correlation between GL and the overall expression level, whereas the genomic design hypothesis predicts the strongest correlation between GL and the breadth of expression. A recent study used these predictions to evaluate the competing hypotheses and found that the selection hypothesis serves as the best explanation for the relationship between GL and expression (Carmel and Koonin 2009).

While the aforementioned studies were ongoing, there was an independent line of research investigating the relationship between gene architecture and gene expression from a different perspective. In eukaryotic genomes, and particularly for mammalian genomes, gene architecture is substantially influenced by the presence of transposable element (TE)–derived sequences. TE-derived sequences are extremely abundant in mammalian genomes; at least 45% of the human genome is made up of TE sequences (Lander et al. 2001; Venter et al. 2001). In addition, TE sequences are nonrandomly distributed across genomes. In the human genome, Alu (SINE) elements are enriched in GC- and gene-rich regions, whereas L1 (LINE) elements are enriched in low-GC and gene-poor regions (Smit 1999; Lander et al. 2001). Finally, individual genes can vary tremendously with respect to the amount and identity of TE sequences that they harbor.

Over the last several years, a series of studies have called attention to a relationship between the TE environment in and around genes and the level and breadth of gene expression. In 2003, the human genome sequence was used together with expression data to construct a human transcriptome map (Versteeg et al. 2003). This map identified colocated clusters of highly expressed genes with specific genomic characteristics. These clusters were gene dense, had high GC content, were enriched for SINEs, Alu elements in particular, and had low LINE densities. The same study found clusters of weakly expressed genes with low SINE and high LINE densities. Shortly thereafter, Han et al. (2004) confirmed that the most highly expressed human genes were depleted for L1 elements and demonstrated a mechanism that could partially explain this pattern. They showed that L1 elements can disrupt transcriptional elongation based on the presence of strong polyA signals in their sequences.

Kim et al. made an important contribution to this body of work by distinguishing between TE effects on the level of expression and the breadth of expression (Kim et al. 2004). They measured overall expression level as the peak expression (PE) over all tissues and breadth of expression (BE) as the number of tissues in which a gene is expressed over some basal threshold. Their work revealed that Alu element gene densities are more highly correlated with BE, whereas L1 densities are most negatively correlated with PE. These results suggested that different families of TEs may have specific effects on different aspects of gene expression. Consistent with these results, Eller et al. showed that highly and broadly expressed housekeeping genes can be distinguished by their TE content, being primarily enriched for Alus and depleted for L1s (Eller et al. 2007). In addition to the level and breadth of expression, the TE environment of mammalian genes has also been related to expression in cancer tissues (Lerat and Semon 2007) and the evolutionary divergence of gene expression (Pereira et al. 2009).

As of yet, no one has attempted to consider these two areas of investigation together: 1) the relationship between GL and expression and 2) the relationship between TE environment and gene expression. In this study, we attempt to disentangle the effects of GL and TE environment on gene expression and to evaluate the relative influences of each on expression. Having considered their effects separately, we then more thoroughly evaluate the connections between gene architecture and the selection versus genomic design hypotheses.

Materials and Methods

Defining Gene Loci

To accommodate alternative splice variants of human genes and compute TE fractions for specific loci, we define genes here as distinct transcriptional units (TUs)—genomic regions encompassing all overlapping transcripts from the start of the 5′-most exon to the end of the 3′-most exon (supplementary fig. S1A, Supplementary Material online). To that end, we downloaded RefSeq annotations for the March 2006 build of the human genome reference sequence (National Center for Biotechnology Information [NCBI] build 36.1; University of California–Santa Cruz [UCSC] hg18) from the UCSC Genome Browser (Karolchik et al. 2004; Rhead et al. 2010). A total of 32,128 RefSeq transcripts were merged into 19,123 TUs that represent distinct gene loci.

Determining Genic and Intergenic TE Fractions

To determine the fractions of human genes (TUs) that are made up of TE sequences, human TEs were broken down into six of the major human TE classes or families according to the Repbase classification system (Jurka et al. 2005; Kohany et al. 2006)—Alu, MIR, L1, L2, DNA and LTR (long terminal repeat). RepeatMasker (http://

www.repeatmasker.org) annotations of the genomic coordinates of these TEs were used to map them onto their colocated genes. For each TE type, its fraction in a gene was computed as the number of base pairs occupied by a TE as a fraction of all base pairs in the gene. For each human gene, its intergenic region was taken as the union of the regions upstream of the transcription start site (TSS) and downstream of the termination site to the genomic midpoint between the adjacent upstream and downstream genes. TE intergenic fractions were then calculated in the same way as for TE genic fractions based on these genomic coordinates.

Gene Expression Data

To measure gene expression in different tissues, we used the Gene Expression Atlas from the Genomics Institute of the Novartis Research Foundation, which consists of Affymetrix microarray gene expression values for 44,776 probe sets across 79 human tissues (Su et al. 2004). Affymetrix probe sets were mapped onto their corresponding TUs based on their genomic location coordinates. As suggested previously (Stalteri and Harrison 2007), probes that mapped to more than one TU were discarded, and for TUs with more than one mapped probe, the average expression level per tissue was used. This resulted into a final data set of 15,658 TUs to which expression data could be assigned. Expression data are represented as signal intensity units based on the Affymetrix MAS4 processing and normalization algorithm suite.

Measurement of GL and Gene Expression Parameters

For each TU, the GL was calculated by simply subtracting its start coordinate along the chromosome from the end coordinate and then subjecting the difference to a log2 transformation. The microarray expression data described above were used to calculate three measurements of gene expression: peak expression (PE), breadth of expression (BE) and tissue-specificity (TS). To obtain PE, the signal intensity value from the tissue where the TU is most highly expressed was selected for each TU and subjected to a log2 transformation to accommodate the vast disparity (range = 197,652.4 signal intensity units) in the peak levels of expression between TUs. For each TU, the BE was calculated as the number of tissues in which the expression of the TU exceeded a threshold of 350 expression signal intensity units (Jordan et al. 2005). For each TU, a TS index was computed as described (Yanai et al. 2005). The value of TS varies between 0 and 1 and reflects the number of tissues where the TU is overly expressed relative to its expression in other tissues. The TS index is calculated as follows:

$$\mathrm{TS} = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1},$$

where N is the number of tissues and x_i represents a TU's signal intensity value in each tissue i divided by the maximum signal intensity value of the TU across all tissues.

Comparative Analysis of GL, TE Gene Fractions, and Gene Expression Parameters

The relative effects of GL and the TE gene environment on gene expression were evaluated using pairwise and multiple linear regression analyses where GL and the TE fractions were the independent variables and the gene expression parameters PE, BE, and TS were the dependent variables. For these analyses, parameter values were ranked and binned in order to smooth the signal and reduce the background noise. For each parameter, the 15,658 TUs were ranked and divided into 100 bins of approximately equal size (~157 TUs per bin). Parameter values were averaged for each bin and the averages were used to populate ordered vectors of values (n = 100). Vectors that represent independent and dependent variables were then compared using pairwise regression or combined into a multiple regression model. All data were treated using the same ranking and binning procedure so that the relative effects of the independent variables on the dependent variables could be comparatively evaluated.

