Expression levels and gene function influence transposon occurrence in mammalian introns

Pozzoli U (1), Menozzi G (1), Cereda M (1), Comi GP (2), Sironi M (1)

(1) Scientific Institute IRCCS E.Medea - Bioinformatics Lab.
(2) Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale MaggiorePoliclinico, Mangiagalli and Regina Elena Foundation , 20100 Milan, Italy.

Motivation

Transposable elements (TEs) represent more than 45% and 37% of the human and mouse genomes, respectively. Once considered as merely junk DNA, it is now widely recognized that interspersed repeats have been playing a major role in genome structure evolution. Several studies have suggested that TE integrations have been subjected to purifying selection to limit the genetic load imposed to their host. Yet, a comprehensive analysis of the forces driving TE insertion, fixation and maintenance within mammalian genes is still missing.

Methods

NCBI Reference Sequence genes were selected for human and mouse. Gene and intergenic sequences as well as intron/exon boundaries were derived from the UCSC genome annotation database (http://genome.ucsc.edu/). Gene counts were: 7695 and 5550, for human and mouse, respectively, accountiong for 81989 and 55553 introns. Mulispecies conserved sequences (MCS) were obtained using phastCons predictions, which are available through the UCSC database. Transposable elements were identified and categorized using the UCSC annotation tables that rely on RepeatMasker. Microarray data on expression levels in human and mouse tissues were derived from previous studies based on high-density oligonucleotide arrays (GNF Gene Expression Atlas 2). For human and mouse, SAGE libraries, were obtained from the SAGE Genie website (http://cgap.nci.nih.gov/SAGE). For each transcript entry in our databases we extracted a SAGE tag; tags were then matched to all RefSeq mRNAs and purged if they corresponded to more than one transcript. We than matched our tags to those in libraries, added all counts for libraries representing the same tissue type and converted absolute counts to relative tag counts (c.p.m.). Statistical analysis were performed using R. For lowess smooths, five robustifying iterations were always performed and a smoothing span of 0.5 was used. To allow empirical p value calculations, we performed 100 independent random data permutations.

Results

We had previously suggested that TE insertions might be constrained in human introns by the presence of conserved elements. We analyzed the distribution of different TE families, namely Alu, MIR, L1, L2, LTR and DNA transopsons in human/mouse introns. Correlation analysis revealed that, in both mammals, the frequency of all TE families negatively correlates with MCS density when introns containing at least one MCS are analyzed. We next wished to verify whether different TE families might be differentially represented depending on gene function. TE frequency varies with intron length, GC%, and, as shown above, MCS density. For each TE family we performed multiple regression analysis using intron GC%, intron length and conserved sequence length as covariates; the regression fit was used to predict the expected TE number per intron (nTEiexp). For each gene the TE normalized abundance (TEna) was calculated as follows: TEna=[sum(nTEi exp)-sum(nTEi obs)]/[sum(nTEi exp)+sum(nTEi obs)] where nTEiobs is the observed TE number per intron. Genes displaying TEna>0.5 or TEna<-0.5 were classified as TE-rich or -poor, respectively and significant gene ontology associations were retrieved. Genes involved in morphogenesis/development were found to be overrepresented in all TE-poor groups (as previously noticed for the HOX gene cluster); the same holds true for genes encoding transcription factors and cellular proteins involved in basic functions. Interestingly, we identified another group of genes that in both human and mouse, are over-represented among TE-poor genes: it is the case of hormones and ckemokines/cyotkines. This finding suggests that the majority of TE, including Alus and old

TEs (usually considered relatively benign dwellers of mammlian genomes), might exert disturbing effects on genes that require subtle tuning of expression levels. TEs have been reported to differentially associate with gene regions depending on expression levels. In order to address this issue we analyzed variation in Tena as a function of mean gene expression (obtained from both microarray and SAGE data). Lowess smooths were calculated for each TE family and compared to curves obtained from random data permutations. In both mammals a marked decrease in Tena is observed for genes above the 70th- 80th gene expression percentile. We next calculated the intronic to intergenic normalized frequency difference: again, for highly expressed genes and for all TE families, a decreasing trend is observed when frequency differences are plotted against gene expression. These data suggest that independently of gene isochore and repeat type, a TE exclusion from intronic regions is observed that depends on gene expression level. In summary our data indicate that, although different TE types might exert distinct effects on gene regulation, gene features such as intronic MCS density, gene function and expression have been playing a major role in governing TE fixation and maintenance in mammalian intronic regions.

Contact email: uberto.pozzoli@bp.lnf.it