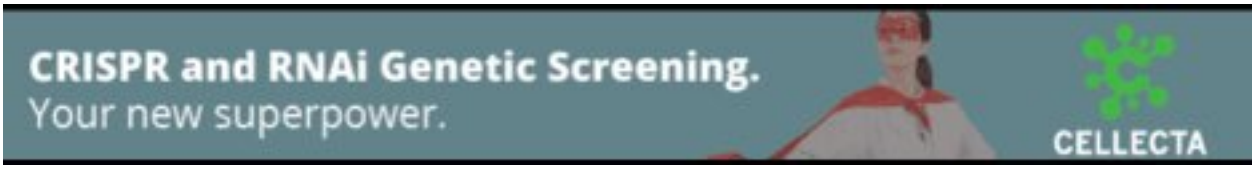
Gene length and expression level shape genomic novelties

Vladislav Grishkevich and Itai Yanai

Genome Res. 2014 24: 1497-1503 originally published online July 11, 2014
Access the most recent version at doi:10.1101/gr.169722.113



Gene length and expression level shape genomic novelties

Vladislav Grishkevich and Itai Yanai

Department of Biology, Technion–Israel Institute of Technology, Haifa 32000, Israel

Gene duplication and alternative splicing are important mechanisms in the production of genomic novelties. Previous work has shown that a gene's family size and the number of splice variants it produces are inversely related, although the underlying reason is not well understood. Here, we report that gene length and expression level together explain this relationship. We found that gene lengths correlate with both gene duplication and alternative splicing: Longer genes are less likely to produce duplicates and more likely to exhibit alternative splicing. We show that gene length is a dynamic property, increasing with evolutionary time—due in part to the insertions of transposable elements—and decreasing following partial gene duplications. However, gene length alone does not account for the relationship between alternative splicing and gene duplication. A gene's expression level appears both to impose a strong constraint on its length and to restrict gene duplications. Furthermore, high gene expression promotes alternative splicing, in particular for long genes, and alternatively, short genes with low expression levels have large gene families. Our analysis of the human and mouse genomes shows that gene length and expression level are primary genic properties that together account for the relationship between gene duplication and alternative splicing and bias the origin of novelties in the genome.

[Supplemental material is available for this article.]

The evolutionary lineage leading to metazoans is marked by a rise in organismal complexity evident in the sophistication of the cellular components and the emergence of new cell types (Gerhart and Kirschner 1997). This gain of complexity must have arisen through genomic changes, with the mechanisms of gene duplication and alternative splicing playing a major role (Lynch 2007). Both of these sources of evolutionary novelty allow for the appearance of functions without affecting preexisting ones, and thus, can contribute to the generation of variation required for the rise of complexity (Force et al. 1999; Johnson et al. 2003; Makova and Li 2003; Koonin and Wolf 2010). For example, gene duplications have led to the proliferation of transcription factor families, such as the HOX homeodomain gene family, which underlie the sophisticated developmental patterning of the organism (McGinnis et al. 1984; Garcia-Fernandez 2005; Duboule 2007), and alternative splicing contributes to the explosion in the transcriptomic potential from a limited gene complement (Graveley 2001; Chothia et al. 2003).

Previous work has suggested that these two seemingly unrelated mechanisms may actually be coupled: A negative correlation was detected between a gene's family size and its number of splice isoforms (Kopelman et al. 2005). Originally, it was proposed that splice variants are subfunctionalized between duplicates (Kopelman et al. 2005; Su et al. 2006). Support for this model came with the observation of asymmetric partitioning of the splice variants (Su et al. 2006). A more detailed study on the functional determinants of the variants, though, did not find evidence for subfunctionalization (Talavera et al. 2007). Furthermore, a recent work proposed that the relationship follows from a bias for duplication in genes depleted in splice variants and the subsequent gain of splice variants with time (Roux and Robinson-Rechavi 2011), though this view was challenged (Su and Gu 2012). Here we explored the possibility that gene duplication and alternative splicing are a consequence of the interaction between two more basic genic properties: gene length and expression level.

Between *E. coli* and humans the average gene length differs by nearly two orders of magnitude (Lynch 2007). The lengths of the human protein-coding genes are widely distributed (Fig. 1A) and range from a few hundred bases up to a few million: The keratin associated protein 12-4 gene (*KRTAP12-4*) is 447 nucleotides long, whereas the *CNTNAP2* gene has 2,304,637 nucleotides. The extraordinary length of some genes is enabled by the presence of introns, which, as a substrate for insertion of transposable elements, provides a mechanism for the generation of transcriptional diversity through alternative combinations of exons. While different models have been proposed to explain the emergence of introns in eukaryotes, what is well evidenced is that their proliferation is consistent with drift in relatively small populations (Lynch and Conery 2003). Genetic drift also likely facilitated the proliferation of gene families by gene duplication. Indeed, the number of genes in single-cellular eukaryotic genomes is substantially higher than that in prokaryotic genomes, and higher still in multicellular eukaryotes. However, the increase in the gene number is not distributed uniformly among the different gene families. Katju and Lynch (2006) demonstrated that the average size of the duplicated DNA regions in the *C. elegans* genome is shorter than the average gene length. In fact, from their analysis of ancestral and duplicated paralog pairs they were able to determine the completeness of the duplication events in only 40% of observed copies. As we explore in this work, such a bias for duplications of a particular length in principle favors the duplication of shorter genes.

