

Concluding remarks

We have identified protein domains that are preferentially found to be encoded by either tissue-specific or widely expressed genes. Domains that are enriched in tissue-specific genes are also under-enriched in essential genes, and are more likely to be absent from the proteome of a unicellular eukaryote. Hence, the families of genes encoding these protein domains are probably important for the development or terminal differentiation of many different tissue types in metazoan organisms. These observations are consistent with a model in which the development of a multicellular body plan is largely controlled by genes encoding only a limited number of protein domains that function either as DNA-binding domains or in intercellular communication.

Acknowledgements

B.L. is supported by a Sanger Institute Postdoctoral Fellowship. A.G.F. is supported by the Wellcome Trust.

Supplementary data

Supplementary data associated with this article can be found at [doi:10.1016/j.tig.2004.08.002](https://doi.org/10.1016/j.tig.2004.08.002)

References

- 1 Kasprzyk, A. *et al.* (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* 14, 160–169
- 2 Mulder, N.J. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* 31, 315–318
- 3 Su, A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 4465–4470
- 4 Huminiecki, L. *et al.* (2003) Congruence of tissue expression profiles from gene expression atlas, SAGEmap and TissueInfo databases. *BMC Genomics* 4, 31
- 5 Mootha, V.K. *et al.* (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* 115, 629–640
- 6 Eisenberg, E. and Levanon, E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.* 19, 362–365
- 7 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29
- 8 Davidson, E.H. (2001) *Genomic Regulatory Systems: Development and Evolution*, Academic Press, San Diego
- 9 Winter, E.E. *et al.* (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.* 14, 54–61
- 10 Anderson, K.V. and Ingham, P.W. (2003) The transformation of the model organism: a decade of developmental genetics. *Nat. Genet.* 33 (Suppl.), 285–293
- 11 Lu, D. *et al.* (2003) Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature* 426, 96–100
- 12 Lespinet, O. *et al.* (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048–1059
- 13 Kamath, R.S. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421, 231–237

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.08.002

Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy

Yi Xing and Christopher J. Lee

Molecular Biology Institute, Center for Genomics and Proteomics, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095-1570, USA

The importance of alternative splicing in many genomes has raised interesting questions about its role in evolution. We analyzed 13 384 full-length transcript isoforms from human and 2227 isoforms from mouse to identify sequences containing premature termination codons (PTCs) that are likely targets of mRNA nonsense-mediated decay. We found that alternatively spliced isoforms have a much higher frequency of PTCs (11.1%) compared with the major transcript form of each gene (3.7%). On the X chromosome, which is generally expressed as a single copy, the overall PTC rate was much lower (3.5%, versus 8.9% on diploid autosomes), and the effect of alternative splicing was enhanced.

Thus, diploidy and alternative splicing each increased tolerance for PTC by about threefold, as approximately additive effects. These data suggest that nonsense mediated decay might itself reduce negative selection pressure during evolution, via rapid degradation of aberrant transcripts that might yield dominant negative phenotypes.

Recently, it was proposed that alternative splicing might have a special role in evolution, by reducing negative selection pressure against large-scale mutations such as exon creation and loss [1–3]. Whereas orthologous gene structures in human, mouse and rat are highly similar, alternatively spliced exons frequently are not conserved between these genomes [3,4]. Such divergent features indicate an exon creation or loss event subsequent to the

Corresponding author: Christopher J. Lee (leec@mbi.ucla.edu).

separation of these genomes during evolution. The fact that such exon creation and/or loss events are observed at a much higher rate for alternatively spliced exons suggests that alternative splicing increases tolerance for such large-scale alterations of gene structure. Intuitively, this makes sense; insertion or deletion of an exon is likely to disrupt the reading frame of a protein, its structure or function. However, if a new exon is added as an alternatively spliced exon that is included in only small fractions of the total transcripts (e.g. 10%), the majority of transcripts will still encode the original product, greatly reducing negative selection pressure against the new exon form.

To assess the validity of this hypothesis, it would be useful to find measures of negative selection pressure that are applicable to large-scale mutations such as exon creation. This requires somewhat different measurements of selection pressure than are typically applied to small-scale mutations (e.g. Ka/Ks for single nucleotide polymorphisms [5]). One simple measurement that can be applied to large-scale mutations is the frequency of occurrence of truncated protein reading frames that are likely targets of nonsense-mediated decay (NMD) [6]. If the open reading frame (ORF) of a transcript has a premature termination codon (PTC; defined in connection with NMD as a STOP codon if > 50 nt upstream from its last exon–exon junction [7]), it is likely to be degraded by NMD [8]. Many alternatively spliced isoforms are potential candidates of NMD based on this ‘50 nucleotide rule’ [9,10]. Thus, NMD-candidate transcripts are likely to have reduced function relative to the wild-type transcript, and their rate of occurrence (as a fraction of observed transcript forms) gives a simple measure of this negative selection pressure.