Gene Expression Clustering Analysis

TS patterns for the top 10% MIR-rich genes were analyzed using hierarchical clustering based on pairwise Euclidean distances between vectors of tissue-specific gene expression levels over 79 tissues. This analysis was conducted using the program Genesis (Sturn et al. 2002) with signal intensity values median normalized across tissues.

Statistical Analyses Used

For the pairwise regression analyses, independent and dependent variable vectors were compared using pairwise Pearson correlation (r values in figs. 1–5; individual coefficient of determination R^2 values in tables 1–5), and the significance of the correlations (P values in figs. 1–5 and tables 1–5) was determined using the Student's t-distribution. Partial correlation analyses were used to control for the effects of correlated pairs of independent variables (tables 1, 2 and 4). Multiple regression analyses were conducted to determine the combined coefficient of determination for all TE fractions (R^2 values in table 3) and the partial correlation values (r values in table 3). Significance values for the multiple coefficients of determination ("all TE" P values in table 3) were determined using the F distribution. Significance values for the partial correlations (P values in tables 1–4) were determined using the Student's t-distribution.

Results and Discussion

TE Environment of Human Genes

Gene and TE annotations from the reference sequence of the human genome (NCBI build 36.1; UCSC hg18) were analyzed together to characterize the TE environment of human genes. A total of 19,123 TUs, which reconcile alternative splice variants and represent discrete gene loci, were derived from RefSeq annotations as described in the Materials and Methods (see also supplementary fig. S1A, Supplementary Material online). The fraction of each human gene locus derived from TE sequences was determined using RepeatMasker annotations. Six of the most abundant classes (families) of TEs were considered in this analysis—Alu, MIR, L1, L2, DNA and LTR. The frequencies of other classes of TEs were found to be too low to substantially affect the overall TE environment of human genes.

Human genes show an average TE fraction of 34% and a standard deviation (SD) of 18% (fig. 1*A*). Human TE gene fractions show a broad distribution that is fairly bell shaped with the exception of a sharp peak of genes that are devoid of TEs (0% TE fraction in fig. 1*A*). The presence of these TE-free genes is consistent with the removal of genic TEs by purifying selection (Simons et al. 2006). The TE gene fractions observed for individual TE families are consistent with previous results (Medstrand et al. 2002) in which Alu elements were found to be the most abundant family of TEs in human genes, whereas LTR elements are found in the lowest frequency within human genes (supplementary fig. S1B, Supplementary Material online). The length distributions of TEs in genes (supplementary table S2, Supplementary Material online) reveal that they are mostly short (<400 bp) as would be expected in transcribed regions where long TEs are less tolerated owing to their higher propensity to be deleterious.

Overall, intergenic regions show higher TE fractions (average = 46%; fig. 1*A*) and also have a more normal distribution with lower variation than seen for genic regions (SD = 14%; fig. 1*A*). For individual human genes, genic and intergenic TE fractions are highly positively correlated ($r = 0.95$, $P = 6.3 \times 10^{-53}$; fig. 1*B*), consistent with the notion that the local genomic environment strongly influences TE gene fractions (Smit 1999; Lander et al. 2001).

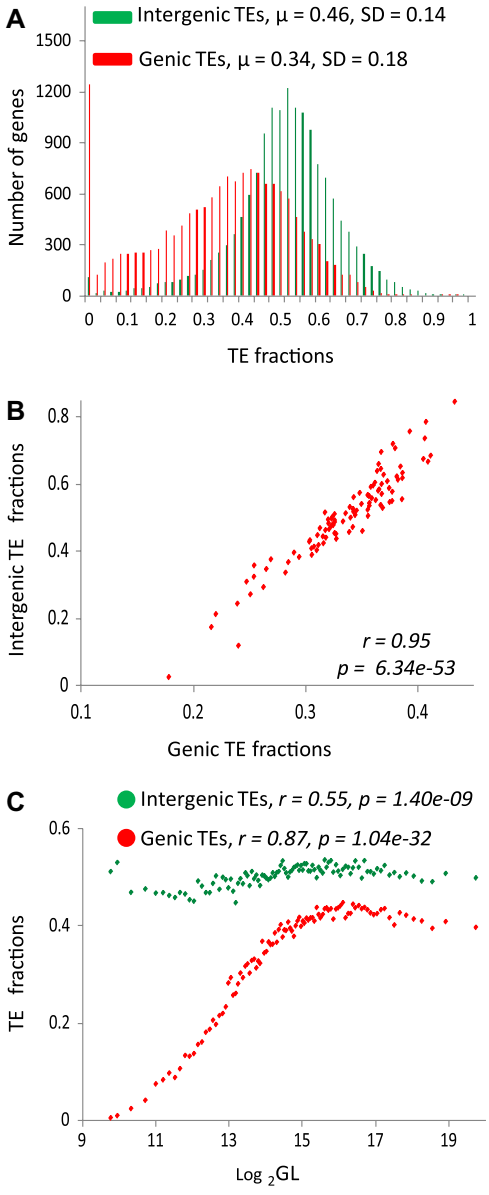

Fig. 1.—TE fractions in and around human genes. (*A*) Distributions of intergenic (green) and genic (red) TE fractions. (*B*) Relationship between intergenic TE fractions and the corresponding genic TE fractions. (*C*) Relationship between intergenic TE fractions and GL (green) and relationship between genic TE fractions and GL (red). Pearson correlation coefficient values (*r*) along with their significance values (*P*) are shown for all pairwise regressions.

TE Fractions are Related to GL

As noted in the introduction, the relationship between GL and expression has been investigated separately from the relationship between the TE environment of genes and their expression. However, GL and gene TE fractions may be related if genes increase in length due, at least in part, to an accumulation of TE-derived sequences. If genes increase in length due to the acquisition of TEs, then we expect to see a positive correlation between gene TE fractions and GL. On the other hand, if GL increases via mechanisms that do not involve TEs, there should be no correlation between gene TE fractions and GL. To distinguish between these two possibilities, we compared the TE fractions of human genes with their length (as described in Materials and Methods).

When all human TEs are considered together, there is a strong and significantly positive correlation between gene TE fractions and GL ($r = 0.87$, $P = 1.0 \times 10^{-32}$; fig. 1*C*).

Table 1
Relationship between the Local TE Environment and GL

	TE Fractions	*r*	*P* Value
GL	Genic TE[a]	0.87	1.04E-32
	Intergenic TE[a]	0.55	1.40E-09
	Genic TE | Intergenic TE[b]	0.82	6.80E-45
	Intergenic TE | Genic TE[c]	−0.18	7.02E-02

[a] TE fractions within genes (genic) and between genes (intergenic) are correlated with GL.
[b] Partial correlation between genic TE fractions and GL controlling for intergenic TE fractions.
[c] Partial correlation between intergenic TE fractions and GL controlling for genic TE fractions.

