Intron Size, Abundance, and Distribution within Untranslated Regions of Genes

Xin Hong, Douglas G. Scofield, and Michael Lynch

Department of Biology, Indiana University

Most research concerning the evolution of introns has largely considered introns within coding sequences (CDSs), without regard for introns located within untranslated regions (UTRs) of genes. Here, we directly determined intron size, abundance, and distribution in UTRs of genes using full-length cDNA libraries and complete genome sequences for four species, *Arabidopsis thaliana*, *Drosophila melanogaster*, human, and mouse. Overall intron occupancy (introns/exon kbp) is lower in 5′ UTRs than CDSs, but intron density (intron occupancy in regions containing introns) tends to be higher in 5′ UTRs than in CDSs. Introns in 5′ UTRs are roughly twice as large as introns in CDSs, and there is a sharp drop in intron size at the 5′ UTR-CDS boundary. We propose a mechanistic explanation for the existence of selection for larger intron size in 5′ UTRs, and outline several implications of this hypothesis. We found introns to be randomly distributed within 5′ UTRs, so long as a minimum required exon size was assumed. Introns in 3′ UTRs were much less abundant than in 5′ UTRs. Though this was expected for human and mouse that have intron-dependent nonsense-mediated decay (NMD) pathways that discourage the presence of introns within the 3′ UTR, it was also true for *A. thaliana* and *D. melanogaster*, which may lack intron-dependent NMD. Our findings have several implications for theories of intron evolution and genome evolution in general.

Introduction

Ever since the unexpected discovery of introns (Berget et al. 1977; Chow et al. 1977; Evans et al. 1977; Goldberg et al. 1977), there has been intense debate about their origins, stability, and adaptive significance. Much early attention to the evolution of introns focused on the timing of their origins. The introns-early or exon theory of genes proposes that introns are ancient, and that early diversification of genes in the progenote, the genome ancestral to all prokaryotes and eukaryotes, was greatly accelerated by the shuffling of exons at intron-induced boundaries (Blake 1978; Gilbert 1978, 1987). The subsequent loss of introns in prokaryotes alone then occurred through selection for more streamlined genes and genomes (Doolittle 1978; Darnell and Doolittle 1986; Senapathy 1986; Roy and Gilbert 2005). The introns-late hypothesis maintains that introns appeared later and at random in early eukaryotic genomes, and that any adaptive role in gene evolution was gained following insertion (Orgel and Crick 1980; Cavalier-Smith 1985; Palmer and Logsdon 1991; Frugoli et al. 1998). More recent theory has emphasized mutational and population-genetic processes that are likely to govern the establishment and retention of introns (Lynch 2002; Lynch and Richardson 2002; Lynch and Kewalramani 2003). A so-called "synthetic" introns-early theory proposes the coexistence of ancient introns situated to promote exon shuffling, with more recently gained introns conforming to introns-late expectations (Fedorov et al. 2001; de Souza 2003; Fedorova and Fedorov 2003).

Hypotheses addressing the abundance and locations of introns within the protein-coding sequence (CDS) of transcripts have figured prominently in early as well as more recent theories of intron establishment, maintenance and proliferation. In the introns-late view, initial intron positions are largely random within transcripts or occur at "proto-splice sites" that carry short sequences similar to conserved exon sequences flanking introns but that are otherwise context free (Cavalier-Smith 1991; Dibb 1991; Cho and Doolittle 1997). Intron positions may thus be purely fortuitous or involve selection for features that influence transcription or translation, for example, regulation of transcription initiation (Fong and Zhou 2001; Le Hir et al. 2003), efficiency of mRNA export (Luo and Reed 1999), enhancement of splicing efficiency (Berget 1995; Nissim-Rafinia and Kerem 2002), chromatin assembly (Lauderdale and Stein 1992; Liu et al. 1995), and recognition of premature termination codons via nonsense-mediated decay (NMD) (Kim et al. 2001; Lynch and Kewalramani 2003; Maquat 2004a). Under the introns-early view, ancient genes and exons consisted of relatively short polypeptide sequences of limited secondary or tertiary structural extent. Rearrangement of these discrete protein "modules" at intron boundaries are hypothesized to have been the primary means by which early proteins acquired diverse structures and functions (Darnell and Doolittle 1986; Gilbert 1987; Gilbert et al. 1997). Because introns-early hypotheses predict correlations between gene and protein structure resulting from these rearrangements (de Souza et al. 1996), predictions concerning introns found outside the CDS are less clear.

The 5′ and 3′ untranslated regions (UTRs) that bracket CDSs are fundamental structural and regulatory regions of eukaryotic genes (Ptashne and Gann 2001; Larizza et al. 2002; Mignone et al. 2002; Wilkie et al. 2003). UTRs are known to contain large numbers of introns (Pesole et al. 2001), yet intron abundance and distribution in UTRs have received little study, and there is a lack of hypotheses specifically addressing the evolution of introns within UTRs. Our goal here is to begin to address these gaps in our knowledge. The few summary data available (Pesole et al. 2001) indicate that 22%–26% of metazoan 5′ UTRs carry introns, with lower frequencies in plants (14%) and fungi (5%). The observation that ~4–5× fewer 3′ UTRs carry introns in these taxa is especially curious given that, within taxa, 3′ UTRs are generally 2–3× longer than 5′ UTRs and would thus be expected to form larger targets for random intron insertion. Data from Pesole et al. also suggest the possibility of a strong barrier against carrying >1 intron in 5′ UTRs for all taxa, and a similar but less stringent barrier for introns in 3′ UTRs that does not appear to be as consistent among taxa. These patterns stand in

Key words: untranslated region, intron, genome evolution.

E-mail: dgscofie@indiana.edu.

Mol. Biol. Evol. 23(12):2392–2404. 2006
doi:10.1093/molbev/msl111
Advance Access publication September 15, 2006

Table 1
Sources and Versions for Genome and Full-Length cDNA Sequence Data

Species	Genome Sequence	Full-Length cDNA	Reference
Arabidopsis thaliana	The Institute for Genome Research (version 13 June 2001)	Knowledge-based Oryza Molecular biological Encyclopedia (24 October 2003)	(Castelli et al. 2004)
Drosophila melanogaster	Berkeley Drosophila Genome Project (release 3)	Berkeley Drosophila Genome Project (10 July 2003)	(Stapleton et al. 2002)
Human	GenBank (build 34.3)	Mammalian Genome Consortium (28 January 2004)	(Strausberg et al. 2002)
Mouse	GenBank (build 32.1)	Mammalian Genome Consortium (28 January 2004)	(Strausberg et al. 2002)

sharp contrast to those for the CDS in most multicellular species, the vast majority of which carry multiple introns (Lynch and Conery 2003).

Additional differences between UTRs and CDSs may affect intron size, abundance, and distribution. UTR regions are under less stringent substitutional constraint (vs. nonsynonymous sites) than CDSs, and have a higher indel frequency and length heterogeneity (Graur and Li 2000; Larizza et al. 2002; Shabalina et al. 2004). As a result, introns in UTRs may experience less stabilizing selection for some traits than introns in CDSs, in which case sharp discontinuities in intron traits at CDS-UTR boundaries may be expected. Furthermore, so-called "ancient" CDS introns that were shared by multiple eukaryotic lineages tended to be found in more conserved regions of the CDS (Rogozin et al. 2003); the more dynamic nature of the UTRs may thus promote intron loss and result in lower intron abundance than in CDSs. An additional consideration is that intron distributions that may promote CDS quality via NMD (Lynch and Kewalramani 2003) may provide no benefit within the 5′ UTR, though they may have direct effects on intron abundance and distribution in the downstream 3′ UTR (Nagy and Maquat 1998).

With the explosive growth in genome sequencing projects, a variety of computational methods have been developed to indirectly infer gene structure from genome sequence data, including the detection of intron–exon boundaries; see Zhang (2002) for a comprehensive review. These methods have reached a high level of performance such that they can recognize the large majority of intron–exon boundaries within the CDS. However, despite significant recent advances, recognition of intron–exon boundaries within UTRs, which lack the strong contextual signal provided by a valid open reading frame, is considerably more error prone (Eden and Brunak 2004). The recent availability of large libraries of full-length cDNA transcripts, when rigorously aligned to complete genome sequences for the same species, allows for the direct determination of intron–exon structure within UTRs. As a result, we are able to directly examine intron size, abundance, and distribution in UTRs of thousands of transcripts from each of four species, *Drosophila melanogaster*, *Arabidopsis thaliana*, human, and mouse.

Materials and Methods

Data Sources

We obtained publicly available genome and full-length cDNA sequence data for *D. melanogaster*, *A. thaliana*, human, and mouse (table 1). Boundaries between UTRs and CDSs in full-length cDNA sequences were determined using annotations from GenBank (*D. melanogaster* and *A. thaliana*) and the Mammalian Genome Consortium (human and mouse).

