

Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms

Naama M Kopelman¹, Doron Lancet¹ & Itai Yanai^{1,2}

Gene duplication and alternative splicing are distinct evolutionary mechanisms that provide the raw material for new biological functions. We explored their relationships in human and mouse and found an inverse correlation between the size of a gene's family and its use of alternatively spliced isoforms. A cross-organism analysis suggests that selection for genome-wide genic proliferation might be interchangeably met by either evolutionary mechanism.

New gene functions are generally acquired through redundancy^{1–3}. This occurs through two mechanisms for the production of new genetic isoforms: an increase in transcript diversity^{4,5} or the emergence of additional genomic copies^{1,2}. These processes are each accompanied by divergence in spatial and temporal gene expression^{6,7}. Here we examine how the two modes of functional diversification are related to each other as evolutionary mechanisms.

We found an inverse correlation between the size of a gene's family and its propensity to produce alternative splice variants (Fig. 1). Genes of family size 1 (singletons) are more likely to use alternative splicing

than genes that have undergone duplication. Furthermore, genes with large families (more than ten members) use alternative splicing roughly half as frequently as do singletons.

As the age of a duplication may be in the range of >2,000 million years, it is useful to reduce confounding effects by examining duplications with reference to increasingly narrow time frames. We therefore delineated human gene families defined with respect to four model organisms (yeast, fruit fly, puffer fish and mouse) that have widely different times of divergence from humans. We found that the negative correlation was accentuated for more recent duplicates (Fig. 1). For example, the rate of alternative splicing of genes duplicated since the divergence from mouse (~90 million years ago, MYA; ref. 8) was two times lower than that of genes that have not duplicated. The inverse relationship between gene duplication and alternative splicing was independent of family-size binning (Supplementary Fig. 1 online), number of exons per gene (Supplementary Fig. 2 and Supplementary Table 1 online) and expressed-sequence tag coverage (Supplementary Fig. 3 online). This relationship was also apparent when gene families rather than genes were used as the units of analysis (Supplementary Fig. 4 online).

An extreme example of this trend can be seen in the ten largest human gene families in our analysis, defined relative to divergence from puffer fish. Of the 415 genes in these families for which we have splicing data, less than 3% use alternative splicing, compared with more than 50% for singletons. Comparing frequencies of genes with alternative splicing among duplicates and singletons, a χ^2 test showed significant depletion of alternative splicing ($P < 0.0001$) for nine of these families, with the exception of a family of genes with zinc-finger domains.

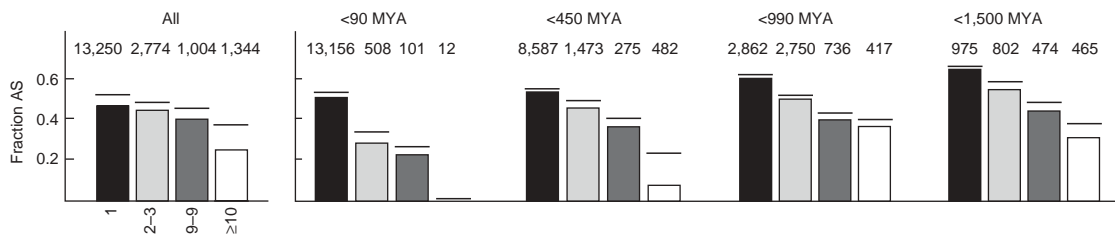


Figure 1 Inverse relationship between the size of a gene's family and production of alternatively spliced variants. Histogram bars indicate the fraction of genes containing more than one splice variant (or fraction with alternative splicing, AS), shown for singletons (1) and for small (2–3 members), medium (4–9 members) and large (≥ 10 members) gene families. 'All' shows all human gene families defined by single linkage clustering (Supplementary Methods online). Other panels show human gene families delineated with respect to another organism using Inparanoid¹⁴ such that the age of the duplications is estimated⁸ with reference to mouse (*Mus musculus*, <90 MYA), puffer fish (*Fugu rubripes*, <450 MYA), fruit fly (*Drosophila melanogaster*, <990 MYA) and yeast (*Saccharomyces cerevisiae*, <1,550 MYA). Horizontal lines above bars show results after removing intronless genes. The number of genes in each bin is indicated above the bars. Estimates of the number of splice variants per human gene were derived from ASD's AltSplice database¹⁵. Similar results are found using other gene family definitions (Supplementary Fig. 5 online) and splice-variant estimation method (Supplementary Fig. 6 online). Alternative splicing estimation and gene family delineation are further described in Supplementary Methods online.

¹Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. ²Present address: Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, 02138, USA. Correspondence should be addressed to I.Y. (yanai@mcb.harvard.edu).

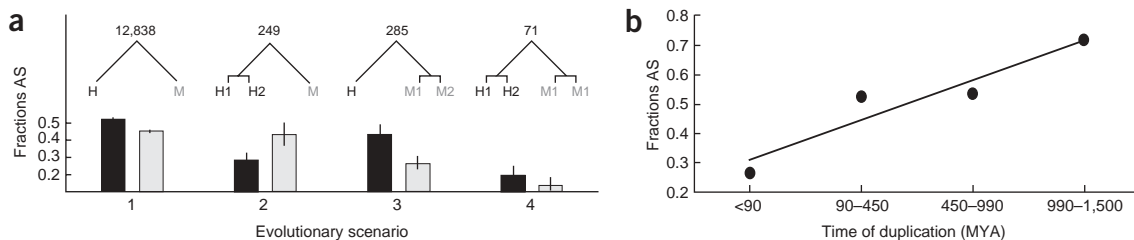


Figure 2 Interchangeability between alternative splicing and gene duplication. **(a)** Dark and light gray bars indicate the fraction of human and mouse genes, respectively, with alternative splicing (AS) under four evolutionary scenarios of gene duplication: 1, conserved singletons (*i.e.*, genes that have not duplicated in either mouse or human); 2, genes that have duplicated (at least once) in human but not in mouse; 3, genes that have duplicated in mouse but not in human; and 4, genes that have duplicated in both human and mouse. Error bars indicate 95% confidence intervals obtained by bootstrapping. **(b)** The fraction of alternative splicing (AS) is given for four sets of 'staged' gene duplications, fitted by a linear regression. For example, human duplicates whose duplication is inferred to have occurred before the human-mouse split but after the human-puffer fish split are placed in the 90–450 MYA bin. The method for inferring 'staged' duplications is described in **Supplementary Methods** online.