Although only 0.55% of the average GL for the bin with the 1% shortest genes is constituted by TEs, the percentage progressively increases to 39.73% for the bin with the top 1% longest genes, a >72-fold increase in the average fractions of genes occupied by TEs. However, the positive relationship between gene TE fractions and GL is not strictly monotonic. Specifically, in 77% of all genes, the percentage of GL constituted by TEs progressively increases from 0.55% in genes of about 850 bp to 44.79% for genes spanning about 70.9 kb (>81-fold increase in gene TE fraction; fig. 1*C*). For the remaining genes beyond this length (23% of all genes), the percentage of GL constituted by TEs levels off and remains more or less constant with increasing length.

As noted in the previous section, TE genic and intergenic fractions are highly correlated (fig. 1*B*). These data are consistent with previous studies showing that TE fractions and family distributions differ among genomic compartments and thus may depend on regional factors such as GC content and recombination rate (Medstrand et al. 2002; Versteeg et al. 2003). Therefore, it is possible that the relationship between genic TE fractions and GL simply reflects such regional genomic features. To test for this possibility, we compared intergenic TE fractions with GL. Intergenic TE fractions are significantly positively correlated with GL ($r = 0.55$, $P = 1.4 \times 10^{-9}$); however, the correlation is substantially weaker than seen for genic TE fractions and the slope of the relationship is far more flat (fig. 1*C*). Furthermore, partial correlation analysis shows that TE genic fractions remain positively correlated with GL when intergenic TE fractions are controlled for, whereas the positive correlation between intergenic TE fractions and GL disappears when genic TE fractions are controlled for (table 1). In other words, the relationship between TE gene fractions and GL does appear to have some gene-specific, as opposed to genomic regional, component.

To evaluate the correlation between TE genic fractions and GL more closely, we focused on individual TE families and found that Alus dominate the leveling off in gene TE fractions seen for the longest genes. Alus are the most abundant TE sequence within gene boundaries (supplementary fig. S1B, Supplementary Material online), and Alus also show a unique TE fraction distribution with GL. The fraction of Alus within genes rises sharply and peaks for midsize genes (~23.3 kb) followed by an almost equally precipitous decline in frequency, yielding a bell-shaped distribution (fig. 2*A* and supplementary fig. S3A, Supplementary Material online). However, the distribution of TE gene fractions for all other TE families analyzed tends to be generally linear in relation to GL (fig. 2*B*; supplementary fig. S3B–F, Supplementary Material online), increasing from an average percentage of 0.34% in the shortest genes to 32.83% in the longest genes (a >96-fold increase in the fractions of genes occupied by TEs).

It is not immediately apparent while Alu fractions, unique among all classes of TEs considered here, decline for the longest genes. One possibility is that Alus are known to be prevalent in GC-rich regions, whereas larger genes (introns) tend to have lower GC content (fig. 2*C*). Thus, it may be that the decline in Alu content for longer genes is based on regional genomic biases in GC content. If this is the case, then genes with low GC content should also have low Alu fractions and vice versa. We found that genes with low GC content do in fact have lower Alu content as expected (fig. 2*D*). However, the relationship between genic Alu fractions and GC content is not monotonic; Alu fractions peak for genes in the middle of the GC content range and decrease for both low– and high–GC content genes. We performed partial correlation in an attempt to further tease apart the relationship between Alu gene fractions and GC content as they relate to GL. GC content is much more strongly correlated with GL than Alu fractions are (fig. 2*A* and *C*). If the relationship of Alu genic fractions with GL mainly reflects regional changes in GC content, then the correlation of Alu fractions with GL should decrease when GC content is controlled for. However, when GC content is controlled for with partial correlation, the positive correlation between Alu gene fractions and GL actually increases (table 2). Similarly, when Alu gene fractions are controlled for, the correlation between GC content and GL becomes more negative. These data suggest that both Alu gene fractions and GC content are independently related, to some extent, with GL in the human genome.

Overall, the positive correlations between TE gene fractions and GL indicate that longer genes have disproportionately more TEs relative to other sequence elements. Considering all TE families together, TEs make up only 0.55% of the shortest genes and yet account for ~40% of the increase in GL when assessed in the longest genes. For three-fourth of all genes, the contribution of TEs to increases in GL is >45%. These results underscore the contributions of TEs to the length differences among human genes and suggest that the influences of TE environment and GL on gene expression cannot be adequately considered separately.

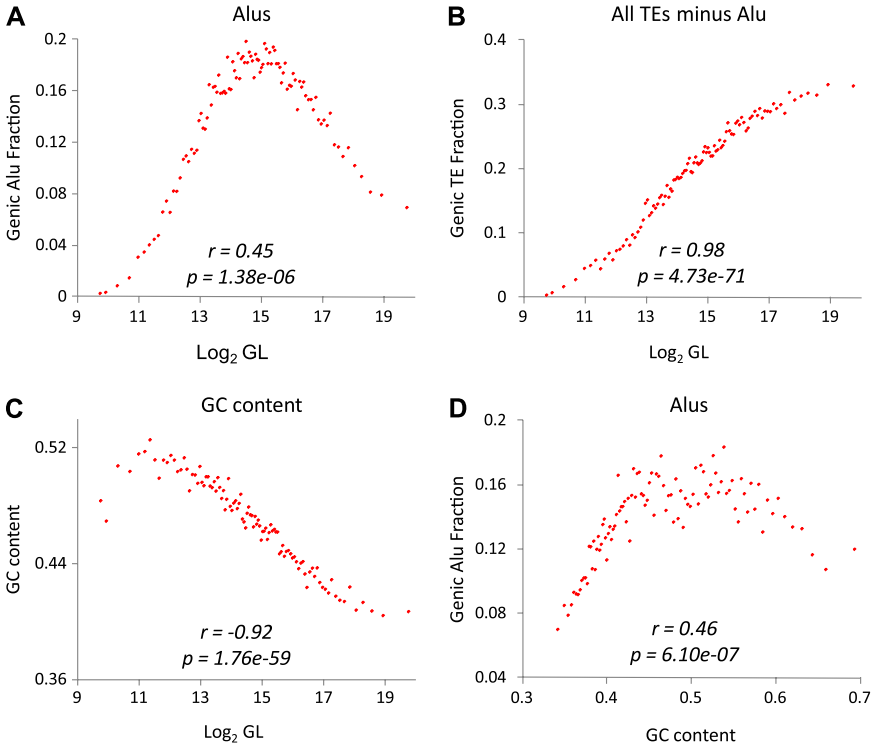

Fig. 2.—Relationships between the Alu fractions of human genes, GL, and GC content. (*A*) Relationships between Alu gene fractions and GL. (*B*) Relationship between TE gene fractions for all TEs except Alu and GL. (*C*) Relationship between GC content and GL. (*D*) Relationships between Alu gene fraction and GC content. Pearson correlation coefficient values (*r*) along with their significance values (*P*) are shown for all pairwise regressions.