Similar to gene length, expression levels are also distributed with a wide range across the genes of the organism. Interestingly, a strong correspondence was reported between the level of a gene's expression and the length of its introns (Castillo-Davis

Corresponding author: yanai@technion.ac.il
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.169722.113.

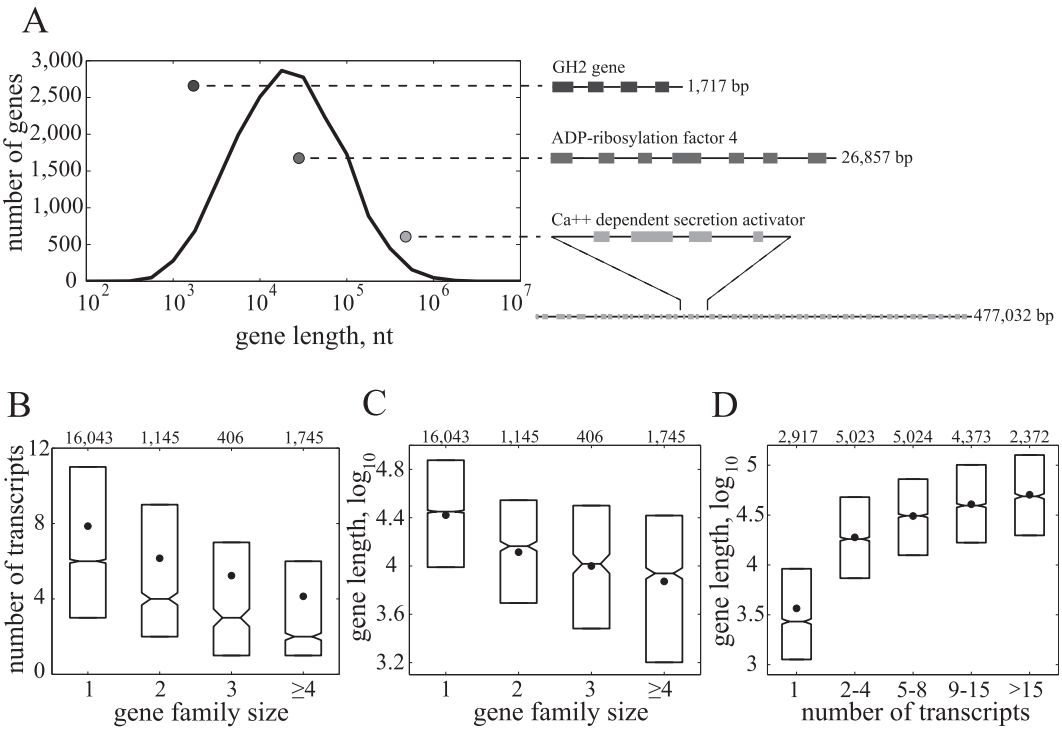


Figure 1. The correlation between gene duplication and alternative splicing is influenced by gene length. (*A*) Histogram of the lengths of human genes, computed as the number of transcribed bases, prior to splicing. Three particular genes are indicated as examples of short (black), medium (dark gray), and long (light gray) genes (boxes indicate exons). (*B*) An inverse relationship between a gene's number of alternatively spliced variants and its family size, as previously characterized (Kopelman et al. 2005). The box plot indicates the median, mean (circle), and 25% and 75% quartiles. *Above* each box plot the number of genes in the set is shown. (*C*) Gene family size is negatively correlated with gene length. (*D*) The number of alternative transcripts is positively correlated with gene length.

et al. 2002). Highly expressed genes tend to have shorter introns, likely reflecting their adaptation for high transcription efficiency. In another work, gene length was examined in a set of constitutively expressed genes (Chiaromonte et al. 2003). The investigators found that genes with the highest level of expression tended to be those that were also the shortest in length. These findings demonstrate that a requirement for high expression imposes a considerable constraint on gene length. Similar to gene length, the level of gene expression also has a direct effect on gene duplication. For example, lowly expressed genes, most of which are nonessential for the organism, have a significantly increased duplication rate (Woods et al. 2013). In contrast, gene duplication also affects gene expression level, as it has been reported that gene duplicates have lower levels of gene expression (Qian et al. 2010).

Here we report evidence that gene length and gene expression level are associated with both a gene's family size and its number of splice variants. Together, gene length and gene expression level can account for the relationship between gene duplication and alternative splicing. Our work suggests that the evolution of novelty is thus intrinsically dependent upon gene length and gene expression.