Intron Positions

Intron positions were determined through the recognition of gaps in alignment of full-length cDNA transcripts with genomic sequences. In brief, for a single full-length cDNA aligned against a contiguous stretch of genomic sequence, exons were determined as proximal blocks of homologous sequence alignment between full-length cDNA and genomic sequence, whereas introns were determined as gaps between exons consisting solely of genomic sequence.

We first cleaned the full-length cDNA libraries by removing transcripts with inconsistent annotation and incomplete CDSs. We then aligned the cleaned library of full-length cDNA transcripts to genome sequences with BLAT (Kent 2002). Apart from being very fast, the search algorithm used by BLAT has at least two advantages for alignment of potentially spliced transcripts to genome sequences. First, BLAT begins by searching for high-quality matches of short discrete sequences (K-mers, each 8–16 nt), and attempts to "stitch together" proximal high-quality K-mers by extending the match through intervening sequences that also provide high-quality matches. The scale at which this occurs is that of typical exon size. Second, once K-mers are stitched into blocks of high-quality alignment, gaps between matching blocks are adjusted so that the ends of gaps provide the best match to consensus sequences typical of intron ends (Kent 2002).

The BLAT alignment for each transcript was refined in two steps. First, the best alignment was chosen, defined as that alignment having the highest sequence identity greater than 95%; if no alignment had sequence identity greater than 95%, the transcript was discarded. If there were multiple best alignments with equal sequence identity, the longest alignment was chosen. Second, putative exon blocks separated by fewer than 5 bp were merged. Under some sequence and gap size conditions, BLAT does not stitch together proximal blocks (Kent 2002). We reasoned that gaps with fewer than 5 bp represent indels within full-length transcript sequences rather than actual introns. These small gaps occurred in 0.02% of *A. thaliana* full-length alignments, 15% of *D. melanogaster* alignments, 19% of human alignments, and 25% of mouse alignments. Per gap-containing full-length alignment, the mean total gap

length was 7.7 bp in *A. thaliana*, 2.8 bp in *D. melanogaster*, 6.2 bp in human, and 6.7 bp in mouse. We merged these gaps while moving through each alignment in a 5′–3′ direction along the positive-sense genomic strand. We thus introduced the possibility of a slight bias to intron positions that increases in absolute magnitude from 0 bp at the intron in the 5′-most genomic position within the alignment to a maximum of the total gap length at the intron in the 3′-most genomic position within the alignment. For genes encoded on the positive-sense genomic strand, this bias increased in the 5′–3′ direction within the full-length transcript, whereas for genes encoded on the negative-sense genomic strand, this bias increased in the 3′-to-5′ direction within the transcript. As genes have essentially equal proportions of positive- and negative-sense orientations in these genomes, the data set-wide degree of bias was negligible. We recorded intron positions according to their location from the 5′ end of each full-length transcript.

Following the refinement of the BLAT alignment, we created our set of "qualifying transcripts." A qualifying transcript contained at least one intron within its 5′ UTR, CDS, or 3′ UTR, as indicated by the alignment. Additionally, to avoid potential inconsistencies introduced by our use of automated alignments, we required that all introns in a qualifying transcript were between 20 bp and 100 kbp (100,000 bp) in length. We chose 20 bp as our minimum intron size because we were concerned about the potential inflation of intron numbers due to spurious gaps larger than our merge limit of 5 bp, and because very few introns are known to be less than 20 bp in length. For example, minimum CDS intron size was 13 bp in 2903 genes from 10 eukaryotes (Deutsch and Long 1999) and minimum intron size in ESTs was 27 bp in a diverse collection of fungi (Kupfer et al. 2004). A number of introns as large as 100 kbp and larger are known in, for example, humans (Nobile et al. 1997; Bärlund et al. 2002) but we did not wish to include such introns in our data set without manual confirmation of each such alignment. Such extremely large introns were extremely rare in our data set. For all species, larger data sets constructed using less stringent qualifying rules resulted in equivalent estimates and distributions, though the occurrence of extremely large introns (>100 kbp) for *D. melanogaster*, human, and mouse somewhat increased estimates sensitive to extreme outliers (data not shown).

Statistical Analysis of Intron Distribution

We analyzed the general distribution of introns within each region by examining the distribution of exon sizes, which are directly dependent upon intron locations. For example, the expected mean exon size for a region containing n_i introns is (length of region/(n_i + 1)) regardless of the pattern of intron distribution, so we instead calculate the effective number of exons n_e, which is sensitive to the variance in exon size (Lynch and Kewalramani 2003). When introns are uniformly distributed, resulting in all exons having equal length, then $n_e = n_i + 1$. When introns are closely clumped so that one exon is much longer than the others, then n_e approaches 1. To calculate n_e, for each exon within a region, we determined its size e_i relative to the total length of the region, so that the e_i within each region sum to 1. We then calculated n_e for each species, region, and number of introns using $n_e = 1/\sum_{i=1}^{n} e_i^2$, which is equivalent to the classical formula used in population genetics to calculate the effective number of alleles for a locus (Kimura and Crow 1964).

Falling between the two extremes of $n_e = 1$, for unusually high variation in exon length, and $n_e = n_i + 1$, for no variation in exon length, are expected values of n_e for a random distribution of introns within a region, which serve as a null model for comparison. Introns positioned at random create a distribution of exon sizes that follows the "broken-stick" distribution for random partitions of a finite distance (MacArthur 1957; Goss and Lewontin 1996; Lynch and Kewalramani 2003). For each set of species' genes containing n_i = 1–5 introns in a region, we calculated mean $n_e \pm$ standard error. We then calculated the broken-stick expectation of n_e for n_i random intron locations via simulation. To create this null expectation, we randomly chose a region length in bp from the set of all observed regions having n_i introns for each species, with replacement. Within this randomly chosen observed region length, we then randomly chose locations for n_i introns and calculated the relative length e_i of each resulting exon. We calculated n_e for this simulated region and repeated this for 10^5 iterations for each combination of species, region, and number of introns n_i. We call this random distribution of n_e values the "unrestricted random distribution." Other than an absolute minimum exon size limit of 1 bp, we did not place a lower limit on exon size in these simulations, so this approach assumes an absence of exon-size constraints. However, such extremely short exons are quite rare (Deutsch and Long 1999) and introduce the possibility of numerous splicing difficulties (Dominiski and Kole 1991, 1992; Sterner and Berget 1993; Carlo et al. 1996). We thus also simulated "minimum-exon-size random distributions" for each combination of species, region, and number of introns. In these simulations, the random draw of intron locations was repeated until the size of the smallest resulting exon was ≥20 bp; this is somewhat smaller than the smallest exon size that can apparently be constitutively spliced reliably without additional splicing enhancers (~50 bp; Dominiski and Kole 1991, 1992). The minimum-exon-size random distribution has a larger mean n_e than the unrestricted random distribution, and the difference between the mean n_e of the two distribution increases as the length of the simulated region decreases (see also Goss and Lewontin 1996). As there are species-specific relationships between mean total exon length of a region and the number of introns found therein (Lynch and Kewalramani 2003), our minimum-exon-size distributions are not independent of species identities. There is negligible difference among species in the two random distributions within CDSs (data not shown). However, the shorter lengths of 5′ UTRs accentuate the among-species differences (fig. 3), such that we present separate minimum-exon-size random distributions for human and mouse as a group, and for *A. thaliana* and *D. melanogaster* as a group.

Here, we will call observed intron distributions overdispersed or underdispersed in comparison to one of these random distributions if the mean n_e value is greater than or

Table 2
Summary Statistics for Introns in the 5′ UTR, CDS, and 3′ UTR, from Qualifying Full-Length Transcripts that Contained At Least One Intron; Qualifying Transcripts are Those that Met the Data Consistency Conditions Described in the Text