TE Gene Environment and the Selection Hypothesis

In order to relate the TE environment of human genes and GL to gene expression, three expression parameters for human genes were measured using microarray data over 79 tissues as described in the Materials and Methods: 1) peak expression (PE), 2) breadth of expression (BE) and 3) TS. PE is the maximum expression level observed for a gene over all 79 tissues and is taken to represent the overall gene expression level; BE is the number of tissues in which a gene can be considered to be expressed, and TS is a measure of tissue specificity described previously (Yanai et al. 2005). PE and BE were measured here because they can be used to distinguish between the selection versus genomic design hypotheses. The selection hypothesis predicts a stronger positive correlation of PE with GL, whereas the genomic design hypothesis predicts a stronger correlation of BE with GL. However, BE has been criticized as an overly simplistic measure that may not distinguish genes that are expressed in the same sets of tissues albeit at very different relative levels. For this reason, we also use a measure of TS that explicitly reflects the number of tissues where a gene is overly expressed relative to its expression in other tissues (see Materials and Methods). Genes overly expressed in a few tissues (i.e., tissue-specific genes) have high TS indices, whereas more broadly and evenly expressed genes have low values of TS.

Regression analysis was used to individually compare values of these expression parameters with TE gene fractions for all six families and GL (figs. 3–5), and the effects of TE gene fractions and GL were also considered jointly using multiple regression (table 3). Consistent with previous results (Eisenberg and Levanon 2003; Carmel and Koonin 2009), GL can be seen to have a much stronger association with PE than BE. Whereas 48% of the variability in PE is attributable to GL, only about 4% of the variability in BE is attributable to GL (table 3). Furthermore, it can be seen that the nonmonotonic shape of the relationship between GL and PE (fig. 3*H*) is similar to what has been reported previously (Carmel and Koonin 2009) and also closely resembles the shape of the Alu gene fraction versus PE

Table 2
Effect of GC Content on the Relationship between Alu Genic Fractions and GL

	Feature[a]	*r*	*P* Value	Control[b]	*r*	*P* Value
GL	Alu	0.45	1.32E-06	Alu | GC	0.58	1.69E-12
	GC	−0.92	5.93E-42	GC | Alu	−0.94	2.99E-152

[a] Alu genic fractions and genic GC content values are correlated with GL.
[b] Partial correlation analyses control for effect of GC content on Alu fractions (Alu | GC) and Alu fractions on GC content (GC | Alu), respectively.

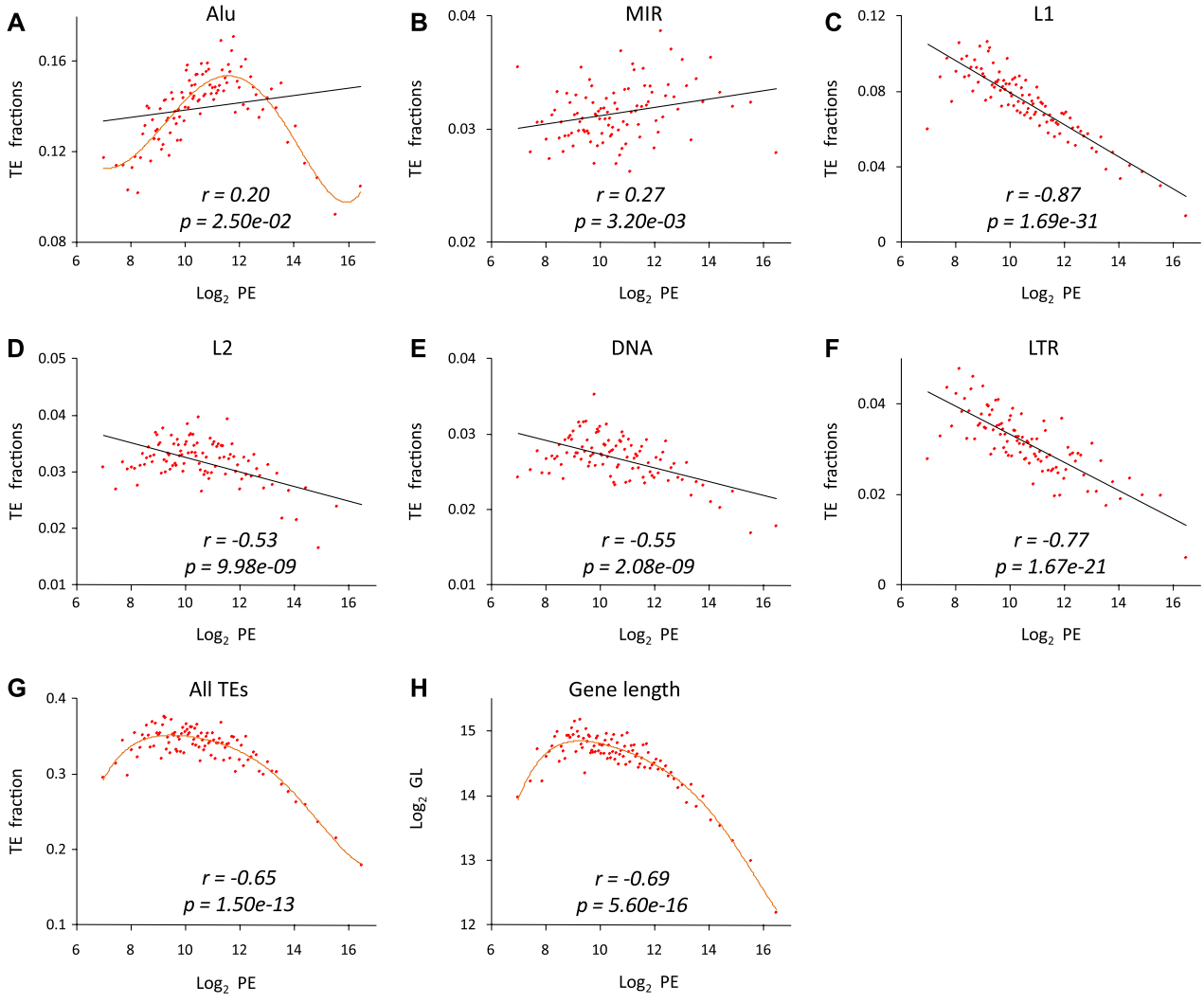

Fig. 3.—TE fractions, GL, and the peak expression (PE). Relationships between the TE gene fractions for (*A*) Alu, (*B*) MIR, (*C*) L1, (*D*) L2, (*E*) DNA, (*F*) LTR, and (*G*) all TEs and the PE of human genes. (*H*) Relationship between GL and PE. Pearson correlation coefficient values (*r*) along with their significance values (*P*) are shown for all pairwise regressions.

distribution (fig. 3*A*). The strongest individual TE family correlation with PE is the negative correlation seen for L1 fraction versus PE (fig. 3*C*). L1 also has the largest negative partial correlation value with PE in the multiple regression analysis as well as the largest coefficient of determination (table 3). When all TEs are analyzed together, 78% of the variability in PE can be attributed to variability in TE gene fractions, whereas only 48% is attributable to variability in GL (table 3).

Although these data do lend support to the selection hypothesis, they also indicate that TE-derived sequences within genes are more highly correlated with their expression level than the overall GL. Thus, the selective mechanism for streamlining highly expressed genes may be related more to the elimination, or shortening, of TE sequences per se rather than the overall shortening of genes.