Results

Gene length is correlated with both gene duplication and alternative splicing

A gene's length—defined as the number of bases in the primary transcript prior to splicing—ranges over four orders of magnitude in the human genome (Fig. 1A). We asked whether the previously reported inverse relationship between alternative splicing and gene duplication (Kopelman et al. 2005; Su et al. 2006) is an indirect consequence of their correlation with gene length. Figure 1B recapitulates the gene duplication and alternative splicing relationship, showing that a single copy gene has on average 7.9 unique transcripts, while genes in large families (four or more copies) have on average 4.1 transcripts. Querying for a relationship with gene length, we found that genes present in multiple copies tend to be shorter in length than those from small gene families (Fig. 1C). Short genes also tend to have fewer splice variants (Fig. 1D). Thus, a gene's length may be intimately related to how it generates novelty either by additional copies or by alternative splice variants. To further examine this relationship we inquired into the mechanisms by which gene lengths change.

Gene length generally increases in evolutionary time, due in part to transposable elements

To further study gene lengths, we studied the gene's evolutionary dynamics across a range of age categories. To estimate a gene's age, we invoked the phylostratigraphy approach which examines the phyletic distribution of a gene's orthologs (Domazet-Loso and Tautz 2010). For example, the human gene *AASS* (aminoadipate-semialdehyde synthase) is annotated to the "Cellular" age category because orthologs of it can be detected in archaea, eubacteria, and eukaryotes; while the gene *TMEM167B* (transmembrane protein 167B) is annotated to the "Primates" age category because its orthologs are restricted to primates. Examining the distribution gene lengths across the gene age categories, we found that older genes are longer (Fig. 2A) ($R = 0.3$; P-value $< 10^{-250}$; $N = 12{,}939$), suggesting that genes increase in length over evolutionary time. This plot also suggested that the most recent human genes are longer than expected, indicating a potential acceleration of gene length evolution in this age category.