	Arabidopsis thaliana			*Drosophila melanogaster*			Human			Mouse		
	5′ UTR	CDS	3′ UTR	5′ UTR	CDS	3′ UTR	5′ UTR	CDS	3′ UTR	5′ UTR	CDS	3′ UTR
Number of Qualifying Transcripts		10,562			3,424			5,236			4,527	
Number of Occupied Regions (% of Qualifying)	1,805 (17.1)	10,184 (96.4)	269 (2.5)	1,130 (33)	3,203 (93.5)	37 (1.1)	1,988 (38)	4,956 (94.7)	38 (0.7)	1,827 (40.4)	4,292 (94.8)	33 (0.7)
Number of Introns	2,012	55,510	382	1,490	10,507	63	2,721	37,508	75	2,490	35,376	54
Mean Number of Introns per Qualifying Region (SE)	0.2 (0.005)	5.3 (0.046)	0.03 (0.003)	0.4 (0.013)	3.1 (0.043)	0.02 (0.004)	0.5 (0.012)	7.2 (0.078)	0.01 (0.003)	0.5 (0.014)	7.8 (0.097)	0.01 (0.002)
Mean Introns per Occupied Region (SE)	1.1 (0.01)	5.5 (0.05)	1.4 (0.09)	1.3 (0.02)	3.3 (0.04)	1.7 (0.22)	1.4 (0.02)	7.6 (0.08)	1.7 (0.23)	1.4 (0.02)	8.2 (0.10)	1.3 (0.15)
Intron Occupancy, Introns/Exon kbp	1.44	4.17	0.14	1.55	2.06	0.05	3.21	5.84	0.02	3.49	5.82	0.01
Intron Density, Introns/ Exon kbp in Occupied Regions	5.27	4.30	2.80	2.73	2.14	1.66	5.85	6.09	1.91	6.29	6.05	1.06
Median Intron Size (i.q.d./2)	268 (149)	98 (37)	103 (5)	553 (1,067)	68 (93)	142 (311)	2,643 (3,968)	1,334 (1,525)	1,303 (1,333)	2,328 (3,181)	1,095 (1,131)	619 (1,397)
Mean Intron Size (SE)	332 (7)	158 (1)	205 (14)	2,889 (183)	818 (35)	1,309 (369)	8,223 (266)	3,749 (42)	2,891 (496)	7,205 (253)	2,874 (35)	2,603 (561)
Coefficient of Variation Intron Size, SD / Mean	0.98	1.00	1.36	2.45	4.38	2.24	1.69	2.15	1.49	1.75	2.28	1.58

NOTE.—Occupied regions are those regions that contained at least one intron. i.q.d./2 = (75th quantile − 25th quantile)/2.

less than, respectively, the simulated values of n_e for the appropriate random distribution. Overdispersed introns are more uniformly distributed in the region in comparison to a random distribution, whereas underdispersed introns are more clumped.

Results

We found introns to be approximately as abundant in 5′ UTRs as reported by Pesole et al. (2001), with correspondence to overall patterns of Pesole et al. including approximately equal percentages of 5′ UTRs carrying introns in *D. melanogaster*, human, and mouse, about half that number in *A. thaliana* (table 2). We found introns to be much less abundant in 3′ UTRs than did Pesole et al. (2001); this may be due to our sampling a comprehensive set of full-length cDNAs for each species, as more recent versions of the associated nonredundant databases (Pesole et al. 2002) contain fewer records of 3′ UTRs carrying introns (e.g., 1.3% of human 3′ UTRs vs. 7.9% in the original report).

Intron size distributions for all species were strongly positive skewed in all regions (fig. 1), consistent with several other studies (Mount et al. 1992; Deutsch and Long 1999; Adams et al. 2000; Comeron and Kreitman 2000; Lander et al. 2001; Yu et al. 2002). There is no clear bimodal distribution of "small" and "large" introns (e.g., Maroni 1994) in any region for any species. Modal intron sizes were similar across regions within species, and the overall shape of the distributions is similar, but the right tail in the 5′ UTR carries more density in all species than the CDS. In all regions, the right tail of the distribution for human and mouse carries more density than for *A. thaliana* and *D. melanogaster*, as was observed for the initial release of the human genome (Lander et al. 2001). In both *A. thaliana* and *D. melanogaster*, the strongly peaked distribution of intron sizes in the CDS was duplicated in the 5′ UTR and 3′ UTR. *Arabidopsis thaliana* is distinct from the other three species in having the right tail of the size distribution to be shorter by about an order of magnitude or more than the right tail for the other species. Low sample sizes preclude similar summaries for introns in human and mouse 3′ UTRs.

In all species, median intron sizes (table 2) were greater in the 5′ UTR than in the CDS (Mann–Whitney $U > 1.1 \times 10^7$, $P \sim 0$ for all) and the 3′ UTR ($U > 5.8 \times 10^4$, $P < 0.001$ for all). Median intron size in the 3′UTR was greater than in the CDS for *A. thaliana* and *D. melanogaster* ($U > 2.5 \times 10^5$, $P \sim 0.001$ for both). For human and mouse, despite larger difference in median intron size between the CDS and 3′ UTR in comparison to *A. thaliana* and *D. melanogaster*, intron sizes did not differ significantly between these regions (human, $U = 1.4 \times 10^5$, $P = 0.76$; mouse, $U = 9.6 \times 10^5$, $P = 0.91$) because of greater variation in intron size (table 2). Similarly, larger median intron sizes in human and mouse are due to the greater frequency of larger intron sizes rather than due to a larger modal intron size (fig. 1). In *A. thaliana* and *D. melanogaster*, median and mean intron lengths are short in comparison to human and mouse and the centers of the distributions are much more tightly constrained, as is apparent both from figure 1 and the much shorter interquartile distance (i.q.d.) (table 2). *A. thaliana* had the lowest Coefficient

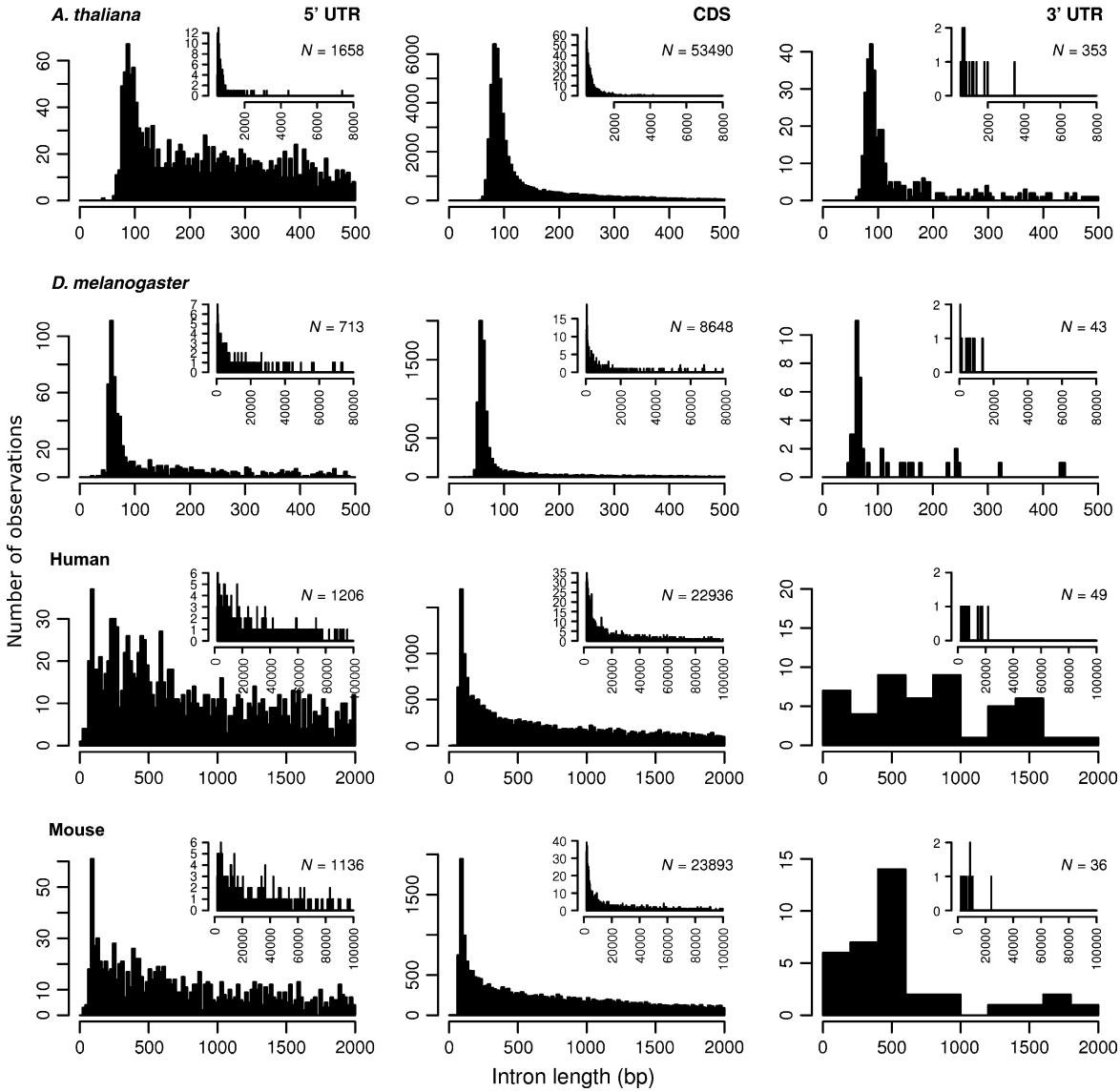

FIG. 1.—Intron length distributions in the 5′ UTR (left), CDS (center), and 3′ UTR (right) of full-length cDNA transcripts for *Arabidopsis thaliana*, *Drosophila melanogaster*, human, and mouse. The inset histogram continues the right tail of the main histogram with the identical bin size, so that the ordinate axes of the two histograms are on the same frequency scale. Bin sizes are 5 bp for *A. thaliana* and *D. melanogaster* and 20 bp for human and mouse, except for human and mouse 3′ UTRs for which bin sizes are 200 bp due to small sample sizes.

of Variation (C.V. = standard deviation/mean) for intron size in all regions, due to the much shorter right tail of its intron size range, and *D. melanogaster* the largest, owing to its combination of a tightly constrained peak of intron size density and an extremely long right tail. Human and mouse C.V.s were similar in all regions. Consistent with previous work (Abril et al. 2002), we found mouse introns to be slightly smaller than human introns in all regions (table 2).