A major concern is the validity of this hypothesis for diploid genes that are present in two copies in each cell (one from each of the two copies of its chromosome, in a diploid genome). Even if a new mutation eliminated production of the original transcript form by its gene, in a heterozygote the wild-type copy of the gene would ensure that the original transcript form would still be produced at 50% of its original level (instead of 0%, as would be the case if there was only one copy of the gene). This might alleviate most of the negative effects of the mutation, so that alternative splicing of the mutation would not produce much additional relief of its negative selection pressure. This issue was not addressed in our original model [3] but is crucial for its evaluation.

Fortunately, this can be tested by measuring negative selection pressure against PTC-containing isoforms on diploid chromosomes (e.g. autosomes) versus those on haploid chromosomes (e.g. sex chromosomes). The human X chromosome is haploid in males and, due to X inactivation, its gene expression is typically limited to a single chromosome in female cells [11]. In this article, we analyze the frequency of PTCs in the canonical splice form for each gene (which we will refer to as the ‘major’ splice form) and in alternative splice forms for these genes (which we will refer to as the ‘minor’ splice forms), on both autosomal and X chromosomes in human and mouse.

Detecting alternatively spliced transcripts

We detected alternative splice forms for human and mouse by mapping mRNA and EST sequences onto genomic sequences as previously described [12], using the following data: (i) a download of UniGene EST data <ftp://ftp.ncbi.nih.gov/repository/UniGene/> (from January 2002) [13]; and (ii) human and mouse genome sequences downloaded from NCBI <ftp://ftp.ncbi.nih.gov/genomes/> (from January 2002). A database of alternatively spliced transcript isoforms from human and mouse was constructed from mRNA-EST-genomic multiple sequence alignments using our isoform generation algorithm described previously [10]. Major isoforms were characterized as isoforms with the largest number of ESTs for a given gene. Premature transcripts that are likely targets of NMD were identified by checking for a STOP codon located > 50 bp upstream of the last exon–exon junction site.

Analysis of PTCs in alternatively spliced transcripts

In our set of 13 384-transcript isoforms for 4422 human genes, we found that 3.7% of major transcript isoforms (165 out of 4422) had PTCs, similar to the percentage reported for human mRNAs [9,10]. However, 11.1% of the alternative-splicing (minor) isoforms had PTCs, approximately a threefold increase. We observed the same pattern in the mouse genome (3.6% for major isoforms versus 9.3% for minor isoforms; Table 1). These data indicate that alternative splicing is indeed associated with a substantial reduction in selection pressure against PTCs.

Analyzing the human data by chromosome, we found a large decrease in the frequency of PTCs on the X chromosome (3.5%) compared with autosomal chromosomes (8.9%) (Table 1). Chromosome X had the lowest PTC rate of all 23 chromosomes that were compared (Figure 1), and its difference versus the autosomal chromosomes was statistically significant ($P < 0.000006$). Moreover, we found the same result on the mouse X chromosome (2.9%) versus the mouse autosomal chromosomes (7.2%). Thus, diploidy also is associated with a significant reduction in selection pressure against PTCs, compared with haploid chromosomes (X). This pattern was observed even when we limited our analysis to major splice forms: in human, 1.2% had PTCs on the X chromosome compared with 3.8% on autosomal chromosomes. The Y chromosome

Table 1. Percentage of human and mouse isoforms with PTCs

Percentage of chromosomes with PTCs	Major isoforms	Minor isoforms	All isoforms
Human and mouse isoforms with PTCs			
Human	3.7%	11.1%	
Mouse	3.6%	9.3%	
Human isoforms with PTCs on autosomes and X chromosome			
Autosomes	3.8% (160/4221)	11.3% (975/8602)	8.9% (1135/12 823)
X chromosome	1.2% (2/161)	4.6% (14/299)	3.5% (16/460)
Mouse isoforms with PTCs on autosomes and X chromosome			
Autosomes	3.7% (32/871)	9.7% (121/1253)	7.2% (153/2124)
X chromosome	2.6% (1/39)	3.1% (2/64)	2.9% (3/103)

Abbreviation: PTCs, premature termination codons.