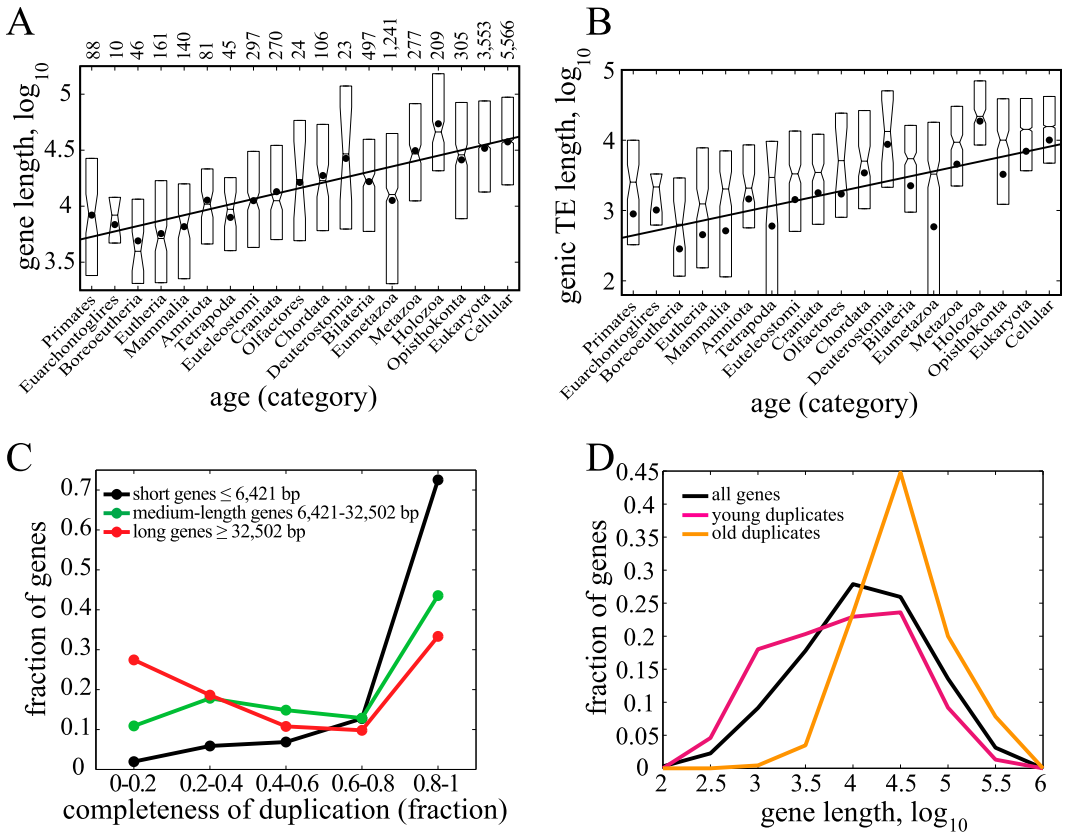


Figure 2. Dynamics of gene length in evolutionary time. (*A*) Gene length box plots for different gene age categories show that older genes are longer, suggesting that genes increase in length with time. The line is based on the least square fit of the mean. (*B*) The length of genic transposable elements also correlates with gene age (shown in the same format as *A*). (*C*) Truncation of gene lengths following incomplete gene duplication. Young duplicates (duplicated since separation of primates) of a family size of two were categorized into three equally sized gene length categories (short, medium, and long). The distributions show the ratio of the gene lengths of the paired duplicates (shorter gene length/longer). The set of the lengths of the longer genes in each pair was used to set the length categories. (*D*) Preferential duplications of shorter genes. The distributions of gene lengths for young duplicates (241 gene pairs), old duplicates (duplicated prior to the formation of Chordates, 234 gene pairs), and all human genes. The distribution is of the lengths of the longer gene of each pair.

One possible mechanism for the increase in gene length over time is the insertion of transposable elements (TEs). Indeed it has been previously noted that TEs are enriched for introns in the human genome: While protein-coding genes comprise only 24% of the human genome, 60% of the TEs map to these genomic regions (Sela et al. 2007). We thus asked whether TEs (SINE, LINE, LTR, and DNA repeats) in the human genome are more numerous in older genes. For each gene in each age category, we quantified the contribution of TEs to the gene length. We found that older genes have more TEs than younger genes (Fig. 2B), suggesting that the insertion of TEs contributes to the lengthening of genes over evolutionary time. Moreover, we also observed that older genes have a higher fraction of TEs than more recent genes (Supplemental Fig. S1). From this set of analyses, we concluded that there is a general bias in the genome for genes to increase in length over evolutionary time-scales, due in part to the insertion of TEs.

Gene duplications tend to reset gene lengths and are biased for short genes

Gene length may also be a dynamic genic property due to the effect of gene duplications. The observation that shorter genes have larger gene families may follow from the partial nature of duplications (Katju and Lynch 2006). To test this we asked whether there is evidence that longer genes produce duplicate genes that are more incomplete than shorter genes. Examining recently duplicated genes, we computed for each pair of genes the ratio in their lengths, relative to the longer gene of the pair. We found that for longer genes this ratio tends to be smaller, suggesting that duplications are increasingly incomplete for longer genes (Fig. 2C). We further characterize this relationship here by showing that longer genes have increasingly partial duplications which may thus explain why recent duplicates are shorter.

Partial duplication may also constrain which genes can successfully duplicate. To test for this, we examined the lengths of the recently duplicated genes. As a proxy for the lengths of genes prior to duplication, we again examined the longer gene of paired duplicates. We found that genes that have recently duplicated are shorter than the average gene ($P < 10^{-5}$, Kolmogorov-Smirnov test) (Fig. 2D). In contrast, genes that have not duplicated since the inception of the *Chordate* phylum are longer than average, suggesting a bias for the fixation of duplicates of shorter genes. This is likely due to the relative lack of survival of the partial duplications of longer genes. Collectively, the two processes of partial duplication of longer genes and higher survival rates of gene duplications of shorter genes account for the decreased average length of recent duplicates.

Gene expression level correlates with gene duplication and alternative splicing

We detected important outliers to the relationships of gene length with both gene duplication and alternative splicing. One characteristic property that distinguished these outlier genes was their high level of expression (Supplemental Fig. S2A–D). For example, the highly expressed ribosomal genes are short, yet have high levels of alternative splicing and few gene duplicates (Supplemental Fig. S2E–H). A plausible hypothesis is that due to their high levels of expression, many splice forms are generated. Indeed, a gene's expression level strongly correlates with its number of splice forms (Fig. 3A). Highly expressed genes also tend to be essential, and consequently also less prone to duplication (Woods et al. 2013). Thus, similarly to gene length, gene expression level is inversely correlated with both gene duplication and alternative splicing. However, gene length and gene expression level themselves have a complicated relationship (Fig. 3B). Overall, the two are significantly correlated ($R = -0.14$; $P < 10^{-33}$), but genes with a mid-range level of expression do not have expression levels that correlate with gene length ($P = 0.05$) (Fig. 3B). This suggests that while high expression imposes a constraint on the growth of a gene's length—resulting in short genes—longer gene lengths similarly restrict gene expression levels.

To further explore how gene length and expression level together impose an effect on gene duplication and alternative splicing, we defined 10 equally populated categories of genes according to gene lengths and 10 equally populated categories of genes according to expression levels. For genes present in each combination of gene length and expression level bins we computed the average gene family size and average number of transcripts (Fig. 3C,D). We found that highly expressed genes have very low rates of duplication (Fig. 3C, top row). This reflects the strong negative selection against gene duplication imposed by gene expression level. With medium to low expression levels, gene duplication is biased toward short genes. This indicates that in the relative absence of a bias against the duplication of highly expressed genes, duplications are enriched for short genes. Examining alternative splicing in the same analysis framework, it is apparent that the highest levels of alternative splicing occur for genes that are both highly expressed and long in their gene length (Fig. 3D).