Intron Sizes across the 5′ UTR-CDS Boundary

It has been proposed that 5′-ward introns tend to be larger because of the possibility of their general use as hosts for regulatory elements (Duret 2001), although it is unclear what the null expectation for intron size should be. If this is true, then we should expect to find a gradient of decreasing intron size while moving downstream within the full-length transcript, based on empirical data that suggest a gradient of decreasing intron regulatory effects moving downstream within a gene (Nott et al. 2003; Rose 2004). To test for a decreasing trend in intron size from the 5′ UTR, across the 5′ UTR-CDS boundary, into the CDS, we examined median intron size in 50 bp windows from 500 bp upstream to 1,000 bp downstream of the start codon (fig. 2). Surprisingly, within transcripts of all four species, we found a strong discontinuity in median intron size moving from the 5′ UTR to the CDS (fig. 2). In the immediate vicinity of the start codon, moving across the 5′ UTR-CDS boundary, there is a drop in median intron size of ~180 bp in *A. thaliana*, ~500 bp in *D. melanogaster*, and ~1700 bp in both human and mouse (fig. 2). Note that within the 5′ UTR of all species, starting at around 200 bp upstream of the start codon, there is a tendency for median intron size to increase as one moves downstream toward the start codon (fig. 2). Median size of

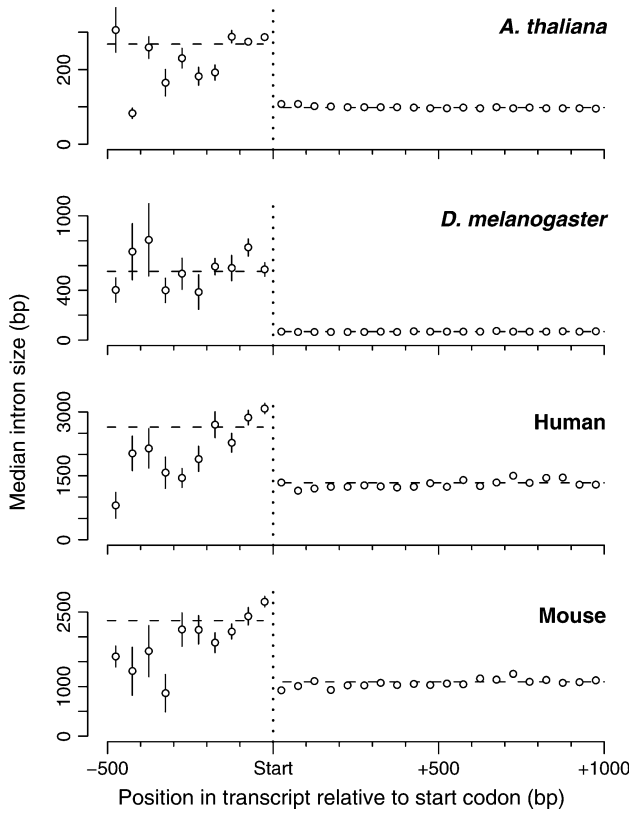

FIG. 2.—Median intron size versus relative position within transcript for all introns. For each species, the region to the left of the vertical dotted line is the 5′ UTR, the region to the right is the CDS. Median intron sizes within each 5′ UTR and CDS are indicated by horizontal dashed lines. Bin sizes are 50 bp. The total length of each error bar is equal to the interquartile distance divided by the square root of the sample size within each bin; note that some error bars are more narrow than the height of the plotting symbols.

5′ UTR introns within the region from the start codon to 200 bp upstream of the start codon is significantly greater than the median intron size of 5′ UTR introns upstream of the 200 bp partition (Mann–Whitney $U > 1.2 \times 10^5$, $P < 0.0025$ for all species). As the partition is moved farther upstream in 50-bp increments, median intron size remains significantly greater (Mann–Whitney U, $P \leq 0.05$ for all species) downstream of the partition up to a partition position of 450 bp upstream of the start codon (data not shown).

Intron Numbers and Positions within UTRs

As we previously observed for the CDS (Lynch and Kewalramani 2003), there is a clear linear relationship between 5′ UTR length and intron number for all species (fig. 3). In accordance with Pesole et al. (2001), we found sharply reduced numbers of 5′ UTRs carrying >1 intron (fig. 3). For our data set, regressions of region length versus number of introns reveal that slopes for 5′ UTRs are ~1.2–2.5× greater than in CDSs, and intercepts are ~470–830 bp lower in 5′ UTRs than CDSs (table 3). For the 3′ UTR, only *A. thaliana* had sufficient sample size across the range of intron numbers to perform the regression ($n = 264$); the slope was 281 (SE = 26) and the intercept was 133 (34), both of which were significant ($P < 0.001$) and both of which were more similar to this species' estimates for the 5′ UTR than the CDS.

Another trend that distinguishes intron distributions within the 5′ UTR from those in the CDS is that absolute intron positions are independent of length of the region (fig. 3). For example, mean position of the first intron is ~144 and ~217 bp downstream of the 5′ end of the 5′ UTR in *A. thaliana* and *D. melanogaster*, respectively, and ~130 and ~116 bp downstream in human and mouse, and this is relatively constant regardless of the total length of the 5′ UTR or the number of introns found there (fig. 3). This is in contrast to the pattern observed in the CDS of all four species, in which the mean position of the first intron shifts increasingly upstream as more introns occupy the region (fig. 3); this trend continues for transcripts having >5 introns in the CDS (data not shown, see also Lynch and Kewalramani 2003). Similar trends are apparent for introns in other ordinal positions of the 5′ UTR, except for *A. thaliana* where sample size is prohibitively small for 5′ UTRs with ≥3 introns (fig. 3).

Intron densities (introns/exon kbp) among all occupied regions (those regions containing at least one intron) were greater in the 5′ UTR than the CDS (table 2; Mann–Whitney $U, P < 10^{-10}$ for all species), whereas among all regions in the data set, intron densities were greater in the CDS than the 5′ UTR ($P < 10^{-10}$ for all species). Intron densities were lower in the 3′ UTR than both the 5′ UTR and CDS both among all occupied regions and among all regions for all species, except for *D. melanogaster* where intron density did not differ between occupied 3′ UTRs and occupied CDSs ($P = 0.11$).

Dispersion of Introns within UTRs

We reasoned that intron distributions that promote CDS quality via NMD (Lynch and Kewalramani 2003) would provide no benefit within the 5′ UTR; thus we expected to find a random linear distribution of introns within 5′ UTRs. Introns were dispersed at random in the 5′ UTRs of all species in accordance with expectations of the minimum-exon-size random distribution, and among-species differences tended to be somewhat less distinct than in the CDS (fig. 4). To further test this observation, we compared n_e values for observed regions containing 1–5 introns against n_e values for a null distribution created by assembling a random data set containing 1×10^5 regions drawn from the species' minimum-exon-size random distribution with the same number of introns. Mean n_e estimates within 5′ UTRs were not significantly different from the minimum-exon-size random distribution (Mann–Whitney U test) except for human 5′ UTRs with ≥3 introns ($P < 0.05$ for all) and mouse 5′ UTRs with ≥3 introns ($P < 0.005$ for all). In contrast, mean n_e estimates within CDSs containing 1–5 introns were significantly different from the random distribution for *A. thaliana* (all $P < 0.0001$) and for all human and mouse CDSs ($P < 0.001$) except those containing one intron ($P = 0.8$ for human, $P = 0.2$ for mouse). Interestingly, as is apparent in figure 4, mean n_e values for CDSs of *D. melanogaster* did not significantly differ from the random distribution (all $P > 0.8$), except for those CDSs containing 5 introns ($P = 0.007$).