TE Gene Environment and the Genomic Design Hypothesis

The relationship between GL and BE seen here is generally weak; GL has one of the lower individual correlations with BE (fig. 3*G*), and variability in GL only contributes 9% of the variability seen in BE (table 1). In addition, the results show that although all the longest genes are narrowly expressed, there are about as many compact narrowly expressed genes as there are compact broadly expressed genes (fig. 4*H*). Even more surprising is the fact that the partial correlation value for GL versus BE is positive, albeit marginally (table 3), and not negative as can be expected if more narrowly expressed genes are in fact longer.

To interrogate the genomic design hypothesis more closely, we used TS as an alternate measure for the tissue specificity of expression. The genomic design hypothesis

Table 3
The Relationship between TE Fractions, GL, and Gene Expression

Expression Parameter	TE and GL	Coefficient of Determination		Partial Correlation	
		R^{2a}	*P* Value	r^b	*P* Value
PE	All TEs	0.78	<2.2E-16	−0.13	2.1E-01
	L1	0.75	<2.2E-16	−0.86	2.6E-63
	LTR	0.60	<2.2E-16	−0.20	4.5E-02
	GL	0.48	1.1E-15	−0.13	2.2E-01
	DNA	0.29	4.2E-09	−0.01	9.4E-01
	L2	0.27	2.0E-08	−0.25	1.4E-02
	MIR	0.06	6.3E-03	0.25	1.1E-02
	Alu	0.03	5.0E-02	0.32	1.1E-03
BE	All TEs	0.76	<2.2E-16	−0.10	3.1E-01
	Alu	0.59	<2.2E-16	0.52	3.0E-09
	LTR	0.57	<2.2E-16	−0.37	1.0E-04
	L1	0.47	2.8E-15	−0.52	2.4E-09
	MIR	0.12	2.2E-04	−0.28	3.6E-03
	GL	0.04	3.2E-02	0.15	1.5E-01
	L2	0.02	7.4E-02	0.08	4.4E-01
	DNA	0.01	1.3E-01	0.14	1.7E-01
TS	All TEs	0.66	<2.2E-16	−0.32	8.8E-04
	L1	0.63	<2.2E-16	−0.67	9.5E-19
	GL	0.53	<2.2E-16	−0.05	6.3E-01
	L2	0.30	3.0E-09	−0.21	3.3E-02
	Alu	0.29	5.0E-09	−0.13	2.2E-01
	LTR	0.28	9.4E-09	−0.24	1.8E-02
	MIR	0.27	2.1E-08	0.31	1.6E-03
	DNA	0.24	1.8E-07	−0.04	7.3E-01

[a] R^2 (the coefficient of determination) is the fraction of variability in each expression parameter that can be attributed to the variability in each sequence feature (individual TE families, GL, or all TEs combined).

[b] *r* is the partial correlation of each feature with the expression parameters, taking into account the presence of the other elements. For each expression parameter, the TEs and GL are ranked by their predictive value for the parameter.

posits that increasing GL is based on the requirement for additional regulatory sequences in genes that are expressed more narrowly. Thus, in the case of TS, a positive correlation is expected between GL and TS; in other words, longer genes are expected to be more tissue specific. For the pairwise regression analysis, there is actually a strongly negative correlation between GL and TS (fig. 5*H*). This negative trend holds when the TE fractions are controlled for in the partial correlation, and GL also has a high coefficient of determination for TS (table 3). It should be noted that the negative correlation between GL and TS may be related to the analytical formulation used to compute TS (see Materials and Methods) because genes with high expression levels in one or a few tissues (i.e., high PE) will often, but not always, have high TS as well. Nevertheless, when taken together, the data for both GL versus BE and GL versus TS seem to argue against the genomic design hypothesis as originally conceived.

With respect to the TEs, there are strongly positive (Alu; fig. 4*A*) and negative (L1; fig. 4*C*) correlations between TE gene fractions and BE, and 76% of the variability in BE can be attributed to variability in all TE gene fractions (table 3). Overall, TE gene fractions also have the highest coefficient of determination for TS. Consistent with what was previously shown for PE, these data suggest that the combinatorial impact of TEs in human genes is more important than the overall GL with respect to the number of tissues in which a gene is expressed and the tissue specificity of genes.

L1 Elements and Gene Expression Levels

As described previously, the data analyzed here provide support for the selection hypothesis because GL is more strongly (negatively) correlated with PE than BE. However, the strongest negative correlation with PE in the pairwise regression analysis is seen for L1 gene fractions (fig. 3*C*). L1 also has the highest negative partial correlation with PE in the multiple regression analysis and the highest coefficient of determination (table 3); 75% of the variability in PE is attributable to L1 gene fractions compared with the 48% explained by GL. Thus, L1 gene fractions are more predictive of PE than GL, indicating that variation in the gene fractions of L1s is associated with a higher change in gene expression than variation in GL.

It is also possible that regional genomic features, such as GC content, contribute to the apparent effect of L1 gene content on PE. It is known that L1 elements are enriched in GC-poor regions (Smit 1999; Lander et al. 2001), whereas GC content is strongly positively correlated with PE and BE (Vinogradov 2005). Thus, one may expect to see the kind of negative correlations between L1 and PE/BE seen here based solely on regional biases in GC content. We performed partial correlation to separate the effects of L1 gene fractions and GC content on both PE and BE. When we control for GC content, the partial correlation of L1 fractions with PE remains highly significant (table 4). Conversely, when we control for L1 fractions, the partial correlation of GC with PE is rendered insignificant (table 4). Both L1 fractions and GC content show similar levels of relatedness with BE and partial correlation analysis does not remove either effect (table 4). Thus, the relationship between L1 gene fractions and PE/BE cannot be explained solely by the genomic distribution of L1s among different GC content regions.

L1 elements are an abundant and recently active family of LINEs that make up 17% of the human genome sequence (Lander et al. 2001; Venter et al. 2001). Experimental studies have demonstrated that the presence of L1 sequences within genes can lower transcriptional activity (Han et al. 2004; Ustyugova et al. 2006). The effect of the presence of L1s on PE observed here may be attributed to the fact that the disruptive activity of L1s on transcription inhibits gene expression more than an overall increase in GL does. However, this finding is not entirely inconsistent with the selection hypothesis, rather it suggests a specific mechanism, namely the elimination of L1 sequences, for selectively

