Introns containing conserved elements are evolutionary preserved in size

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

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

Motivation

Different explanations have been proposed to account for within-genome intron size variations. Since we have previously demonstrated that fixation of multispecies conserved sequences (MCSs) influences intron size in humans, we wished to analyze whether this finding might be exploited to define a model for intron size evolution in mammals.

Methods

All genomic sequences were obtained from the UCSC genome annotation database (http://genome.ucsc.edu/). Only NCBI Reference Sequence genes were selected for human and mouse. Gene counts were: 7695 and 5550, for human and mouse (81989 and 55553 introns), respectively. Chimpanzee and rat genomic sequences were derived from UCSC; human/mouse mRNA alignments onto chimpanzee/rat genomic sequences were retrieved from the UCSC genome annotation database. Orthologous introns were then aligned using EMBOSS Stretcher with penalties of 48 and 1 for gap opening and extension, respectively. Human sequences mapping to gaps on the chimpanzee genome and covered by a Transposable element (TE, at least 85% of bases in TE and TE sequence extending no more than 10 bp 3' and 5' the gap) were considered humanspecific insertions. Conversely, gaps longer than 80 bp in the human genome that were not accounted for by TE insertions or microsatellites in chimpanzee were classified as deletions. The same procedures were applied to mouse/rat intron alignments. For the identification of humanmouse and human-chicken orthologous pairs, the EnsMart database was interrogated and only entries representing unique best reciprocal hits (UBRH) were selected. MCS were obtained using phastCons predictions (available through the UCSC database). Transposable elements were identified and categorized using the UCSC annotation tables that rely on RepeatMasker. All statistical analysis were performed using R (http://www.r-project.org/). For lowess smooths, five robustifying iterations were always performed and a smoothing span of 0.5 was used.

Results

In order to study the impact of MCS presence on intron size, we aligned 27008 mouse-rat orthologous intron pairs and recorded mouse TE insertions and deletion events. Intron length before mouse-rat divergence (original length) was then estimated. In the case of mouse, TE insertion and deletion frequencies were analyzed after dividing introns in 4 length classes and, independently, in 4 groups depending on MCS density. Deletion frequency increase with intron size but, within each length class, both deletion and TE insertion frequencies diminish at the increase of MCS content (Kruskall-Wallis $p< 0.01$ for differences within all size groups). Interestingly, the deletion/insertion frequency ratio also decreases with MCS density increase (Kruskall-Wallis $p< 0.01$ for the second and third length classes). Analysis of the net variation in mouse intron size relative to the common mouse-rat ancestor indicated an average shrinkage, stronger for longer introns; yet, size contraction is progressively reduced, within each length class , for increasing MCS densities (Kruskall-Wallis $p< 0.01$), with introns being almost invariant when extremely rich in conserved sequence. For primate introns, we identified 280 TE insertions and 592 deletions in 39323 introns. Given the paucity of events and in order to analyze deletion and TE insertion frequency distribution as a function of MCS content, we applied a simulation based approach. For both insertions and deletions (separately) 1000 independent simulations were performed by randomizing TE insertion positions or deletion events. In the case of deletions we did not allow resulting introns shorter than 25 bp. For each intron, simulated TE insertion and deletion frequencies were assumed to conform to a Poisson distribution and lambda was calculated. The Wilcoxon test for paired samples was used to compare

the observed frequency with the expected (lambda). MCS-containing introns displayed significantly ($p<0.001$) less TE insertions compared to simulations. The same result applied to deletion frequency (paired Wilcoxon test, $p<0.001$). Conversely, MCS-lacking introns had significantly (paired Wilcoxon test, $p<0.001$) higher TE insertion and deletion frequencies. Consistent with the above findings is the analysis of normalized size variation (Δlength = [human - mouse length]/[human + mouse length]) in human-mouse orthologous intron pairs; lowess curves indicated that MCS-containing introns are extremely conserved in length irrespective of their size; conversely intronic regions lacking MCSs display remarkable size conservation until below length values ~1 kb and then diverge rapidly. Stronger size conservation for MCS-containig vs MCS-lacking introns also occurs over longer evolutionary periods as we were able to show by the analysis of humanchicken orthologous intron pairs. These results clearly indicate that MCSs have been posing a strong constraint to intron size evolution

Contact email: manuela.sironi@bp.lnf.it