Discussion

Larger Introns in 5′ UTRs

Our three primary observations related to intron size in 5′ UTRs are: 1) markedly larger introns in the 5′ UTR than

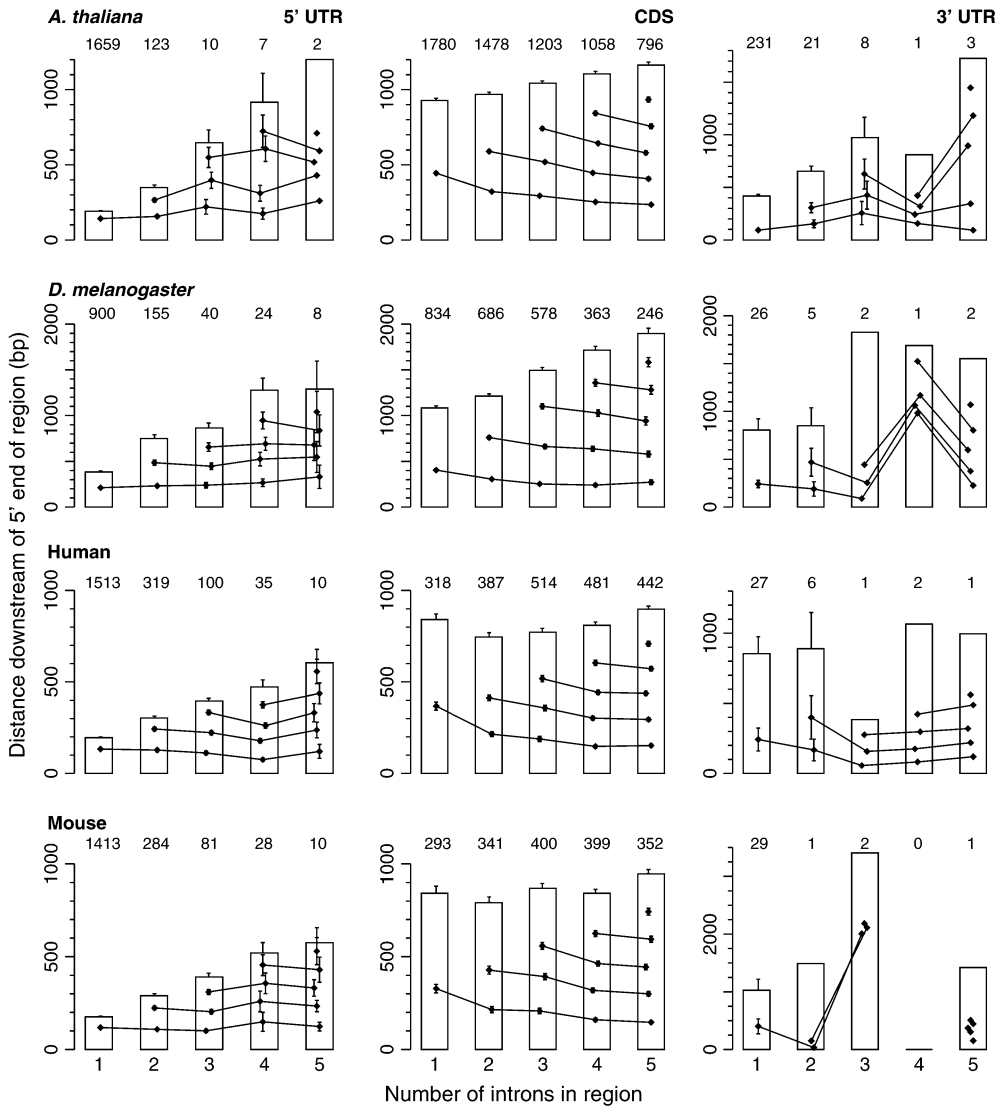

FIG. 3.—Mean intron positions within regions. Bars indicate mean length ± SE of regions containing that many introns. Mean intron position ± SE is plotted within each transcript bar so that its relative position within the mean region is apparent. Bars are oriented so that the 5′ end of each region starts at 0. Introns occupying the same ordinal position within each region (first, second, etc.) are connected by lines across transcripts. The ordinate is identical for the 5′ UTR and CDS of each species, and varies for the 3′ UTR. Numbers above bars indicate sample sizes; error bars are omitted where sample size <5. Error bars that are not visible when sample size ≥5 are more narrow than the height of the plotting symbols.

in the adjacent CDS for all species (table 2 and fig. 1); 2) a threshold-like drop in intron size across the 5′ UTR-CDS boundary (fig. 2); and 3) an increase in intron size in proximity to the start codon (fig. 2). Taken together, these observations indicate markedly different forms of selection on intron size in the 5′ UTR versus the CDS that may occur over very short distances.

We propose a straightforward mechanism for the occurrence of selection differences that may explain these three observations, driven by the potentially deleterious

Table 3
Regressions of Region Length against Number of Introns, for all 5′ UTRs in the Data Set with 1–5 Introns, and all CDSs in the Data Set with 1–10 Introns. All Slopes and Intercepts are Significant at $P < 0.0001$, except for *Arabidopsis thaliana* 5′ UTR Intercept, Which is not Significantly Different from Zero ($P = 0.2$)

	5′ UTR (1–5 Introns)			CDS (1–10 Introns)		
Species	*n*	Slope (SE)	Intercept (SE)	*n*	Slope (SE)	Intercept (SE)
Arabidopsis thaliana	1801	202 (9)	−13 (10)	8849	79 (2)	817 (11)
Drosophila melanogaster	1127	282 (17)	110 (25)	3140	203 (7)	864 (25)
Human	1977	100 (5)	97 (8)	3820	81 (3)	569 (15)
Mouse	1816	110 (5)	67 (8)	3131	79 (3)	623 (19)

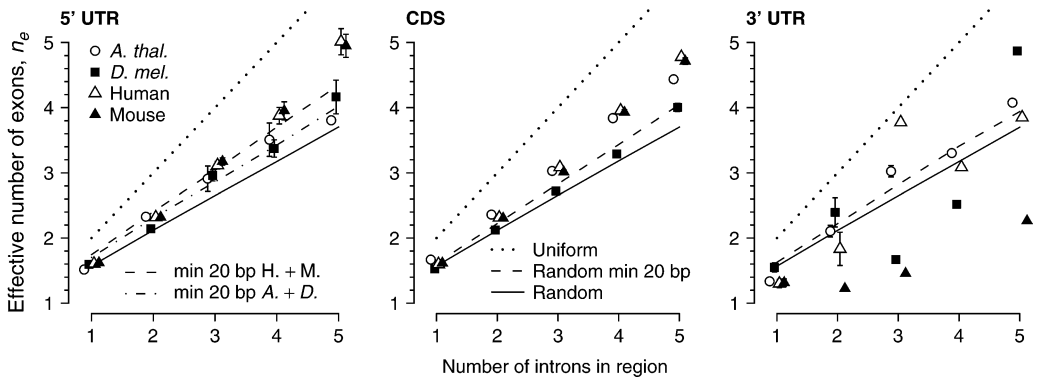

FIG. 4.—Intron dispersion within UTRs. Mean effective number of exons $n_e \pm$ SE for regions containing 1–5 introns in the UTRs and the CDS. Lines indicate expected values of n_e for intron distributions having uniform exon sizes (dotted), random exons of minimum size 20 bp (dashed), and random exons of minimum size 1 bp (solid). For the 5′ UTR, separate minimum 20 bp distributions are provided for human and mouse as a group (dashed) and *Arabidopsis thaliana* and *Drosophila melanogaster* as a group (dash-dot). Values above a random distribution indicate an overdispersed (more uniform) distribution, whereas those below a random distribution indicate an underdispersed (more clumped) distribution. See figure 3 for sample sizes; error bars are omitted where sample size <5. Error bars that are not visible when sample size ≥5 are more narrow than the height of the plotting symbols.

effects of upstream AUG (start) codons (uAUGs) within 5′ UTR exons. Because uAUGs cannot by definition occur in the CDS, this selection does not exist downstream of the start codon. Splice site locations may be less conserved (i.e., less static) for 5′ UTR introns than for CDS introns; thus, both intron contraction—the shift of 5′ UTR intron sequence to an adjacent 5′ UTR exon—and intron expansion—the shift of 5′ UTR exon sequence to an adjacent 5′ UTR intron—may be more likely to occur in the 5′ UTR than the CDS. We propose the existence of: 1) selection against intron contraction, due to the potential introduction of uAUGs residing in 5′ UTR introns at nearly neutral proportions and 2) selection for intron expansion, due to the beneficial effects of both removing uAUGs from 5′ UTR exons and preventing the appearance of new uAUGs by reducing the total 5′ UTR exon length. As will be outlined below, selection against intron contraction is expected to be stronger than selection for intron expansion, which under many circumstances may be effectively neutral. Thus, if a splice-site shift results in an intron contraction or an intron expansion at approximately equal rates, increases in 5′ UTR intron size are due to greater rates of loss of specific intron contraction events, rather than greater rates of fixation of specific intron expansion events.