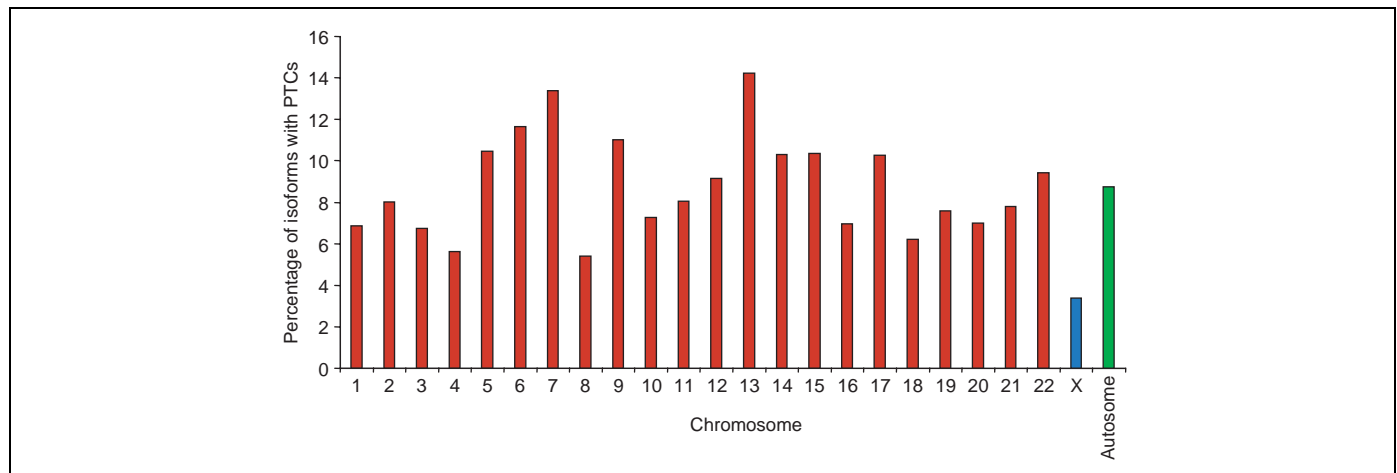


Figure 1. The percentage of isoforms with premature termination codons (PTCs) on individual human chromosomes. The percentages are given for each chromosome. The X chromosome has the lowest PTC rate of all 23 chromosomes compared here. Three and a half percent of transcript isoforms from the X chromosome have PTCs compared with 8.9% from autosomes. The autosomes are shown in red, the X chromosome is shown in blue and the average percentage for all human autosomes is shown in green.

has fewer genes and was not included in this comparison. However, we obtained similar results for the Y chromosome, none of the seven transcript isoforms we constructed from Y-chromosome genes have PTC.

Does alternative splicing still provide relief from negative selection even on diploid chromosomes? In human autosomes, the frequency of PTCs was 3.8% for major splice forms and 11.3% for minor splice forms (a threefold increase). In mouse autosomal chromosomes, the frequency of PTCs was 3.7% for major splice forms versus 9.7% for minor splice forms. However, the strength of this effect appears to be strongest on the X chromosome. On this chromosome, the frequency of PTCs was 1.2% for major splice forms versus 4.6% for minor alternative splice forms, almost a fourfold increase.

Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy

These data provide independent evidence for the hypothesis that alternative splicing can relieve negative selection pressure. Whereas the original evidence for this hypothesis was based on comparative genomics analysis of exon creation and/or loss in mammalian genomes [3], in this article, we have focused on a different phenomenon – the incidence of PTCs that are likely to be targets for NMD. Sorek *et al.* have reported that *Alu* sequences sometimes occur in exons but were always associated with alternative splicing (i.e. *Alu* sequences were only found in alternatively-spliced exons) [14]. This also indicates reduced negative selection pressure in alternatively spliced exons.

Alternative splicing and diploidy appear to produce approximately the same magnitude of reduction (three times) in negative selection pressure against PTCs during recent mammalian evolution. In both cases, the effect of ‘having another functional copy’ appears to increase tolerance for PTC isoforms by approximately threefold. The effects of alternative splicing and diploidy appear to be independent and additive. Alternative splicing still relieves negative selection pressure even on diploid chromosomes, but this effect appears to be stronger on haploid chromosomes (e.g. almost fourfold on the human X

chromosome). The combined effect of alternative splicing and diploidy yields a more than ninefold increase in tolerance for PTCs (1.2% in major splice forms on the X chromosome; 11.3% in minor splice forms on autosomes). The percentage of transcripts with PTCs is probably an underestimate because the transcript isoforms we analyzed are those that have escaped NMD, and because we used stringent criteria during the isoform generation process to filter out potential EST artifacts (e.g. retention of an entire intron) [10]. In this analysis, we have generally assumed that the occurrence of PTC (and thus the likelihood of NMD) can cause a negative impact on phenotype, namely, failure to produce a functional protein product. There is good evidence for this, but it has not yet been broadly demonstrated. Brenner and colleagues have also advanced the interesting hypothesis that NMD might constitute a functional form of regulation: that is, using the coupling of alternative splicing and NMD (instead of transcriptional control) to regulate the amount of protein product [9].

