The Evolution of Transcription-Initiation Sites

Michael Lynch, Douglas G. Scofield, and Xin Hong
Department of Biology, Indiana University

Unlike the situation in prokaryotes, most eukaryotic messenger RNAs contain a moderately long 5′ untranslated region (UTR). Such leader sequences impose a burden on eukaryotic genes by providing substrate for the mutational origin of premature translation-initiation codons, which generally result in defective proteins. To gain an insight into the expansion of 5′ UTRs in eukaryotic genomes, we present a simple null model in which the evolution of transcription-initiation sites is entirely driven by the stochastic mutational flux of core-promoter sequences and premature translation-initiation codons. This model yields results consistent with a variety of heretofore disconnected observations, including the form of length distributions of 5′ UTRs, the relatively low variance in UTR features among distantly related eukaryotes, the universal reliance on relatively simple core-promoter sequences, and the elevated density of introns in the 5′ UTR. We suggest that the reduced effective population sizes of most eukaryotes impose a population-genetic environment conducive to the movement of core promoters to random positions, subject to the constraint imposed by the upstream accumulation of premature translation-initiation codons. If this hypothesis is correct, then selection for gene-specific regulatory features need not be invoked to explain either the origin of lengthy eukaryotic 5′ UTRs or the 1,000-fold range of 5′-UTR lengths among genes within species. Nevertheless, once permanently established, expanded 5′ UTRs may have provided a novel substrate for the evolution of mechanisms for posttranscriptional regulation of eukaryotic gene expression. These results provide a further example of how an increase in the power of random genetic drift can passively promote the evolution of forms of gene architecture that ultimately facilitate the evolution of organismal complexity.

Introduction

The origin of eukaryotes was accompanied by numerous changes in gene and genome organization, including an expansion in gene number, the loss of operons, the colonization of protein-coding genes by introns, and the proliferation of mobile genetic elements (Lynch and Conery 2003). Associated with these changes were modifications in the ways in which genes are transcribed and processed into mature messenger RNAs (mRNAs). However, although we now know a great deal about protein-sequence evolution (Li 1997) and a fair amount about the temporal and spatial regulation of gene expression (Carroll et al. 2002; Davidson 2001), little attention has been given to the noncoding features of genes that are essential to the development of productive transcripts.

Eukaryotic gene expression is generally initiated by the recruitment of one or more transcription factors to multiple upstream regulatory elements, which help activate the basal transcription machinery in the vicinity of the correct transcription-initiation site (Ptashne and Gann 2002). This key event must be sufficiently accurate to insure that transcripts initiate upstream of the translation-initiation site, and elongation must proceed far enough to insure the incorporation of the translation-termination site. Both processes are generally guided by sequence information within the noncoding regions of genes, the refinements of which impose conflicting advantages and disadvantages.

Any benefits of elaborate transcriptional-control sequences must be weighed against the elevated mutation rate to defective alleles resulting from the increased size of the mutational target. For example, a long region between the point of transcription initiation and the translation-initiation site in the mature mRNA provides potential material for the evolution of mechanisms to finely tune gene expression at the level of mRNA localization, stabilization, and/or translation. However, a leader sequence on an mRNA also provides substrate for the mutational origin of inappropriate premature translation-start sites. In addition, although a lengthy localization signal for the transcription-initiation site will reduce the subversion of the transcription machinery to inappropriate locations, it also increases a gene's sensitivity to mutations that eliminate such elements.

This paper explores a number of central questions about the evolution of one particular feature of mRNAs, the transcribed but untranslated upstream region, i.e., the 5′ untranslated region (UTR). As with many other genomic features (Lynch and Conery 2003), it is unclear whether the noncoding regions of eukaryotic genes arose because the immediate selective advantages outweighed the negative mutational consequences. An alternative view is that nonadaptive processes play a key role in structuring eukaryotic gene organization in ways that provide novel physical substrate for the subsequent evolutionary refinement of gene-regulation mechanisms. The merit in exploring this alternative view is embodied in the following question. If the origin of eukaryotic gene structure awaited the appearance of rare beneficial mutations, then why do all well-studied prokaryotes harbor much simpler forms of gene organization, despite their enormous population sizes and greater opportunities for mutation? Resolution of this issue requires a consideration of the direction in which mutational processes will drive the evolution of gene architectural features when the power of natural selection is overwhelmed by genetic drift.