The available empirical evidence for 5′ UTR introns suggests that splice sites within 5′ UTRs are less conserved than within adjacent CDSs. Among rice cultivars, splice sites of 5′ UTR introns have been shown to be relatively less conserved than splice sites for introns in adjacent CDSs for the *waxy* gene (Cai et al. 1998). Additionally, in an evolutionary study involving several species of plants, the 5′ UTR intron found in the *PgiC* gene showed 6 splice-site shifts, 4 of which resulted in intron expansion. The splice sites of the 20+ introns found in the CDS of the same gene were entirely conserved across all species (Gottlieb and Ford 2002). Furthermore, it is intuitively clear that alleles are more likely to remain functional following splice-site shifts involving 5′ UTR introns than those involving CDS introns. It is well known that 5′ UTRs have higher indel frequency, length heterogeneity, and DNA substitution rates than CDSs (Graur and Li 2000; Larizza et al. 2002; Shabalina et al. 2004), whereas in the CDS, any movement of a splice site is likely to create a null allele owing to the introduction of a frame shift, codon gain, or codon loss (Stoltzfus et al. 1997; Lynch 2002). Although the empirical studies are consistent with our proposed mechanism, with just two plant genes having been examined, and with the observed splice-site shifts all being <10 bases in size (Cai et al. 1998; Gottlieb and Ford 2002), any conclusions regarding the commonness or rarity of splice-site shifts as well as their size distribution are tentative.

We now briefly consider the strength of AUG-driven selection involving intron contraction and expansion in the 5′ UTR, and will present a more detailed examination of the model in future work. AUG triplets appear within 5′ UTR introns at ~0.82× the neutral expectation regardless of the position of the intron within the 5′ UTR, whereas within 5′ UTR exons there is a gradient of underrepresentation of uAUGs that increases while moving downstream to ~1/3× the neutral expectation just upstream of the start codon (Rogozin et al. 2001; Lynch et al. 2005). Consider an intron contraction involving n bases. For $n \geq 3$, the probability, $P[n, q]$, that a sequence of n bases includes at least one AUG when AUG occurs at $q = 0.82\times$ the neutral expectation is approximately $P[n, q = 0.82] = 1 - [1 - 0.82(1/64)]^{(n-2)}$. Using the diffusion approximation (Kimura 1962), an intron contraction that introduces an uAUG has a chance of fixation of $-2s/(1 - e^{4Ns})$, where N is the effective population size and s is the selective disadvantage of an uAUG. We will derive approximate values for $4Ns$ below. An intron contraction that does not introduce an uAUG will still experience very weak negative selection due to the additional n exonic bases that serve to increase the exon length–dependent mutation rate to null alleles caused by point-mutational gain of uAUGs. This negative selection is expected to be swamped by drift under most population sizes (Lynch et al. 2005), so this fraction $(1 - P[n, q = 0.82])$ fixes at the neutral rate, $1/2N$. The overall fixation rate of an intron contraction F_C involving n bases is thus approximately $1/2N(1 - P[n, q = 0.82] \times [1 + 4Ns/(1 - e^{4Ns})])$.

An intron expansion of n bases may experience positive selection if it converts an exonic uAUG into an

intronic AUG. This benefit will vary with the probability of occurrence of an uAUG within 5′ UTR exons, so we choose two "bookend" spots along the gradient within 5′ UTR exons where uAUG occurs at 2/3 and 1/3 of the neutral expectation and calculate the probabilities $P[n, q = 2/3]$ and $P[n, q = 1/3]$. For sake of simplicity within this brief analysis, we assume that the selective disadvantage of exposing a previously hidden AUG is equal in absolute magnitude to the selective advantage of hiding a previously exposed uAUG; we will relax this assumption in future work. With the selective advantage associated with the removal of an uAUG equal to $-s$, the chance of fixation is $P[n, q] \times 2s/(1 - e^{-4Ns})$. Assuming that any selection associated with an intron expansion that does not encompass an uAUG will be swamped by drift, we have an overall fixation rate of an intron expansion F_E involving n bases of between $F_E = 1/2N(1 - P[n, q = 2/3] \times [1 - 4Ns/(1 - e^{-4Ns})])$ and $F_E = 1/2N(1 - P[n, q = 1/3] \times [1 - 4Ns/(1 - e^{-4Ns})])$.

We compare the relative strength of selection favoring intron expansion versus intron contraction by first estimating appropriate values for $4Ns$. The ratio of the fixation rates of the birth b and death d of AUG triplets under the diffusion approximation with mild negative and positive selection, respectively, is $[b \times -2s/(1 - e^{4Ns})]/[d \times 2s/(1 - e^{-4Ns})] = (b/d) \times e^{-4Ns}$. The neutral expectation ($s = 0$) is simply b/d, so due to the underrepresentation of AUG at the same bookend locations within 5′ UTR exons chosen above, we have $e^{-4Ns} \sim 2/3$ and 1/3, resulting in $4Ns \sim 0.41$–1.1, respectively. If physical splice-site shifts resulting in intron expansions or contractions are equally likely, then the selective bias for intron expansion can be examined by calculating the scaled probability of fixation of intron expansion versus intron contraction, $\Theta_{E/C} = F_E/F_C$, which is <1 when selection favors intron contraction and >1 when selection favors intron expansion. With $q = 2/3$, for $n = 3$, 10, 20, and 50 bases, $\Theta_{E/C} \sim 1.00$, 1.04, 1.08, and 1.19, respectively. With stronger selection against uAUG giving $q = 1/3$, $\Theta_{E/C} \sim 1.01$, 1.07, 1.17, and 1.44, respectively. Regardless of the strength of selection against uAUG, the strength of selection favoring intron expansion increases with increasing n. However, the very limited applicable empirical data do not show splice-site shifts of $n > 10$ (Cai et al. 1998; Gottlieb and Ford 2002); thus more data on splice-site shifts will assist in judging the applicability of our specific model and of models of intron-size evolution generally.

These calculations also help to explain at least two other observations. First, we observed the occurrence of gradients of increasing 5′ UTR intron size with increasing proximity to the start codon (fig. 2) that accompanies the previously noted underrepresentation of uAUG. Following our calculations, the stronger selection against uAUG in proximity to the start codon increases $\Theta_{E/C}$ and thus increases the relative likelihood of intron expansion. Second, as noted above, we have previously shown that introns within the 5′ UTR harbor AUGs in nearly (but not completely) neutral proportions (Lynch et al. 2005). The slight underrepresentation of uAUGs that we observed within 5′ UTR introns may be an expected side effect of intron expansion via incorporation of sequences from uAUG-poor 5′ UTR exons.

Our model predicts that the dynamics of both splice-site shifts within the 5′ UTR and sequence evolution in the vicinity of these splice sites may be rather complex. For example, our diffusion approximations address the selective environment favoring the first fixation event. Subsequent shifts involving the same splice-site experience different and size-dependent patterns of selection arising from sequence changes in both the exon and intron flanking the splice site. There are two basic cases to consider. In the first, an intron expansion that does not convert an exonic uAUG to an intronic AUG will reduce the strength of selection against intron contraction via a second splice-site shift, provided that the second shift involves no more bases than the first. In the second case, an intron expansion converts an exonic uAUG into intronic AUG, which increases selection against a second shift that results in intron contraction provided that the shift is large enough to incorporate the former uAUG. There are similar considerations for the third and subsequent splice-site shifts. Thus, our model emphasizes the complexity of uAUG/AUG-related dynamics around splice sites within the 5′ UTR. A thorough analysis of these dynamics is beyond the scope of the present paper. As an initial contribution of empirical data to this problem, we have examined the distribution of AUG triplets within 5′ UTR introns (see Supplementary Material online). The degree of asymmetry in under- and overrepresentation of AUG triplets within introns is surprisingly consistent across species, and suggests a common set of evolutionary forces. Because our model invokes the effects of selection against uAUG in the mature transcript, our model in its current formulation does not predict asymmetry in under- or overrepresentation of AUG at different ends of 5′ UTR introns. Within the context of our model, such a pattern may reflect true biases favoring or disfavoring splice-site shifts at different ends of 5′ UTR introns arising from, for example, underlying constraints on sequence evolution in these locations.

That said several alternative models for intron-size evolution are unable to explain our observations. As noted above, it has been proposed that 5′-ward introns would tend to be larger because of the possibility of their general use as hosts for regulatory elements (Duret 2001; Nott et al. 2003; Rose 2004), in which case we should have found a gradient of decreasing intron size while moving downstream within the full-length transcript. However, our observation of both markedly larger 5′ UTR introns and a size discontinuity across the 5′ UTR-CDS boundary are inconsistent with the existence of a continuous gradient of regulation-driven selection. The size difference is also unlikely to be due to a fundamental bias in indel rates between UTRs and CDSs, though indels clearly fix at higher rates in UTRs than CDSs (Graur and Li 2000; Larizza et al. 2002; Shabalina et al. 2004) and thus experience less negative selection than in more tightly constrained CDSs. Greater intron size in 5′ UTRs also seems unlikely to be due to within-gene differences in selection on intron size related to reduced gene expression (Castillo-Davis et al. 2002) or altered recombination (Carvalho and Clark 1999; Comeron and Kreitman 2000). Although there is likely to be strong

selection against overly short introns due to structural constraints related to splicing efficiency (Mount et al. 1992; Comeron and Kreitman 2000), it is unlikely that these structural constraints differ to any large degree between the 5′ UTR and CDS. In fact, if splicing-related structural constraints do differ between the 5′ UTR and CDS, it seems that such constraints would produce intron size patterns that are opposite of those observed here, due to less-conserved 5′ UTR exons being more likely to host splicing enhancer sequences than codon-constrained CDS exons and thus more able to facilitate the removal of smaller introns and exons (Sterner and Berget 1993; Carlo et al. 1996; Sterneret al. 1996).

