Conserved elements shape both size and base composition of noncoding regions in Drosophila

ID - 229

Menozzi Giorgia[1], Manuela Sironi[1], Uberto Pozzoli[1]

[1]Scientific Institute I.R.C.C.S. E. Medea, Via Don Luigi Monza 20, 20842 Bosisio Parini (LC), Italy

Motivation

The great majority of multicellular eukaryote genomes is accounted for by non-coding sequences. Yet, our knowledge of the forces affecting the evolution of such large regions are poorly understood. In particular, we have previously studied the evolution of intron size in mammals and proposed that the need to maintain multispecies conserved noncoding sequences (MCS) has played a major role in shaping intron and intergenic sequence size. We now wish to verify our previous findings using a model system such as the fruitfly Drosophila melanogaster. The advantages offered by this model organism are mainfold: compact genome, many different Drosophila genomes sequenced allowing exact identification of both micro- and macro-indels, availability of gene expression and recombination data. Recent works (Halligan and Keightley, 2006; Haddrill et al., 2005) have identified a negative correlation between divergence and both intron size and GC content. Here we wish to demonstrate that the relation between constraint and size is causal, with constrained sequences having a role in size increase; with respect to GC content we suggest that constrained sequences contribute to GC content possibly as a result of a strong GC to AT mutation bias at neutral sites.

Methods

Genomic sequences and annotations were derived from the UCSC genome database (http://genome.ucsc.edu/, dm2 assembly). A total of 11154 genes were retrieved, accounting for 42789 introns. MCSs were obtained through the UCSC database (phastConsElements9way table) and derive from 9-way species alignments. Insertions and deletions were identified in D.melanogaster/D.simulans alignments using D.yakuba as an outgroup.

Insertion/deletion data were also used to infer the original sequence length in the D.melanogaster/D.simulans ancestor.

In order to analyze insertion and deletion frequencies, introns and intergenic spacers were divided in MCS density and size classes, independently. Given the bimodal distribution of intron size, introns shorter than 80 bp (corresponding to the 54-th quantile) were ascribed to the first length class; similarly, for MCS density the first class was accounted for by sequences displaying no conserved elements.

Expression data at different developmental stages were derived from a previous work (Arbeitman et al., 2002).

Results

We calculated MCS density in intronic and intergenic regions: a highly significant correlation was observed between MCS density and intron or 5/3intergenic size (Spearman rho ranging from 0.67 to 0.75, p< 0.0001), indicating that MCS density explains 45-50% of noncoding sequence size variation. As expected the frequency of both micro- and macro-indels was significantly lower in MCSs compared to non-conserved noncoding sequence (Wilcoxon Rank Sum Test, p<0.001).

We next verified that both micro- and macro-indels are counterselected by the presence of MCSs. Most importantly, the relative (to the common D.melanogaster/D.simulans ancestor) size variation is negligible in small introns, while it goes from negative to positive in larger introns as MCS density increases.

In short, MCSs have been playing a role in intron size evolution through their differential impact on insertions and deletions. We next wished to determine whether MCS densities varied with gene expression patterns. In analogy to what we observed in mammals, genes expressed in one or few developmental stage display higher MCS densities compared to housekeeping or widely expressed genes (Wilcoxon Rank Sum test p< 0.001); the same considerations hold true for both 5 and 3 intergenic spacers, indicating that developmental-stage specific genes need increased regulatory elements.

In contrast to observations in mammals, we failed to verify any sign of selection to minimize intron size in highly expressed genes: when MCS number per intron was fixed, no decrease in intron size was observed for highly or widely expressed genes. With respect to GC content, we verified that MCSs display a significantly higher GC content compared to flanking intron sequences (Wilcoxon Rank Sum Test for paired samples, p< 0.0001). The positive correlation previously described (Haddrill et al., 2005) between sequence constraint and GC content might therefore result from the selection against mutation in MCSs compared to flanking neutral sites. Still, a positive correlation between non-conserved GC content and MCS density is still observed (Spearman rho = 0.33, p<0.001). This might reflect the presence of functional elements within introns that we do not identify as MCSs but are subjected to selective pressure against mutation and, therefore (given the bias of GC to AT mutations), against GC loss.

We suggest that MCS presence shapes both size and base composition of noncoding regions in the fruitfly.

Email: giorgia.menozzi@bp.lnf.it