Gene length and expression level together account for the relationship between gene duplication and alternative splicing

The synergy between gene length and expression level relates, at one extreme, to the generation of genic novelties in the form of gene duplications, and on the other to an increase in isoforms by alternative splicing. We thus asked whether gene length and expression level together account for the observed relationship between these two mechanisms. Figure 4A maps the relationships among these four main parameters. We first quantified the partial correlation between gene duplication and alternative splicing when controlling for gene length (see Methods). We found that the correlation is significantly weakened (Fig. 4C), though it still remained significant. Other genic factors such as chromosome position, chromosome number, strand, protein isoelectric point,

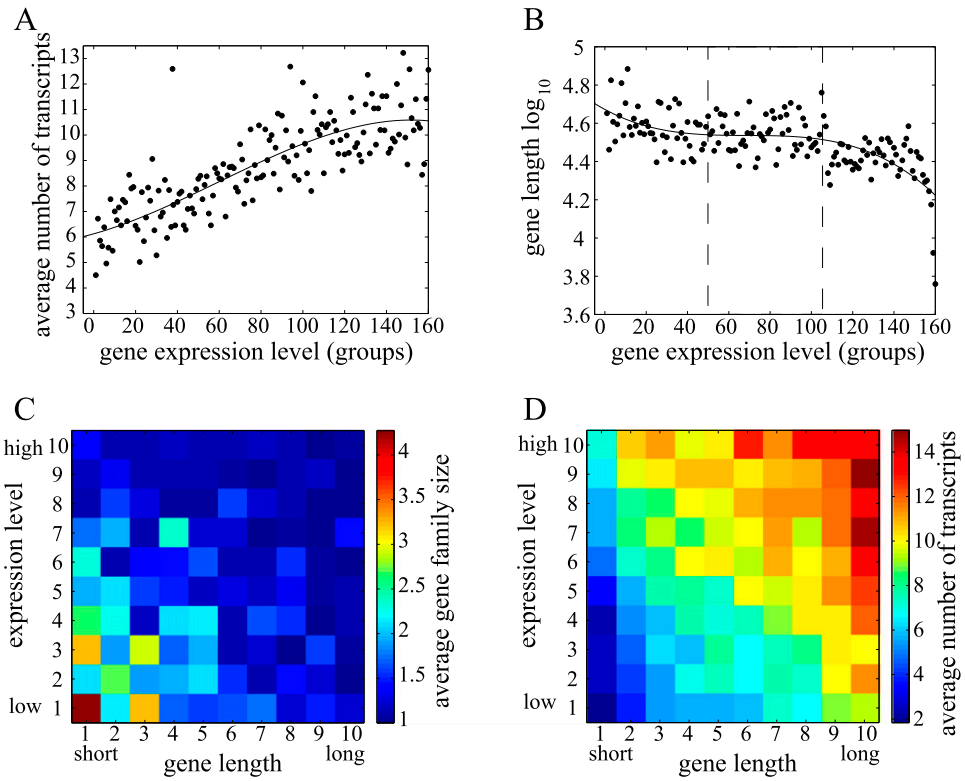


Figure 3. Gene length and expression level as primary genic properties. (*A*) Expression level is positively correlated with the number of transcripts. Genes were binned into 160 expression groups (each with 50 genes) and the average number of transcripts is shown for each expression group. (*B*) Gene length is generally negatively correlated with expression level, but not for mid-range expression level (bounded by dashed lines, see text). (*C*,*D*) The landscape of gene length and expression show inverse patterns of gene family sizes and alternative splicing. For each combination of gene length and expression level bins, the average number of gene family size (*C*) and the average number of transcripts (*D*) is shown.