The mechanism that we propose for intron-size evolution in the 5′ UTR emphasizes the aggregated result of relatively small individual shifts in splice sites. More extreme shifts in splice sites are likely to produce null transcripts whether they occur in the 5′ UTR or the CDS. For example, exon skipping (Berget 1995) in the 5′ UTR may leave the transcript without a valid start codon because there are typically just two exons—one noncoding and one partially coding—in the large majority of intron-bearing 5′ UTRs (table 2 and fig. 3). Nonetheless, there is need for caution in constructing models proposing distinct evolutionary trajectories for large and small introns (e.g., Maroni 1994; Vinogradov 2002) that do not also consider the potential for differences introduced by 5′ UTR-CDS context.

Random Intron Distribution in 5′ UTRs

We found clear support for our expectation that NMD-related selection for overdispersed intron distributions within the CDS did not extend into the 5′ UTR. The low intercepts for the regressions of 5′ UTR length versus number of introns suggest that the number of introns is largely a function of the length of the 5′ UTR. Thus, intron number within 5′ UTRs may be largely the result of a stochastic process dependent on available 5′ UTR "substrate," with minimum exon size determined by, for example, splicing-related structural constraints (Sterner et al. 1996). Similarly, the low numbers of 5′ UTRs with >2 introns (fig. 3) may result from the low frequency of longer 5′ UTRs (Lynch et al. 2005).

As for spatial distributions of introns, in nearly all cases we examined, 5′ UTR introns were randomly distributed in comparison to the minimum-exon-size random distribution (fig. 4). If it is assumed that 5′ UTR introns initially appear at random positions (Cho and Doolittle 1997), then 5′ UTR introns may be fixed "in place," with little selection on the intron distribution per se, except for those imposed by minimum exon-size constraints. If introns are subsequently lost from 5′ UTRs, then to maintain the random distribution, they must also be lost essentially at random, without regard to their location within the 5′ UTR. These results support the hypothesis that some sort of translation-associated process, for example, NMD, is involved in selection for intron locations in CDSs (Lynch and Kewalramani 2003).

Introns in the 3′ UTR

In contrast to both 5′ UTRs and CDSs, we found introns to be rarer in 3′ UTRs of all species (table 2 and fig. 3) than previous reports would suggest (Pesole et al. 2001). NMD-related selection was expected to keep intron numbers low in mammals such as human and mouse, which rely upon intron-associated exon junction complexes in their NMD pathway (Nagy and Maquat 1998; Maquat 2004b). However, we also observed much reduced numbers of 3′ UTR introns in *A. thaliana* (2.5% of transcripts with introns in the 3′ UTR versus 17.1% with introns in the 5′ UTR) and *D. melanogaster* (1.1% vs. 33%). Mean and median intron sizes for 3′ UTRs were similar to those in other regions (table 2), so it seems unlikely that our qualification criteria for aligned transcripts were biased against intron characteristics common in 3′ UTRs. It may be that low intron numbers in 3′ UTRs of species that do not rely upon introns for NMD may be maintained by homologous recombination with reverse-transcribed cDNAs, which would preferentially cause intron loss in 3′ ends of transcripts (Fink 1987; Feiber et al. 2002; Mourier and Jeffares 2003). If so, however, this is not reflected in a deficit of introns toward the 3′ end of CDSs. An additional possibility is suggested by the tight coupling of transcription, splicing, and other posttranscriptional mRNA processing (Maniatis and Reed 2002). Several mRNA-related processes initiate during or soon after transcription, and the dynamics of transcription termination are complex and time sensitive (Proudfoot 2003). As a result, there may simply be insufficient time or space for proper splicing of introns at 3′ ends of transcripts, and introns are thus inherently unstable in these areas. A final possibility follows from the dynamic nature of the 3′ UTR, in terms of base substitutions, insertions and deletions (Graur and Li 2000; Larizza et al. 2002; Shabalina et al. 2004). Given that the 3′ UTR is downstream of the CDS, it should be better able to tolerate sequence changes that directly affect the splicing of introns than either the CDS or the 5′ UTR. Thus, unless there is unusually strong positive selection for intron maintenance, the rate of intron loss is expected to be particularly high in 3′ UTRs.

Implications for Theories of Genome Evolution

It is readily apparent from our data set that failure to consider introns in the 5′ UTR has at least two implications for theories of genome evolution. First, the possibility of within-gene differences in selection on intron size suggests that purportedly genome-wide estimates of insertion–deletion ratios (e.g., Gregory 2004) may not only be biased by failure to consider 5′ UTR introns but that the context of individual introns has the potential to create conditions that modify selection on intron size (see above, and Ptak and Petrov 2002). The second implication is that for species such as *D. melanogaster* with a greater proportion of introns in the 5′ UTR and with fewer overall introns per transcript, 5′ UTRs carry a relatively greater proportion of intronic bp within genes. From table 2, we find that 33% of *D. melanogaster*'s intronic bp within our data set is found within 5′ UTRs, whereas the same is 7%, 14%, and 15% in *A. thaliana*, human, and mouse, respectively. Although our data set is explicitly limited to a subset of intron-bearing full-length transcripts (see Materials and Methods), it is clear that calculations of genome-wide

intron number and total intron content that are used to test a variety of hypotheses in genome evolution (e.g., Vinogradov 1999; Lynch and Conery 2003) will thus be underestimated to the degree that 5′ UTR introns are not considered in the data set, and this bias will be larger in species with an intron profile similar to that of *D. melanogaster*. One would also expect intron content to be underestimated by similar percentages in sequenced genomes that lack a corresponding full-length cDNA library.

Supplementary Material

Supplementary Material is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by National Science Foundation grants DBI-0434671 to D.G.S. and MCB-0342431 to M.L. We thank K. Wolfe and three anonymous referees for helpful comments that greatly improved the manuscript.

Literature Cited

Abril JF, Agarwal P, Alexandersson M, et al. (109 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.

Adams MD, Celniker SE, Holt RA, et al. (192 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science. 287:2185–2195.

Bärlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi O-P, Kallioniemi A. 2002. Cloning of *BCAS3* (17q23) and *BCAS4* (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes Chromosomes Cancer. 35:311–317.

Berget SM. 1995. Exon recognition in vertebrate splicing. J Biol Chem. 270:2411–2414.

Berget SM, Moore C, Sharp PA. 1977. Splices segments at the 5′ terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci USA. 74:3171–3175.

Blake CCF. 1978. Do genes-in-pieces imply proteins-in-pieces? Nature. 273:267.

Cai XL, Wang ZY, Xing YY, Zhang JL, Hong MM. 1998. Aberrant splicing of intron 1 leads to the heterogeneous 5′ UTR and decreased expression of *waxy* gene in rice cultivars of intermediate amylose content. Plant J. 14:459–465.

Carlo T, Sterner D, Berget S. 1996. An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. RNA (N Y). 2:342–353.

Carvalho AB, Clark AG. 1999. Intron size and natural selection. Nature. 401:344.

Castelli V, Aury JM, Jaillon O, et al. (11 co-authors). 2004. Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. Genome Res. 14:406–413.

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.

Cavalier-Smith T. 1985. Selfish DNA and the origin of introns. Nature. 315:283–284.

Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. Trends Genet. 7:145–148.

Cho G, Doolittle RF. 1997. Intron distribution in ancient paralogs supports random insertion and not random loss. J Mol Evol. 44:573–584.

Chow LT, Gelinas RE, Broker TR, Roberts RJ. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell. 12:1–8.

Comeron J, Kreitman M. 2000. The correlation between intron length and recombination in *Drosophila*: dynamic equilibrium between mutational and selective forces. Genetics. 156:1175–1190.

Darnell JE, Doolittle WF. 1986. Speculations on the early course of evolution. Proc Natl Acad Sci USA. 83:1271–1275.

de Souza SJ. 2003. The emergence of a synthetic theory of intron evolution. Genetica. 118:117–121.

de Souza SJ, Long M, Schoenbach L, Roy SW, Gilbert W. 1996. Intron positions correlate with module boundaries in ancient proteins. Proc Natl Acad Sci USA. 93:14632–14636.

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res. 27:3219–3228.

Dibb NJ. 1991. Proto-splice site model of intron origin. J Theor Biol. 151:405–416.

Dominiski Z, Kole R. 1991. Selection of splice sites in pre-messenger-RNAs with short internal exons. Mol Cell Biol. 11:6075–6083.

Dominiski Z, Kole R. 1992. Cooperation of pre-messenger-RNA sequence elements in splice site selection. Mol Cell Biol. 12:2108–2114.