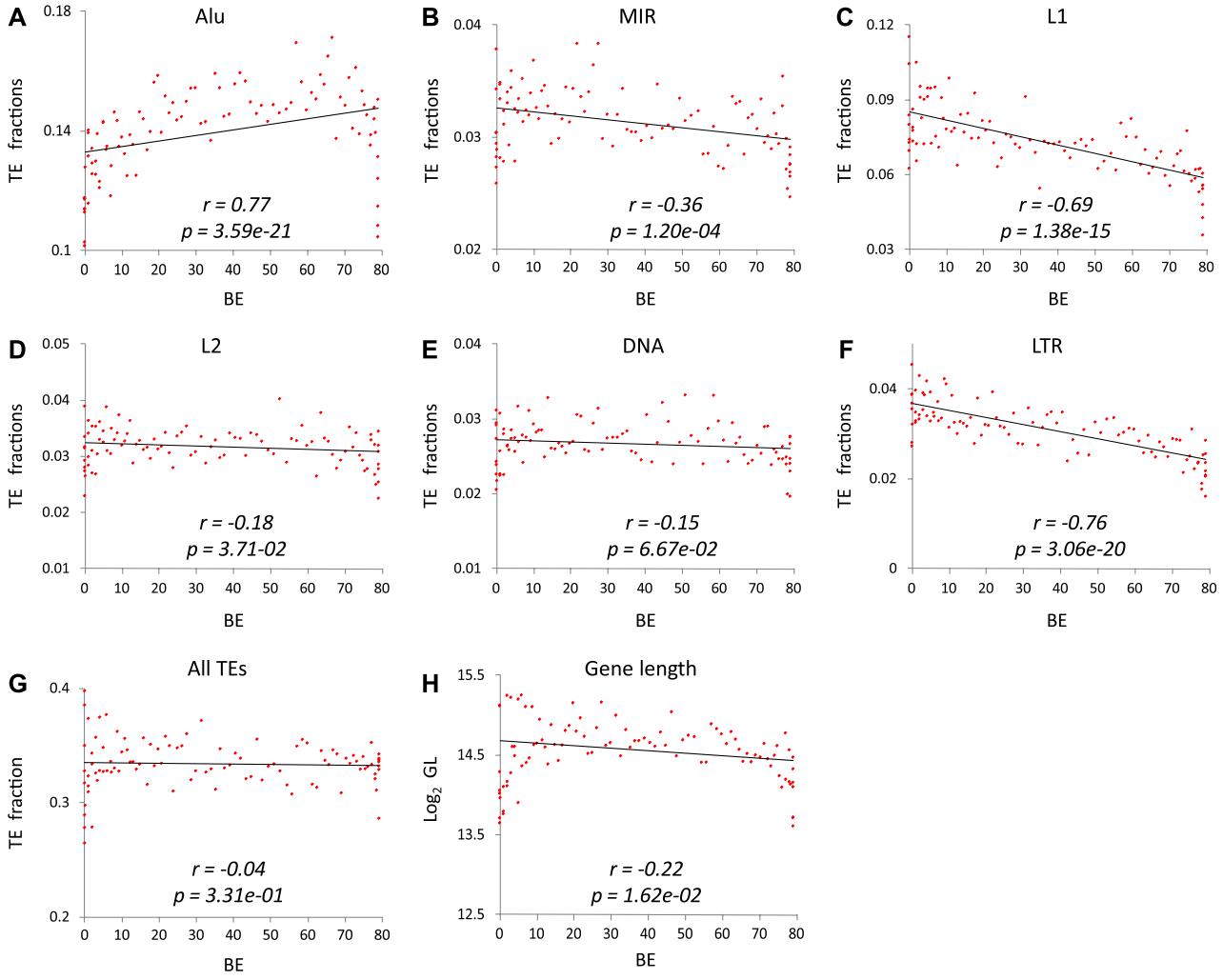

Fig. 4.—TE fractions, GL, and the breadth of expression (BE). Relationships between the TE gene fractions for (*A*) Alu, (*B*) MIR, (*C*) L1, (*D*) L2, (*E*) DNA, (*F*) LTR, and (*G*) all TEs and the BE of human genes. (*H*) Relationship between GL and BE. Pearson correlation coefficient values (*r*) along with their significance values (*P*) are shown for all pairwise regressions.

tuning highly expressed genes that would also result in an overall decrease in their length.

MIR Elements and Tissue-Specific Gene Expression

The genomic design hypothesis posits a requirement for additional regulatory sequence elements that facilitate TS, which in turn leads to an increase in GL. However, data reported here show that the presence of such regulatory elements does not necessarily result in an overall increase in GL as predicted by the genome design hypothesis (fig. 5*H*). In light of this realization, we sought to evaluate whether any specific TE sequence elements might be related to the regulatory complexity entailed by tissue-specific genes. Of all the TE families evaluated, MIRs are the only elements that show the expected trends for the genome design hypothesis for both BE and TS. The fraction of MIRs in human genes is negatively correlated with BE (fig. 4*B*) and positively correlated with TS (fig. 5*B*) as expected. In fact, MIRs are the only TEs positively correlated with TS, and the increase in the MIR gene fraction is not linear with increasing TS. At the high range of TS (>0.7; 58% of all genes), the positive correlation of MIR gene fractions to TS is even stronger ($r = 0.78$, $P = 3.7 \times 10^{-18}$).

These results are interesting in light of what is already known about MIRs. MIR elements (mammalian-wide interspersed repeats) are an ancient family of transfer RNA–derived SINEs (Jurka et al. 1995; Smit and Riggs 1995), and they have previously been implicated as having regulatory significance in a number of studies. Initially, human MIR sequences were shown to be highly conserved over time suggesting that they may encode some unknown regulatory function (Silva et al. 2003). Subsequently, MIR-derived sequences have been shown to donate transcription factor–binding sites (Polavarapu et al. 2008; Wang et al. 2009),