Background

Transcription in eukaryotic genes generally relies on a simple core-promoter sequence in the immediate vicinity of the transcription-initiation site. The core promoter binds the factors associated with the transcription preinitiation complex, which integrates information conveyed

Key words: Inr, genome evolution, TATA, transcription initiation, UTR, eukaryotes.

E-mail: mlynch@bio.indiana.edu.

Mol. Biol. Evol. 22(4):1137–1146. 2005
doi:10.1093/molbev/msi100
Advance Access publication February 2, 2005

Table 1
Average Lengths of 5′-UTR Sequences and CVs (ratio of the SD to the mean)

	Number of Genera	Mean (bp)	CV
Vertebrates			
Mammals	21	139 (19)	1.33 (0.09)
Birds	4	128 (18)	1.21 (0.15)
Frogs	4	103 (17)	1.14 (0.17)
Ray-finned fish	20	125 (7)	1.20 (0.08)
Agnathans	3	139 (11)	0.79 (0.11)
Invertebrates			
Invertebrate chordates	3	125 (24)	1.27 (0.20)
Echinoderms	4	206 (20)	1.09 (0.07)
Arthropoda	24	121 (12)	1.25 (0.09)
Mollusca	6	122 (30)	0.95 (0.11)
Annelida	2	76 (21)	0.77 (0.13)
Nematoda	8	67 (6)	1.15 (0.18)
Platyhelminthes	3	118 (17)	1.75 (0.59)
Porifera	3	85 (9)	0.95 (0.14)
Vascular plants	51	106 (4)	1.22 (0.06)
Unicellular species			
Fungi	17	149 (15)	1.22 (0.10)
Slime molds	2	107 (31)	1.17 (0.06)
Microbial chlorophytes	4	110 (19)	1.22 (0.25)
Apicomplexans	4	227 (32)	1.20 (0.14)
Kinetoplastids	2	163 (26)	1.43 (0.15)
Ciliates	2	126 (16)	1.80 (0.35)

NOTE.—Group-specific data are averages over all genera within the group for which data were available for at least 25 genes. Standard errors are given in parentheses. Statistics are derived from the database of Pesole et al. (2002).

by transcription factors (regulatory proteins) bound at more distant regulatory sites (Smale and Kadonaga 2003). By defining the point of transcription initiation, this process insures the production of a precursor mRNA with a leader sequence upstream of the translation-initiation codon. After the removal of introns during pre-mRNA processing, the retained leader comprises the 5′ UTR of the mature mRNA.

One of the most phylogenetically widespread transcription-initiation signals (hereafter, TISs) is the simple TATA sequence, which generally lies 20–40 nucleotides (nt) upstream of the transcription-initiation sites. The TATA box appears to be an indispensable orientation mechanism in archaeal transcription (Thomm 1996; Soppa 1999*a*, 1999*b*; Bell, Magill, and Jackson 2001; Slupska et al. 2001) and is employed in a substantial fraction of metazoan genes (~42% in *Drosophila* and ~32% in human; Kutach and Kadonaga 2000; Suzuki et al. 2001*a*; Ohler et al. 2002) and most yeast genes (Struhl 1989; Choi et al. 2002). Putative TATA boxes are also associated with the genes of *Paramecium* (Yamauchi et al. 1992), *Dictyostelium* (Hori and Firtel 1994), and *Entamoeba* (Singh et al. 1997). This wide phylogenetic distribution of TATA suggests that the stem eukaryote employed this sequence in transcription orientation.

Remarkably, only a few alternative TISs have been found in eukaryotes, and the only other element known to be self-sufficient is the loosely defined initiator (Inr) sequence, which contains an internal transcription-initiation site within a semiconserved sequence of 5–7 nt. With a consensus sequence of Py-Py-A-N-(T/A)-Py-Py in metazoans (Lo and Smale 1996), with Py denoting pyrimidine and N an arbitrary nucleotide, the complexity of the Inr is even less than that of TATA. In *Drosophila* and humans, 65%–85% of all protein-coding genes employ an Inr sequence (Suzuki et al. 2001*a*; Ohler et al. 2002), and Inr is also used by a number of protists, including *Toxoplasma gondii* (Nakaar et al. 1998), *Trichomonas vaginalis* (Liston and Johnson 1999), and *Entamoeba histolytica* (Singh et al. 1997). A self-sufficient Inr is not known in yeast, although significant information can be encoded at the site of transcription initiation (Hahn, Hoar, and Guarente 1985; Nagawa and Fink 1985). Nevertheless, the wide phylogenetic distribution of Inr elements, the rough similarity of the sequences employed in different eukaryotic lineages, and the utilization of an Inr (albeit highly divergent) in archaea suggests that Inr, like TATA, may have been utilized by the stem eukaryote and then lost from some descendant lineages.