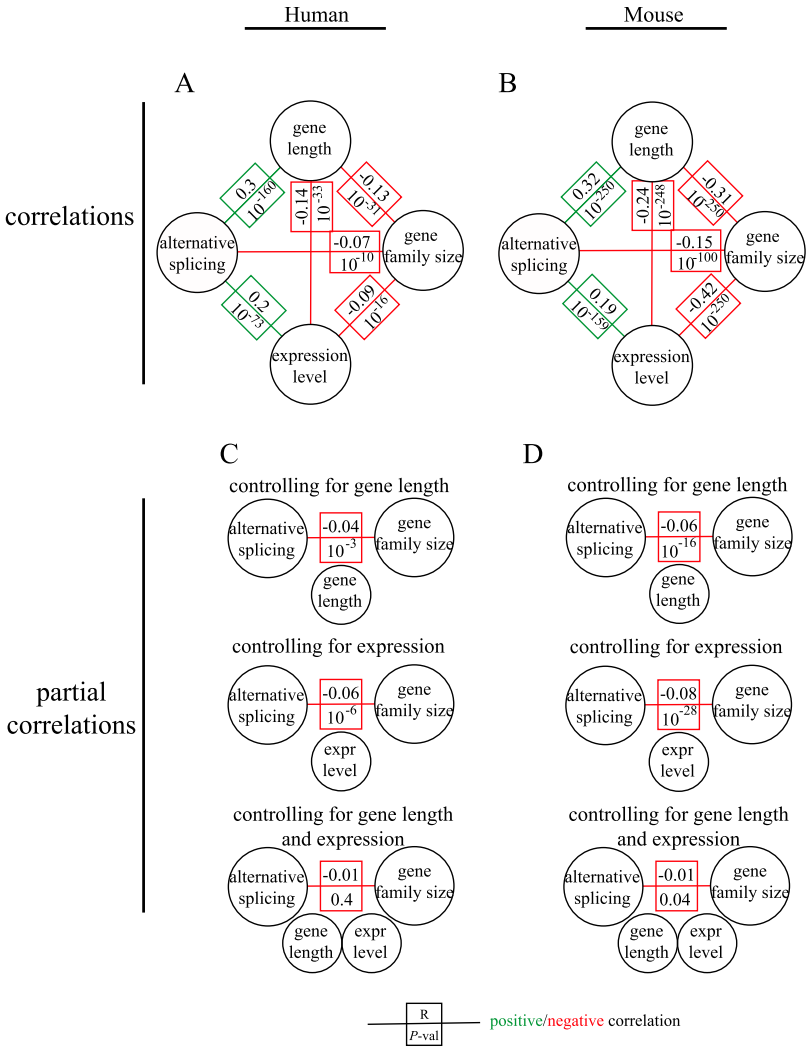


Figure 4. The correlation between alternative splicing and gene duplication is diminished in both human and mouse when controlling for gene length and expression level. (*A*,*B*) Pairwise correlations between gene length, expression level, gene duplication, and alternative splicing in human (*A*) and mouse (*B*). Positive and negative correlations are indicated by green and red lines, respectively. The numbers *above* and *below* each line indicate the correlation coefficient and *P*-value of the significance, respectively. (*C*,*D*) Correlations between alternative splicing and gene duplication when controlling for gene length, expression level, and both, for human (*C*) and mouse (*D*).

and codon adaptation index (CAI) did not weaken the relationship (Supplemental Table S1). We next repeated this test when controlling for expression level, and again found a diminished correlation (Fig. 4C). Finally, controlling for both gene length and expression level, the correlation between gene duplication and alternative splicing is reduced to marginal significance (Fig. 4C) ($P = 0.04$). This suggests that gene length and expression level together dominate the relationship between gene duplication and alternative splicing, since together they weaken the gene duplication and alternative splicing correlation more than each does so individually.

Finally, we tested whether the same relationships among the four gene properties studied here hold for the mouse genome. As in human, we found a negative correlation between gene duplication and alternative splicing. The five other relationships among the gene properties were also observed in mouse as they were found in human (Fig. 4B). In fact, in mouse these relationships were all stronger and more significant, perhaps because the mouse genome is better annotated (Fig. 4B). Again, as in human, when controlling for gene length the gene duplication and alternative splicing relationship was weakened; and it became nonsignificant when also controlling for gene expression level (Fig. 4D). This conserved pattern suggests that organisms with alternative splicing and gene duplications are generally influenced in their generation of genomic novelties by gene lengths and expression level.

Discussion

In this study we explored the manner by which two genic properties—gene length and gene expression level—shape the pattern of genomic novelty. We found that both are positively correlated with alternative splicing and negatively correlated with gene duplication, and thus together exert an influence on the path of generation of novel genic isoforms. Highly expressed long genes are likely to have many splice variants but only a few gene duplicates; in contrast, lowly expressed short genes will have few splice variants but many gene duplicates. We thus propose that gene length and gene expression levels are the "primary genic properties" that underlie the relationship between gene duplication and alternative splicing. While the correlations among these properties are not strong—the highest being −0.42 between gene duplication and expression level in mouse—we note that it is significant that they are at all observed given the many additional factors that operate in the complex process of genome evolution (Wolf et al. 2006). For example, our model neglects whole-genome duplication where longer genes will be completely duplicated without bias, thus adding noise to the relationships studied here. In this section we discuss the dynamic nature of gene length in evolutionary time, the relationship between expression level and alternative splicing, the notion of primary genic properties, and additional implications of our findings.

We presented evidence that gene length is a dynamic property (Fig. 2). The intronic nature of eukaryotic genes allows for a simple mechanism for gene growth and shrinkage, as individual introns can be lengthened with tolerable effects on gene functionality. In addition, the typically small population sizes of eukaryotes result in a weaker selection regime for the maintenance of a short gene length (Lynch and Conery 2003). Due to these two characteristics, gene length may be seen as a dynamic genic property and one that is under the influence of genetic drift. Our comparison of genes of different age categories revealed that such drift is biased toward an overall increase in gene length and that transposable elements are an important factor in this process. This

increase in length comes with implications for alternative splicing and gene duplications. With an increase in length comes the addition of splice variants (Fig. 1D), leading to potentially new functions. In contrast, the increase in length comes at a cost, because long genes are less likely to be successfully duplicated (Fig. 2C,D). Duplications of long genes tend to be of partial length, thereby resetting the length and allowing for the generation of other genic novelty as the truncated gene would inevitably lengthen differently. Katju and Lynch (2006) have provided evidence for the immediate acquisition of functions of partial gene duplicates. While this may seem to provide a mechanism for enhanced rather than constrained evolution, we propose that the overall set of partial duplicates generally leads to nonfunctional transcripts and consequently would not be under positive selection. Thus, in comparison with complete duplications of short genes, longer genes—as a group—would be biased against successful duplication.