Concluding remarks

These data might reflect genomic evidence for an important role for NMD during mammalian evolution [15,16]. The increased incidence of NMD-candidate forms observed on diploid chromosomes is consistent with their function of degrading aberrant transcript forms that might produce dominant negative phenotypes. Because this function is useful for genes on diploid chromosomes (where the second copy of the chromosome can supply a working copy of the gene), but not on haploid chromosomes, this would be expected to yield more NMD-candidate forms on diploid chromosomes.

Acknowledgements

We thank D. Black, C. Grasso, Y. Marahrens, B. Modrek, A. Resch and Q. Xu for their discussions and comments on this work. C.J.L. was supported by NIMH and NINDS Grant MH65166 and by DOE grant DEFG0387ER60615.

References

- 1 Boue, S. *et al.* (2003) Alternative splicing and evolution. *BioEssays* 25, 1031–1034

- 2 Kan, Z. *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837–1845
- 3 Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* 34, 177–180
- 4 Nurtdinov, R.N. *et al.* (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* 12, 1313–1320
- 5 Li, W.H. *et al.* (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–174
- 6 Maquat, L.E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* 5, 89–99
- 7 Nagy, E. and Maquat, L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* 23, 198–199
- 8 Maquat, L.E. (2002) Nonsense-mediated mRNA decay. *Curr. Biol.* 12, R196–R197
- 9 Lewis, B.P. *et al.* (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U. S. A.* 100, 189–192
- 10 Xing, Y. *et al.* (2004) The multiassembly problem reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.* 14, 426–441
- 11 Heard, E. *et al.* (1997) X-chromosome inactivation in mammals. *Annu. Rev. Genet.* 31, 571–610
- 12 Modrek, B. *et al.* (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859
- 13 Boguski, M.S. *et al.* (1993) dbEST—database for expressed sequence tags. *Nat. Genet.* 4, 332–333
- 14 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067
- 15 Lynch, M. (2002) Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6118–6123
- 16 Lynch, M. and Kewalramani, A. (2003) Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol. Biol. Evol.* 20, 563–571

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.07.009

Conserved regulatory motifs in bacteria: riboswitches and beyond

Cei Abreu-Goodger, Nancy Ontiveros-Palacios, Ricardo Ciria and Enrique Merino

Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, 62210 Morelos, México

We present a computational approach that identifies regulatory elements conserved across phylogenetically distant organisms. Intergenic regulatory regions were clustered by orthology of the adjacent genes, and an iterative process was applied to search for significant motifs, enabling new elements of the putative regulon to be added in each cycle. With this approach, we identified highly conserved riboswitches and the Gram positive T-box. Interestingly, we identified many other regulatory systems that appear to depend on conserved RNA structures.

Comparative genomic approaches are central to analyzing the increasing number of whole-genome sequences. Although using this kind of analysis to find regulatory elements is not new, the focus has usually been on one genome or group of closely related genomes [1–3] because sequence conservation of functional intergenic regions (promoters, protein binding sites) is usually low, and quickly diverges.

It came as a surprise to many scientists when specific RNA ‘riboswitches’ were shown to be capable of regulating gene expression by directly sensing a metabolite without the intervention of a protein [4]. RNA riboswitches have since been shown to be involved in various metabolic

processes including thiamine, riboflavin, cobalamine, adenine, guanine and lysine biosynthesis [5–11]. We assumed that this type of regulatory sequence would be easily identified given their broad phylogenetic distribution and highly conserved nature.

Searching for interesting motifs

The starting point for our work is a set of orthologous regulatory regions. To obtain these we used the Cluster of Orthologous Groups (COG) of proteins database (<http://www.ncbi.nlm.nih.gov/COG/>) [12] together with operon predictions based on intergenic distances [13]. In this manner, every protein from 164 fully sequenced bacterial genomes that was associated with a COG was assigned to the intergenic minimal upstream region (iMUR) of the first gene of the predicted operon to which it belongs. To avoid over-representation of similar sequences from related genomes, redundant sequences were eliminated. We obtained ~4000 clusters of orthologous regulatory regions, each belonging to a different COG.

We used the public domain motif discovery tool Multiple EM for Motif Elicitation (MEME) [14] to find a set of over-represented ‘seed motifs’ for each COG (Figure 1a). These motifs were used to identify other members of the putative regulon by searching in all upstream regions using the MEME counterpart Motif Alignment and Search Tool (MAST) [15]. As a result of this

Corresponding author: Enrique Merino (merino@ibt.unam.mx).

Available online 19 August 2004