HUMAN GENOME POLYMORPHISM
AND ALTERNATIVE SPLICING

Ramensky V.*, Nurtdinov R., Neverov A., Mironov A., Gelfand M.

1 Engelhardt Institute of Molecular Biology, RAS, Moscow, 119991, Russia; 2 Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia; 3 State Scientific Center GosNIIGenetika, Moscow, 117545, Russia; 4 Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia

*Corresponding author: e-mail: ramensky@imb.ac.ru

Key words: alternative splicing, single nucleotide polymorphism, McDonald-Kreitman test, positive selection, molecular evolution

SUMMARY

Motivation: Alternative splicing and single nucleotide polymorphism are the phenomena responsible for the organismal molecular diversity. We present the study of polymorphism in the coding regions of 9,125 alternatively spliced human genes aimed at understanding of the forces driving their evolution.

Results: The significant difference in non-synonymous to synonymous SNP ratio is observed: 0.757 in constitutive exons vs. 1.799 in alternative ones from minor isoforms. The alternatively spliced regions therefore experience lower selective pressure at the amino acid level. The results of the McDonald-Kreitman test suggest that, unlike the constitutive regions, they are also subject to the positive selection.

INTRODUCTION

Alternative splicing (AS) of genes has been recognized as one of the major sources of organismal complexity (Graveley, 2001). The AS can be defined as the various ways of splicing out introns in eukaryotic pre-mRNAs resulting in one gene producing several different mRNAs and protein products (isoforms). The estimates of fraction of alternatively spliced genes in the human genome gradually increase from the level of 35 % (Mironov, 1999) up to 70–80 % (Johnson, 2003; Kampa, 2004). The alternative splicing has also been recognized as the mechanism of accelerated evolution by relaxation of purifying selection pressure (Cusack, Wolfe, 2005; Xing, Lee, 2005; Ermakova et al., 2006, in press). Single nucleotide polymorphism (SNP) is another well-recognized phenomenon providing molecular diversity (Brookes, 1999) and comprising approximately 90 % of human DNA variation (Collins et al., 1998). The SNPs in the coding regions of genes fall in two categories, synonymous or silent (sSNPs) and non-synonymous (nsSNPs) that change the corresponding amino acid residue and therefore are thought to be responsible for existence of various phenotypes (Collins et al., 1998). The relationship between the non-synonymous (Pa) and synonymous (Ps) SNP (Zhao et al., 2003) or fixed interspecies variation (Da, Ds) (Xing, Lee, 2005) provides information on selection pressure in the genomic regions under study. The analysis of SNP density in the Celera human sequence assembly showed that the bulk genomic ratio Pa/Ps is less than half of that under neutral expectations, reflecting the acting purifying (negative) selection (Zhao et al., 2003). The simultaneous account for the four abovementioned types of variation (Da, Ds, Pa, Ps) enables the detection of positive darwinian selection in
the presence of negative selection via the McDonald-Kreitman test (McDonald, Kreitman, 1991). In this work, we investigate the differences between patterns of synonymous and non-synonymous polymorphism and human-chimpanzee variation in alternative and constitutive regions of human genes. The novelty of the approach is in the simultaneous analysis of polymorphism and divergence data in the coding regions of two types that reveals not only the relaxation of selection pressure (Cusack, Wolfe, 2005; Xing, Lee, 2005) but also the significant difference between the positive selection strength in the alternative and constitutive regions.

**METHODS AND ALGORITHMS**

Validated SNPs from build 121 of the dbSNP database (Sherry et al., 2001) were mapped to human genes from the EDAS database (Nurtdinov, 2004) with at least two isoforms generated by the IsoformCounter algorithm (Neverov et al., 2005). The human-chimpanzee synonymous and non-synonymous variation were derived from the whole genome alignments (Kent et al., 2002). The conservative isoform generation procedure and subsequent postprocessing (regions between the isoform ends and stops, alternatives with frameshifts, and regions not completely aligned to the chimp genome are not considered) generates 64,742 constitutive and 18,036 alternatively spliced regions from 9,125 genes (Table 1). The considered regions include cassette exons, alternative exons, and alternative 5’ and 3’ splice sites.

**Table 1.** Alternative and constitutive regions. Minors are those included in less than 2/3 of all coding sequences (ESTs, mRNAs, peptides) observed for this regions of a gene

<table>
<thead>
<tr>
<th></th>
<th>Num. of regions</th>
<th>Total length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constitutive</td>
<td>64,742</td>
<td>9,544,015</td>
</tr>
<tr>
<td>Major</td>
<td>11,129</td>
<td>1,193,395</td>
</tr>
<tr>
<td>Minor</td>
<td>6,907</td>
<td>720,845</td>
</tr>
</tbody>
</table>

**IMPLEMENTATION AND RESULTS**

The resulting statistics on intra- and interspecies variation is given in the Table 2. The significant difference between Pa/Ps ratio in constitutive and alternative gene regions is observed. The data on human-chimp variation follows the same trend.

**Table 2.** SNPs and human-chimp mismatches in the various regions of alternatively spliced genes

<table>
<thead>
<tr>
<th></th>
<th>Pa = SNP</th>
<th>Ps = SNP</th>
<th>Pa/Ps</th>
<th>Da = Mismatch</th>
<th>Ds = Mismatch</th>
<th>Da/Ds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonsyn</td>
<td>Synon</td>
<td></td>
<td>Nonsyn</td>
<td>Synon</td>
<td></td>
</tr>
<tr>
<td>Constitutive</td>
<td>4380</td>
<td>5787</td>
<td>0.757</td>
<td>25574</td>
<td>34104</td>
<td>0.750</td>
</tr>
<tr>
<td>Major</td>
<td>588</td>
<td>707</td>
<td>0.832</td>
<td>3164</td>
<td>4268</td>
<td>0.741</td>
</tr>
<tr>
<td>Minor</td>
<td>644</td>
<td>358</td>
<td>1.799</td>
<td>4687</td>
<td>2377</td>
<td>1.972</td>
</tr>
</tbody>
</table>

However, the excess of nonsynonymous substitutions relative to polymorphism (Da/Ds = 1.972 > Pa/Ps = 1.799) in minor alternative regions implies the fixation of advantageous mutations (Fay et al., 2002). This effect increases if the most stringent conditions, when only mRNA and protein sequences are considered: Da/Ds = 1.892 vs. Pa/Ps = 1.659 (data not shown). The P-value of the $\chi^2$-test is in the range 0.1–0.2. This value can be explained by the relatively small number of SNPs.
DISCUSSION

The excess of non-synonymous SNPs and human-chimp mismatches in the alternatively spliced regions of human genes suggests that they are under lower purifying selective pressure. Unlike the constitutive regions, they also experience the positive selection, as shown by the results of McDonald-Kreitman test. The simultaneous action of these two forces shapes the evolution of alternatively spliced genes.

ACKNOWLEDGEMENTS

Authors thank E. Ermakova for valuable discussions. This work was supported by the grants “Human molecular polymorphism” and “Cellular and Molecular Biology” from the Russian Academy of Sciences.

REFERENCES