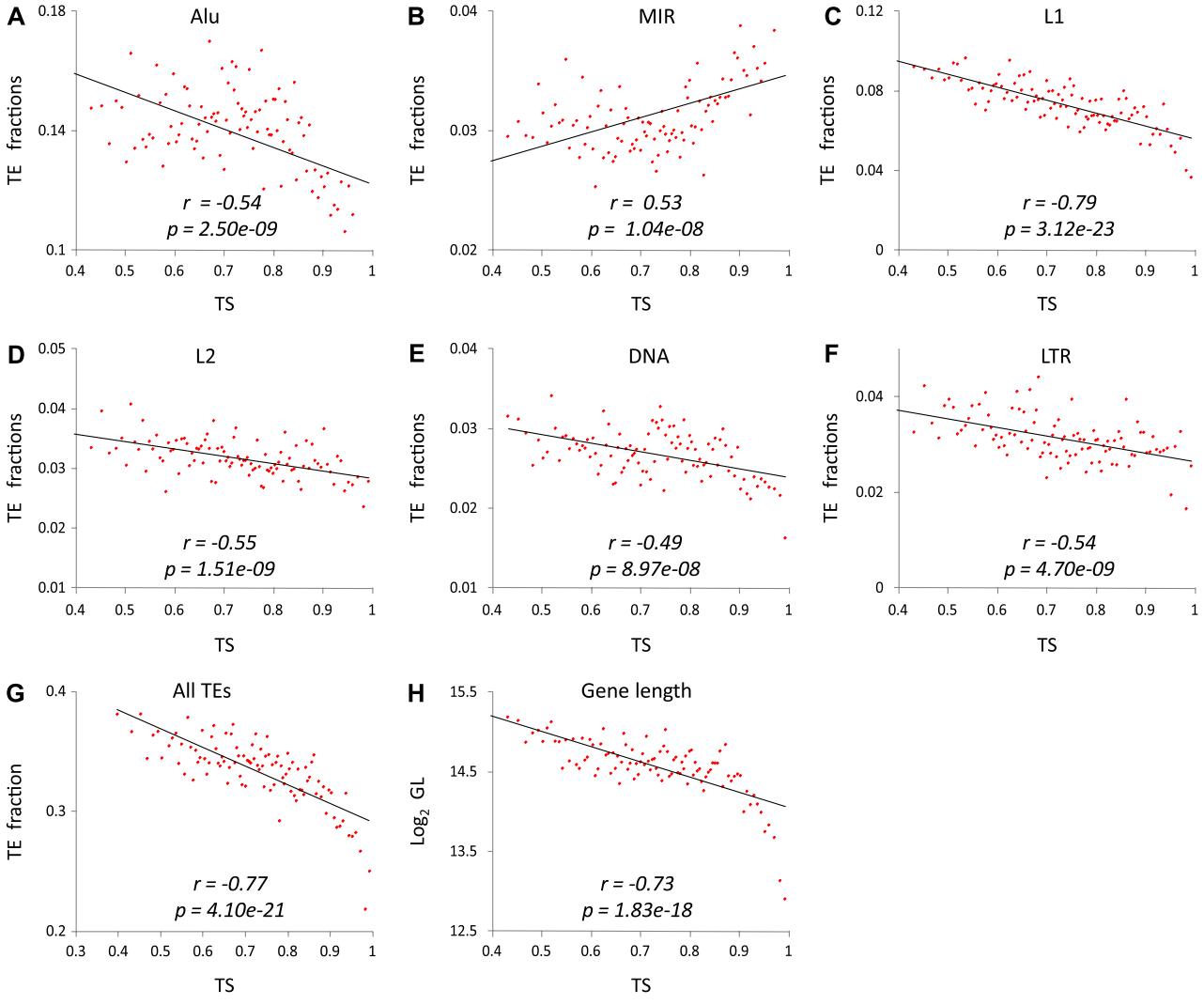

Fig. 5.—TE fractions, GL, and TS. Relationships between the TE gene fractions for (*A*) Alu, (*B*) MIR, (*C*) L1, (*D*) L2, (*E*) DNA, (*F*) LTR, and (*G*) all TEs and the TS of human genes. (*H*) Relationship between GL and TS. Pearson correlation coefficient values (*r*) along with their significance values (*P*) are shown for all pairwise regressions.

enhancer sequences (Marino-Ramirez and Jordan 2006), microRNAs (Piriyapongsa et al. 2007), and cis-natural antisense transcripts (Conley et al. 2008) to the human genome. In addition, it has been shown that, whereas TEs are generally depleted from introns, MIRs are actually significantly enriched within genes that might require subtle regulation of transcript levels or precise activation timing, such as growth factors, cytokines, hormones, and genes involved in the immune response (Sironi et al. 2006). Such genes would be expected to be largely tissue specific.

If MIRs donate regulatory sequences to tissue-specific genes, then one may expect to observe relative increases in MIR density in the regulatory regions upstream and downstream of TSSs. To evaluate this possibility, we took the top 10% tissue-specific genes and evaluated their MIR frequencies at 1-kb intervals along a 20-kb window surrounding the gene TSS. As with all other TEs, MIRs show a marked decline in frequency most proximal to the TSS. However, MIRs show a unique pattern of enrichment both upstream and

Table 4

Effect of GC Content on the Relationship between L1 Genic Fractions and Gene Expression

	Feature[a]	*r*	*P* Value	Control[b]	*r*	*P* Value
PE	L1	−0.87	1.69E-31	L1 | GC	−0.73	1.3E-25
	GC	0.69	1.20E-15	GC | L1	0.12	2.2E-01
BE	L1	−0.69	1.38E-15	L1 | GC	−0.44	1.7E-06
	GC	−0.21	2.00E-02	GC | L1	0.44	1.4E-06
TS	L1	−0.79	3.12E-23	L1 | GC	−0.77	3.0E-32
	GC	0.32	6.81E-04	GC | L1	−0.03	7.5E-01

[a] L1 genic fractions and genic GC content values are correlated with the expression parameters PE, BE, and TS (tissue-specificity).

[b] Partial correlation analyses control for effect of GC content on L1 fractions (L1 | GC) and L1 fractions on GC content (GC | L1), respectively.

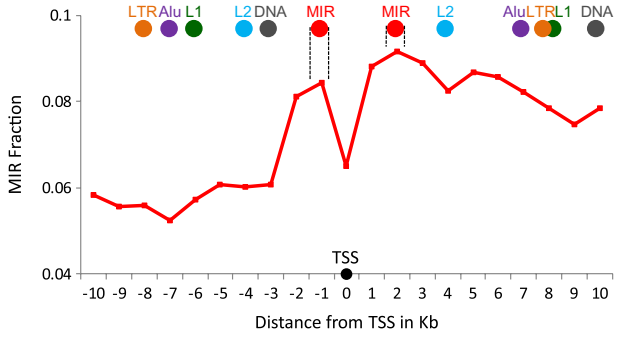

Fig. 6.—The local frequency maxima of TE densities around the TSSs of tissue-specific genes. The red line shows the density distribution of MIRs around TSSs. Colored dots show the locations of the local frequency maxima for the different TE classes/families.

downstream of the TSS, just outside the proximal promoter region, compared with other families of TEs. In fact, MIRs are the only elements that show local frequency maxima at −1 kb and +2 kb with respect to the TSS. All other TEs show their maxima in more distal regions from the TSS (fig. 6). This pattern is consistent with a unique regulatory role for MIRs, perhaps owing to the donation of cis-regulatory elements, as compared with other TEs.

If the regulatory effect of genic MIRs is based on the donation of shared transcription factor–binding sites, then one may expect the tissues in which MIR-rich genes are expressed to be similar. We evaluated this prediction in two ways. First, we took the top 10% MIR-rich genes and for each gene we determined the tissue in which it was maximally expressed. The observed frequency distribution for these tissues was compared with a randomized distribution of the same number of genes among all tissues in the microarray data set analyzed here using a χ^2 test. The observed distribution is far from random (supplementary fig. S4, Supplementary Material online; $\chi^2 = 1{,}406.8$, $P = 1.1 \times 10^{-242}$), and there are a number of specific tissues, and groups of related tissues, that are overrepresented, particularly liver, blood-related tissues, reproductive tissues and nervous tissues. Second, we clustered the expression patterns of the top 10% MIR-rich genes using hierarchical clustering based on the Euclidean distances between their gene expression patterns over 79 tissues. Several of the resulting clusters show groups of MIR-rich genes that are markedly overexpressed among these same related groups of tissues (fig. 7).

MIRs are a relatively ancient family of TEs that are conserved among mammals including mouse. We evaluated TE gene fraction and expression data for mouse, in the same way as was done for humans, to see if the same trends in the relationship between MIR gene fractions and tissue specificity hold for mouse elements. As is the case for the human genome, mouse MIR elements are the only family of TEs with genic fractions that are significantly positively correlated with TS (table 5). This suggests the possibility that MIR elements have been conserved among mammalian genomes, at least to some extent, by virtue of their regulatory contributions.