The existence of multiple copies of retrotransposed genes could make a substantial contribution to the inverse correlation between duplication and alternative splicing. Such intronless genes arise by reverse transcription of mRNA and hence lack alternative splicing. But we found that the observed relationship between gene duplication and alternative splicing was largely undisturbed by excluding all genes without introns (~10%; **Fig. 1**).

To determine what evolutionary forces underlie the inverse correlation, we carried out a detailed comparison of human and mouse orthologs under four gene duplication scenarios (**Fig. 2a**). As expected on the basis of our other analyses, the rate of alternative splicing was roughly three times higher in conserved singletons in human and mouse than in genes that have duplicated in both lineages (**Fig. 2a**). This relationship could reflect an inherent property of certain gene families to prefer proliferation through gene duplication as opposed to alternative splicing. But singletons in one mammalian species whose orthologs in the other have duplicated since divergence had a substantially higher rate of alternative splicing than did their duplicated orthologs (**Fig. 2a**). This finding implies that inherent family inclinations do not underlie the inverse correlation and suggests instead that demand for proliferation is fulfilled independently by either mechanism.

We next sought to distinguish between two opposite models consistent with our results (**Fig. 2a**): either duplicates lose splice variants or singletons acquire them. We examined the relevant genes with respect to organisms of increasing phylogenetic distance. If splice-variant loss prevailed, we would expect anciently duplicated genes to have a relatively small fraction of alternative splicing, having had more time available for such loss. But we detected a positive correlation between the age of a duplication event and the fraction of alternative splicing (**Fig. 2b**). This suggests that splice-variant acquisition, rather than loss, is a better explanation for our results. Despite the general trend, gene duplicates may often partition splice variants through subfunctionalization⁹, as previously documented for *mitf* genes in fish¹⁰.

Further support for the notion of evolutionary gain of splice variants is provided by the observation that alternative splicing frequently evolves through exon duplication^{11–13}. This suggests that acquiring splice variants is concomitant with an increase in the number of exons; in fact, we found that singletons had significantly more exons than duplicates. The inverse correlation (**Fig. 1**), however, was not explained by this exon bias (**Supplementary Fig. 2** and **Supplementary Table 1** online).

Our results have implications for understanding the molecular evolutionary mechanisms leading to genetic innovations. Gene duplication is one evolutionary pathway for the emergence of a new biological function. The fact that singletons whose orthologs have duplicated seem to be compensated with additional splice variants indicates that the pressure to proliferate precedes the duplication event, and is thus consistent with the idea that new function precedes gene duplication².

We show here that gene duplication and alternative splicing rates are not independent evolutionary properties of a gene. This idea is consistent with the suggestion⁵ that an alternatively spliced exon may serve as an 'internal paralog' of a gene. The inverse relationship can be explained by a balanced fulfillment of a requirement for diversification through either of the two mechanisms. Our work therefore justifies the viewing of gene duplication and alternative splicing in the same frame of reference when analyzing evolutionary trends of increasing gene complexity.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank A. Derti for a critical reading of the manuscript and Y. Groner, E. Domany, M.J. Lercher, A. Bar-Even and S.J. Fleishman for discussions. I.Y. is a Koshland Scholar and D.L. is the incumbent of the Ralph and Lois Silver Chair in Human Genomics. This research is supported by the Crown Human Genome Center, the Abraham and Judith Goldwasser Foundation and the Koshland Center for Basic Research.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 7 October 2004; accepted 13 April 2005

Published online at <http://www.nature.com/naturegenetics/>

- Ohno, S. *Evolution by Gene Duplication* (Springer, Berlin, 1970).
- Hughes, A.L. *Proc. R. Soc. Lond. B Biol. Sci.* **256**, 119–124 (1994).
- Lynch, M. & Conery, J.S. *Science* **290**, 1151–1155 (2000).
- Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. *Nat. Genet.* **30**, 29–30 (2002).
- Modrek, B. & Lee, C.J. *Nat. Genet.* **34**, 177–180 (2003).
- Johnson, J.M. *et al. Science* **302**, 2141–2144 (2003).
- Makova, K.D. & Li, W.H. *Genome Res.* **13**, 1638–1645 (2003).
- Hedges, S.B. *Nat. Rev. Genet.* **3**, 838–849 (2002).
- Force, A. *et al. Genetics* **151**, 1531–1545 (1999).
- Lister, J.A., Close, J. & Raible, D.W. *Dev. Biol.* **237**, 333–344 (2001).
- Kondrashov, F.A. & Koonin, E.V. *Hum. Mol. Genet.* **10**, 2661–2669 (2001).
- Letunic, I., Copley, R.R. & Bork, P. *Hum. Mol. Genet.* **11**, 1561–1567 (2002).
- Ast, G. *Nat. Rev. Genet.* **5**, 773–782 (2004).
- Remm, M., Storm, C.E.V. & Sonnhammer, E.L.L. *J. Mol. Biol.* **314**, 1041–1052 (2001).
- Thanaraj, T.A. *et al. Nucleic Acids Res.* **32**, D64–D69 (2004).