Over time, the integration of TEs and other insertions are likely to lead to extensive exon shuffling and the generation of novel exons (Sorek et al. 2002; Lev-Maor et al. 2003; Morgante et al. 2005; Mersch et al. 2007). Furthermore, recent work has provided evidence that intron length is linked to the exonization of TEs in mammals (Sela et al. 2010). Consistently, a gene's length is strongly correlated with the number of exons (Kopelman et al. 2005). The accretion of exons has consequences for the repertoire of splice forms generated by the gene, as genes with more exons were observed to generate more splice forms (Castillo-Davis et al. 2002). Collectively, these analyses indicate that genes increase in length over time—with the introduction of TEs as one mechanism—and that this increase leads to additional splice variants. Indeed this provides one explanation for the observation that genes accumulate splice forms over time (Kopelman et al. 2005; Roux and Robinson-Rechavi 2011).

We found that gene length is not the only determinant of splicing and duplication because the relationship between these two remains after controlling for gene length (Fig. 4B). For example, many ribosomal genes are both short but have many splice variants (Supplemental Fig. S2E–H). The genetic organization of highly expressed genes suggests that they experience strong purifying selection, as can be seen by their biases for common codons (Tuller et al. 2010), short introns (Castillo-Davis et al. 2002), and essentiality (Woods et al. 2013). We thus recognized high expression level as a secondary determinant of the genic potential for alternative splicing and gene duplication. First, highly expressed genes exhibit many splice variants (Fig. 3) that likely result from errors of the transcriptional machinery. In contrast, lowly expressed genes may have few annotated splice variants simply because errors in these are rarer given their low expression. We cannot, however, exclude the possibility that the causal link is actually the inverse: that the necessity for the various splice variants of a gene leads to its high expression level. Under the former explanation, it is easier to understand why most alternative splice variants are not conserved between human and mouse since errors need not be identical (Nurtdinov et al. 2003). Also, many highly expressed genes such as ribosomal genes have many splice variants yet are ubiquitously expressed, thereby obfuscating the need for tissue-specific alternative splice variants. Finally, in a framework where high expression leads to a high number of splice variants, Roux and Robinson-Rechavi's suggestion that a gene with many splice variants is less likely to be duplicated (Roux and Robinson-Rechavi 2011) can be rationalized as follows. A highly expressed gene may be less likely to duplicate—irrespective of its splice variants—because such genes tend to be essential and this would preclude the possibility of a nondisruptive duplicate. Essential genes tend to be highly expressed (Castillo-Davis et al. 2002) and thus the link with alternative splicing is secondary.

In our view, gene duplication and alternative splicing can be considered as "output" parameters, as they are the consequence of these primary properties. The evidence for this notion comes from our demonstration that the reported relationship (Kopelman et al. 2005) is not significant when controlling for gene length and expression level (Fig. 4). Thus, the previously characterized relationship between these latter two properties is not direct. While it was originally suggested that gene duplications yielding a big gene family may be interchangeable with the acquisitions of many splice variants (Kopelman et al. 2005), our results suggest that the correlation follows from the opposing relationships these have with length and expression. Furthermore, a proposed asymmetric partitioning of splice variants among duplicates (Su et al. 2006) may be attributed to partial gene duplications. As we have examined here only the human and mouse genomes, further work is required to address the universality of these results.

Our results imply that the primary genic properties define the course by which new isoforms may be generated, by either gene duplication or alternative splicing. Neutral aspects such as long lengths can eventually yield novel genes by partial duplicates. Similarly, splice variants generated by high expression level may lead to useful splice variants that may then be genetically assimilated. Thus, in the generation of genomic novelty, primary genic properties such as gene length and expression level may lead to surprising emergent properties and shape the course of the novelties available to the agency of natural selection over evolutionary time-scales.

Methods

Gene properties

The set of duplicate genes and their time of origin were retrieved from Ensembl (Flicek et al. 2013); genome release GRCh37.p10 for human and GRCm38.p2 for mouse. Unless otherwise indicated, only duplicates occurring since the appearance of Chordates are included in the analyses. A gene's family size corresponds to the number of genes in its associated Ensembl family. A gene's length and its number of splice variants were also retrieved from Ensembl. Gene age categories were obtained from a previous study that introduced phylostratigraphic analysis to establish gene age (Domazet-Loso and Tautz 2010).