The genomic design hypothesis predicts that additional regulatory sequence elements required by tissue-specific genes will lead to an increase in their overall length. However, with respect to MIRs, our analysis suggests that the enrichment of regulatory elements in tissue-specific genes does not lead to an increase in the overall length of genes. Rather, the regulatory complexity required by tissue-specific genes may be achieved in some cases via the donation of a few key sequence elements provided by TEs that come preequipped with existing regulatory capacity.

Conclusions

The architecture of human genes has important implications for how they are expressed. Previous studies on this topic have focused separately on the influences of GL or the TE environment on gene expression. Here, we show that these two factors are closely related, and we consider them jointly in an attempt to dissect their individual contributions. Consistent with previous results, we observed GL to be strongly correlated with PE and less so with BE. We also show that GL is strongly correlated with TS but not in the direction that is expected according to the genomic design hypothesis. These data provide strong support for the selection hypothesis. However, we show that the TE fraction of human genes has a stronger overall effect on gene expression than does GL. Considered together, TE gene fractions explain 78%,

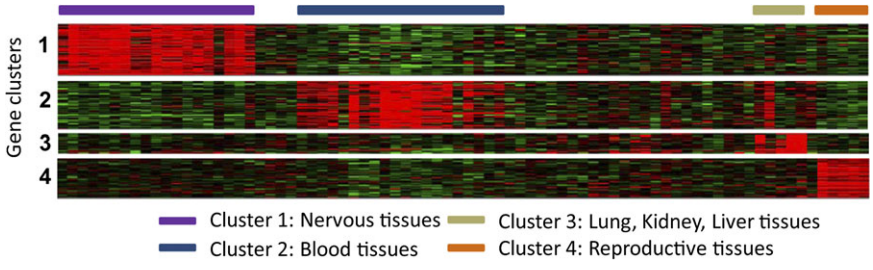

Fig. 7.—MIR-rich genes hierarchically clustered into groups of similar expression profiles across tissues. The clusters show maximum expression in related sets of tissues.

Table 5
Relationship between Genic TE Fractions and Tissue-Specificity in Mouse[a]

TE Family	*r*	*P* Value
MIR	0.37	7.5E-05
LTR	0.12	1.2E-01
L1	0.08	2.2E-01
DNA	0.07	2.6E-01
L2	−0.25	5.6E-03
ID	−0.40	2.1E-05
B4	−0.46	5.9E-07
B1	−0.74	1.6E-18
B2	−0.74	4.9E-19

[a] Genic TE fractions for mouse TE families were correlated with tissue-specificity in the same way as done for human TE families (see fig. 5).

76%, and 66% of the variability observed for PE, BE, and TS respectively, in all cases greater than what is seen for GL. We also uncover examples where individual TE families, L1s, and MIRs respectively, have marked effects on the level and breadth of gene expression.

Consideration of intergenic TE fractions and GC content together with TE gene fractions suggests that the relationships between TE gene fractions and GL and expression are not solely related to regional genomic processes. However, there may be other as yet undetected regional genomic factors that could mitigate the apparent relationships between TE gene fractions and GL and expression. Nevertheless, the results reported here underscore the potential regulatory implications of the TE environment of human genes and also suggest specific mechanisms for how TEs may contribute to gene regulation.

Supplementary Material

Supplementary figures S1, S3, and S4 and table S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Nathan J. Bowen for guidance on the gene expression analysis. We would like to thank members of the Jordan lab for their support and technical assistance. D.J. was supported by a Fulbright predoctoral fellowship. I.K.J. and A.H. were supported by the School of Biology, Georgia Institute of Technology, and an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839). This research was supported in part by the Intramural Research Program of the National Institute of Health, National Library of Medicine, NCBI.

Literature Cited

Carmel L, Koonin EV. 2009. A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. Genome Biol Evol. 2009:382–390.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415–418.

Chen J, Sun M, Hurst LD, Carmichael GG, Rowley JD. 2005. Human antisense genes have unusually short introns: evidence for selection for rapid transcription. Trends Genet. 21:203–207.

Comeron JM. 2004. Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 167:1293–1304.

Conley AB, Miller WJ, Jordan IK. 2008. Human cis natural antisense transcripts initiated by transposable elements. Trends Genet. 24:53–56.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet. 19:362–365.

Eller CD, et al. 2007. Repetitive sequence environment distinguishes housekeeping genes. Gene 390:153–165.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature 429:268–274.

Jordan IK, Marino-Ramirez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. Gene 345:119–126.

Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Jurka J, Zietkiewicz E, Labuda D. 1995. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. Nucleic Acids Res. 23:170–175.

Karolchik D, et al. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32:D493–D496.

Kim TM, Jung YC, Rhyu MG. 2004. Alu and L1 retroelements are correlated with the tissue extent and peak rate of gene expression, respectively. J Korean Med Sci. 19:783–792.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 7:474.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lerat E, Semon M. 2007. Influence of the transposable element neighborhood on human gene expression in normal and tumor tissues. Gene 396:303–311.

Li SW, Feng L, Niu DK. 2007. Selection for the miniaturization of highly expressed genes. Biochem Biophys Res Commun. 360:586–592.

Marino-Ramirez L, Jordan IK. 2006. Transposable element derived DNaseI-hypersensitive sites in the human genome. Biol Direct. 1:20.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 12:1483–1495.

Pereira V, Enard D, Eyre-Walker A. 2009. The effect of transposable element insertions on gene expression evolution in rodents. PLoS One. 4:e4321.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337.

Polavarapu N, Marino-Ramirez L, Landsman D, McDonald JF, Jordan IK. 2008. Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics. 9:226.

Rhead B, et al. 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. 38:D613–D619.

Seoighe C, Gehring C, Hurst LD. 2005. Gametophytic selection in Arabidopsis thaliana supports the selective model of intron length reduction. PLoS Genet. 1:e13.

Silva JC, Shabalina SA, Harris DG, Spouge JL, Kondrashovi AS. 2003. Conserved fragments of transposable elements in intergenic regions:

evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. Genet Res. 82:1–18.
Simons C, Pheasant M, Makunin IV, Mattick JS. 2006. Transposon-free regions in mammalian genomes. Genome Res. 16:164–172.
Sironi M, et al. 2006. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. Genome Biol. 7:R120.
Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657–663.
Smit AF, Riggs AD. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. 23:98–102.
Stalteri MA, Harrison AP. 2007. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. BMC Bioinformatics. 8:13.
Sturn A, Quackenbush J, Trajanoski Z. 2002. Genesis: cluster analysis of microarray data. Bioinformatics 18:207–208.
Su AI, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 101:6062–6067.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. Genome Res. 13:2260–2264.
Ustyugova SV, Lebedev YB, Sverdlov ED. 2006. Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts. Genetica 128:261–272.
Venter JC, et al. 2001. The sequence of the human genome. Science 291:1304–1351.
Versteeg R, et al. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC-content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. 13:1998–2004.
Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. 20: 248–253.
Vinogradov AE. 2005. Dualism of gene GC-content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. Trends Genet. 21:639–643.
Wang J, Bowen NJ, Marino-Ramirez L, Jordan IK. 2009. A c-Myc regulatory subnetwork from human transposable element sequences. Mol Biosyst. 5:1831–1839.
Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.

Associate editor: Marta Wayne