There are substantial differences in the positioning of TISs in prokaryotes and eukaryotes. Eubacterial core promoters are generally located close enough to the translation-initiation codon to yield 5′ UTRs that are <12 base pair (bp) in length. Eubacterial genes are also often aggregated into operons that are transcribed as continuous (polycistronic) mRNAs, with each gene being preceded by a Shine-Delgarno sequence that binds the 16S small ribosomal RNA subunit and directs it to the translation-start codon (Shine and Dalgarno 1974). In the archaea, genes internal to operons also often have Shine-Delgarno sequences, but the leader genes often do not (Tolstrup et al. 2000; Slupska et al. 2001), and in the species that has been investigated in most detail, *Pyrobaculum aerophilum*, almost all mRNAs are leaderless, i.e., the transcription- and translation-initiation sites are identical (Slupska et al. 2001), a feature that has also been found in a few eubacteria (Weiner, Herrmann, and Browning 2000; Moll et al. 2002).

In contrast, the average lengths of 5′ UTRs in eukaryotes range from just under 100 bp to just over 200 bp, and remarkably, there are no obvious differences between unicellular and multicellular species (table 1). The narrow approximately threefold range in average 5′-UTR lengths among eukaryotic lineages and its independence of organismal complexity is in striking contrast to the order-of-magnitude differences that exist among eukaryotes for average numbers and sizes of introns, numbers of mobile elements, and lengths of nontranscribed intergenic spacers (Lynch and Conery 2003).

Because the lengths of individual 5′ UTRs within species can range from tens to thousands of base pairs (Radford and Parish 1997; Pesole et al. 2000, 2001; Suzuki et al. 2000; Rogozin et al. 2001) (fig. 1), the relative invariance of the average length does not appear to be due to a structural constraint. In the diplomonad *Giardia*, the 5′ UTR often consists of just a single nucleotide (Iwabe and Miyata 2001), and although the data are limited, a substantial fraction of 5′ UTRs in a variety of other unicellular eukaryotes have been found to be <25 bp, including those in the ciliate *Euplotes crassus* (Ghosh et al. 1994), the amoeba *E. histolytica* (Singh et al. 1997), and the trichomonad *T. vaginalis* (Liston and Johnson 1999). Furthermore, experimental evidence suggests that such diminutive leader sequences are sufficient to support translation in mammals and yeast,

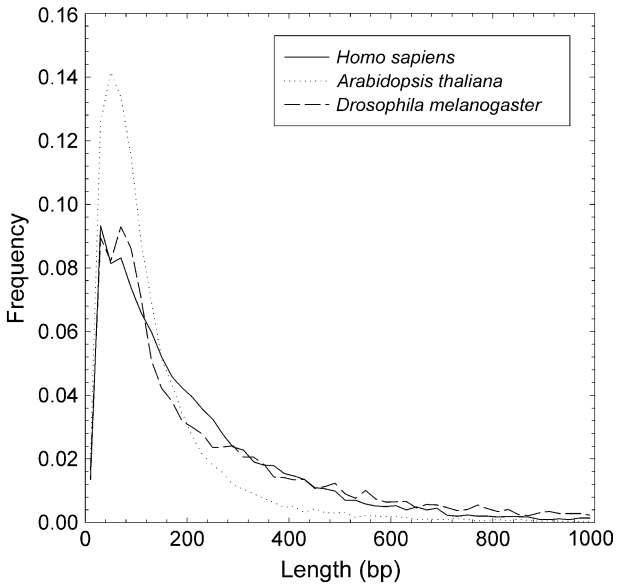

FIG. 1.—Frequency distribution of lengths of the 5′ UTR derived from the database of Pesole et al. (2002). Frequencies are reported in 20-bp bins. Sample sizes are 28,152 for *Homo*, 13,966 for *Arabidopsis*, and 10,548 for *Drosophila*.