Transposable elements (TE)

Genes and TE coordinates were retrieved from the genes and RepeatMasker tracks of the UCSC Genome Browser (Meyer et al. 2013). The contribution of TEs to the gene length was then calculated as the combined length of the TEs that colocalized within genic regions. Only the contribution of the four major classes of TEs (SINE, LINE, LTR, and DNA repeats) was analyzed.

Expression level

Human gene expression data were obtained from an atlas consisting of 5372 samples representing 369 different cell and tissue types, disease states, and cell lines (Lukk et al. 2010). Disease state samples were excluded from analysis. The 7986 genes with expression data in at least one sample were used for expression analysis. Mouse expression data spanning 19 different tissues were downloaded from the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE29184 (Shen et al.

2012), and the 18,929 genes with expression data in at least one sample were used for expression analysis. For both data sets, expression level for each gene was defined as its average expression across all tissues and cell types.

Statistics

Partial correlation coefficients between gene family size and number of splice variants controlling for gene properties was computed using the MATLAB function "partialcorr" from the "Statistics Toolbox." Distributions were compared in Figure 2 using the Kolmogorov-Smirnov test.

Acknowledgments

We acknowledge helpful advice from Yael Mandel-Gutfreund, Itai Sharon, Gal Avital, David Silver, and Leon Anavy. This work was supported by an ERC grant from the EU and the EMBO Young Investigator Program.

References

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet* **31:** 415–418.

Chiaromonte F, Miller W, Bouhassira EE. 2003. Gene length and proximity to neighbors affect genome-wide expression levels. *Genome Res* **13:** 2602–2608.

Chothia C, Gough J, Vogel C, Teichmann SA. 2003. Evolution of the protein repertoire. *Science* **300:** 1701–1703.

Domazet-Loso T, Tautz D. 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol* **8:** 66.

Duboule D. 2007. The rise and fall of Hox gene clusters. *Development* **134:** 2549–2560.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.

Garcia-Fernandez J. 2005. The genesis and evolution of homeobox gene clusters. *Nat Rev Genet* **6:** 881–892.

Gerhart J, Kirschner M. 1997. *Cells, embryos, and evolution: toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability.* Blackwell Science, Malden, MA.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17:** 100–107.

Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141–2144.

Katju V, Lynch M. 2006. On the formation of novel genes by duplication in the *Caenorhabditis elegans* genome. *Mol Biol Evol* **23:** 1056–1067.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet* **11:** 487–498.

Kopelman NM, Lancet D, Yanai I. 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37:** 588–589.

Lev-Maor G, Sorek R, Shomron N, Ast G. 2003. The birth of an alternatively spliced exon: 3′ splice-site selection in *Alu* exons. *Science* **300:** 1288–1291.

Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A. 2010. A global map of human gene expression. *Nat Biotechnol* **28:** 322–324.

Lynch M. 2007. *The origins of genome architecture*. Sinauer Associates, Sunderland, MA.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302:** 1401–1404.

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* **13:** 1638–1645.

McGinnis W, Garber RL, Wirz J, Kuroiwa A, Gehring WJ. 1984. A homologous protein-coding sequence in *Drosophila* homeotic genes and its conservation in other metazoans. *Cell* **37:** 403–408.

Mersch B, Sela N, Ast G, Suhai S, Hotz-Wagenblatt A. 2007. SERpredict: detection of tissue- or tumor-specific isoforms generated through exonization of transposable elements. *BMC Genet* **8:** 78.

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2013. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41:** D64–D69.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37:** 997–1002.

Nurtdinov RN, Artamonova II, Mironov AA, Gelfand MS. 2003. Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum Mol Genet* **12:** 1313–1320.

Qian W, Liao BY, Chang AY, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet* **26:** 425–430.

Roux J, Robinson-Rechavi M. 2011. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res* **21:** 357–363.

Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. 2007. Comparative analysis of transposed element insertion within human and mouse genomes reveals *Alu*'s unique role in shaping the human transcriptome. *Genome Biol* **8:** R127.

Sela N, Kim E, Ast G. 2010. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biol* **11:** R59.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. 2012. A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488:** 116–120.

Sorek R, Ast G, Graur D. 2002. *Alu*-containing exons are alternatively spliced. *Genome Res* **12:** 1060–1067.

Su Z, Gu X. 2012. Revisit on the evolutionary relationship between alternative splicing and gene duplication. *Gene* **504:** 102–106.

Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Res* **16:** 182–189.

Talavera D, Vogel C, Orozco M, Teichmann SA, de la Cruz X. 2007. The (in)dependence of alternative splicing and gene duplication. *PLoS Comput Biol* **3:** e33.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141:** 344–354.

Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc Biol Sci* **273:** 1507–1515.

Woods S, Coghlan A, Rivers D, Warnecke T, Jeffries SJ, Kwon T, Rogers A, Hurst LD, Ahringer J. 2013. Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet* **9:** e1003330.

Received November 18, 2013; accepted in revised form June 23, 2014.