Doolittle WF. 1978. Genes in pieces: were they ever together? Nature. 272:581–582.

Duret L. 2001. Why do genes have introns? Recombination might add a new piece to the puzzle. Trends Genet. 17:172–175.

Eden E, Brunak S. 2004. Analysis and recognition of 5′ UTR intron splice sites in human pre-mRNA. Nucleic Acids Res. 32:1131–1142.

Evans RM, Fraser N, Ziff E, Weber J, Wilson M, Darnell JE. 1977. The initiation sites for RNA transcription in Ad2 DNA. Cell. 12:733–739.

Fedorov A, Cao XH, Saxonov S, de Souza SJ, Roy SW, Gilbert W. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. Proc Natl Acad Sci USA. 98:13177–13182.

Fedorova L, Fedorov A. 2003. Introns in gene evolution. Genetica. 118:123–131.

Feiber AL, Rangarajan J, Vaughn JC. 2002. The evolution of single-copy *Drosophila* nuclear *4f-rnp* genes: spliceosomal intron losses create polymorphic alleles. J Mol Evol. 55:401–413.

Fink GR. 1987. Pseudogenes in yeast. Cell. 49:5–6.

Fong Y, Zhou Q. 2001. Stimulatory effect of splicing factors on transcriptional elongation. Nature. 414:929–933.

Frugoli JA, McPeek MA, Thomas TL, McClung CR. 1998. Intron loss and gain during evolution of the catalase gene family in angiosperms. Genetics. 149:355–365.

Gilbert W. 1978. Why genes in pieces? Nature. 271:501.

Gilbert W. 1987. The exon theory of genes. Cold Spring Harbor Symp. Quant Biol. 52:901–905.

Gilbert W, de Souza SJ, Long M. 1997. Origin of genes. Proc Natl Acad Sci USA. 94:7698.

Goldberg S, Schwartz H, Darnell JE Jr. 1977. Evidence from UV transcription mapping in HeLa cells that heterogeneous nuclear RNA is the messenger RNA precursor. Proc Natl Acad Sci USA 74:4520–4523.

Goss PJE, Lewontin RC. 1996. Detecting heterogeneity of substitution along DNA and protein sequences. Genetics. 143:589–602.

Gottlieb LD, Ford VS. 2002. The 5′ leader of plant *PgiC* has an intron: the leader shows both the loss and maintenance of constraints compared with introns and exons in the coding region. Mol Biol Evol 19:1613–1623.
Graur D, Li W-H. 2000. Fundamentals of molecular evolution. Sunderland (MA): Sinauer Associates, Inc.
Gregory TR. 2004. Insertion-deletion biases and the evolution of genome size. Gene. 324:15–34.
Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.
Kim VN, Kataoka N, Dreyfuss G. 2001. Role of the nonsense-mediated decay factor hUpf3 in the splicing-dependent exon-exon junction complex. Science. 293:1832–1836.
Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics. 47:713–719.
Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. Genetics. 49:725–738.
Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW. 2004. Introns and splicing elements of five diverse fungi. Eukaryotic Cell. 3:1088–1100.
Lander ES, Linton LM, Birren B et al. (271 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature. 409:860–921.
Larizza A, Makalowski W, Pesole G, Saccone C. 2002. Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyl and rodent gene pairs. Comput Chem. 26:479–490.
Lauderdale JD, Stein A. 1992. Introns of the chicken ovalbumin gene promote nucleosome alignment *in vitro*. Nucleic Acids Res. 20:6589–6596.
Le Hir H, Nott A, Moore MJ. 2003. How introns influence and enhance eukaryotic gene expression. Trends Biochem Sci. 28:215–220.
Liu K, Sandgren EP, Palmiter RD, Stein A. 1995. Rat growth hormone gene introns stimulate nucleosome alignment *in vitro* and in transgenic mice. Proc Natl Acad Sci USA. 92:7724–7728.
Luo MJ, Reed R. 1999. Splicing is required for rapid and efficient mRNA export in metazoans. Proc Natl Acad Sci USA. 96:14937–14942.
Lynch M. 2002. Intron evolution as a population-genetic process. Proc Natl Acad Sci USA. 99:6118–6123.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302:1401–1404.
Lynch M, Kewalramani A. 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. Mol Biol Evol. 20:563–571.
Lynch M, Richardson AO. 2002. The evolution of spliceosomal introns. Curr Opin Genet Dev. 12:701–710.
Lynch M, Scofield DG, Hong X. 2005. The evolution of transcription-initiation sites. Mol Biol Evol. 22:1137–1146.
MacArthur RH. 1957. On the relative abundance of bird species. Proc Natl Acad Sci USA. 43:293–295.
Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. Nature. 416:499–506.
Maquat LE. 2004a. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol. 5:89–99.
Maquat LE. 2004b. Nonsense-mediated mRNA decay: a comparative analysis of different species. Curr Genomics. 5:175–190.
Maroni G. 1994. The organization of *Drosophila* genes. DNA Seq. 4:347–354.
Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. Genome Biology. 3:reviews0004.0001–reviews0004.0010.
Mount SM, Burks C, Hertz G, Stormo GD, White O. 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res. 20:4255–4262.
Mourier T, Jeffares DC. 2003. Eukaryotic intron loss. Science. 300:1393.
Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci. 23:198–199.
Nissim-Rafinia M, Kerem B. 2002. Splicing regulation as a potential genetic modifier. Trends Genet. 18:123–127.
Nobile C, Marchi J, Nigro V, Roberts RG, Danieli GA. 1997. Exon-intron organization of the human dystrophin gene. Genomics. 45:421–424.
Nott A, Meislin SH, Moore MJ. 2003. A quantitative analysis of intron effects on mammalian gene expression. RNA (N Y). 9:607–617.
Orgel LE, Crick FHC. 1980. Selfish DNA—the ultimate parasite. Nature. 284:604–607.
Palmer JD, Logsdon JM Jr. 1991. The recent origins of introns. Curr Opin Genet Dev. 1:470–477.
Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, Gissi C, Saccone C. 2002. UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. Nucleic Acids Res. 30:335–340.
Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, Liuni S. 2001. Structural and functional features of eukaryotic mRNA untranslated regions. Gene. 276:73–81.
Proudfoot NJ. 2003. Dawdling polymerases allow introns time to splice. Nat Struct Biol. 10:876–878.
Ptak SE, Petrov DA. 2002. How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. Genetics. 162:1233–1244.
Ptashne M, Gann A. 2001. Transcription initiation: imposing specificity by localization. Essays Biochem. 37:1–15.
Rogozin IB, Kochetov AV, Kondrashov FA, Koonin EV, Milanesi L. 2001. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with 'weak' context of the start codon. Bioinformatics. 17:890–900.
Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol. 13:1512–1517.
Rose AB. 2004. The effect of intron location on intron-mediated enhancement of gene expression in *Arabidopsis*. Plant J. 40:744–751.
Roy SW, Gilbert W. 2005. Complex early genes. Proc Natl Acad Sci USA. 102:1986–1991.
Senapathy P. 1986. Origin of eukaryotic introns: a hypothesis, based on codon distribution statistics in genes, and its implications. Proc Natl Acad Sci USA. 83:2133–2137.
Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ. 2004. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res. 32:1774–1782.
Stapleton M, Carlson J, Brokstein P, et al. (10 co-authors). 2002. A *Drosophila* full-length cDNA resource. Genome Biology. 3:research0080.0081–research0080.0088.
Sterner D, Carlo T, Berget S. 1996. Architectural limits on split genes. Proc Natl Acad Sci USA. 93:15081–15085.
Sterner DA, Berget SM. 1993. *In vivo* recognition of a vertebrate mini-exon as an exon-intron-exon unit. Mol Cell Biol. 13:2677–2687.
Stoltzfus A, Logsdon JM, Palmer JD, Doolittle WF. 1997. Intron "sliding" and the diversity of intron positions. Proc Natl Acad Sci USA. 94:10739–10744.
Strausberg RL, Feingold EA, Grouse LH, et al. (79 co-authors). 2002. Generation and initial analysis of more than 15,000

full-length human and mouse cDNA sequences. Proc Natl Acad Sci USA. 99:16899–16903.
Vinogradov A. 1999. Intron-genome size relationship on a large evolutionary scale. J Mol Evol. 49:376–384.
Vinogradov A. 2002. Growth and decline of introns. Trends Genet. 18:232–236.
Wilkie GS, Dickson KS, Gray NK. 2003. Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. Trends Biochem Sci. 28:182–188.
Yu J, Yang Z, Kibukawa M, Paddock M, Passey DA, Wong GK-S. 2002. Minimal introns are not "junk". Genome Res. 12:1185–1189.
Zhang MQ. 2002. Computational prediction of eukaryotic protein-coding genes. Nat Rev Genet. 3:698–709.

Kenneth Wolfe, Associate Editor

Accepted September 7, 2006