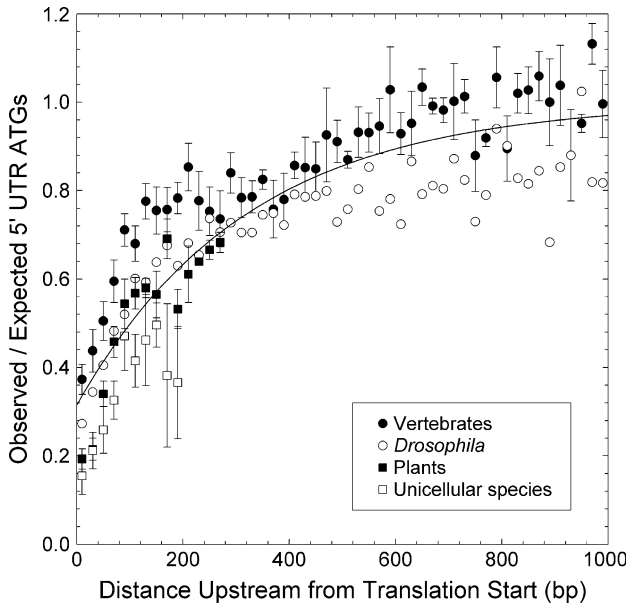

FIG. 2.—The fraction of observed AUG triplets relative to the random expectation in 20-bp intervals increasingly upstream of the true translation-start site derived from the database of Pesole et al. (2002). Vertebrates include *Bos*, *Danio*, *Gallus*, *Homo*, *Mus*, *Macaca*, *Oncorhynchus*, *Rattus*, *Sus*, and *Xenopus*; plants include *Arabidopsis*, *Brassica*, *Glycine*, *Hordeum*, *Lycopersicon*, *Medicago*, *Nicotiana*, *Oryza*, *Pisum*, *Solanum*, *Triticum*, and *Zea*; and unicellular species include *Chlamydomonas*, *Dictyostelium*, *Saccharomyces*, *Schizosaccharomyces*, and *Trypanosoma*. All data points are based on 50 or more estimates. Standard errors are based on results for different genera within taxonomic groups. The solid line denotes the incidence pattern employed in simulations described in the text, evaluated by least-squares regression, $y = 1 - e^{-0.0031(x+121)}$, $r^2 = 0.78$. We observed similar trends within UTR exons from complete full-length cDNA libraries for human, *Drosophila*, and *Arabidopsis* (Stapleton et al. 2002; Strausberg et al. 2002; Castelli et al. 2004).

although the efficiency of translation can be reduced with UTRs shorter than ~30 bp (van den Heuvel et al. 1989; Maicas, Shago, and Friesen 1990; Hughes and Andrews 1997). These observations suggest that a 5′ UTR is not an essential landing platform for the ribosome in eukaryotes. Indeed, no eukaryote is known to employ a Shine-Delgarno sequence for translation-start site localization, although some eukaryotic mRNAs do harbor sequences with complementarity to rRNAs that may function in translation control (Tranque et al. 1998; Mauro and Edelman 2002).

For eukaryotes, lengthy 5′ UTRs impose a mutational burden on their associated alleles by enhancing the rate of acquisition of a premature translation-start codon (PSC). Such nucleotide triplets are problematical because eukaryotic translation initiation generally proceeds by scanning from the 5′ end of a transcript until the first AUG is encountered, although the selection of a specific initiation site is sometimes influenced by a small surrounding set of nucleotides (Kozak 1987, 1994; Pesole et al. 2000; Niimura et al. 2003). If a 5′ UTR acquires an AUG triplet in an inappropriate context, there is a high likelihood that premature translation initiation will result in a nonfunctional allele because two-thirds of random AUGs will be out-of-frame with respect to the downstream coding sequence, and even when in frame, the addition of N-terminal amino acids may lead to a defective protein. In a variety of species, 15%–55% of 5′ UTRs contain upstream AUGs (Suzuki et al. 2000; Peri and Pandey 2001; Rogozin et al. 2001), about equally distributed among the three reading frames and often in a context that satisfies the Kozak criteria (~60% in the case of humans; Suzuki et al. 2000), so it is clear that not all PSCs result in gene malfunction. Nevertheless, several human genetic disorders are known to result from mutations to PSCs (Kozak 2002), and all species have a reduced incidence of PSCs relative to random expectations. Taken over the entire length of the 5′ UTR, the latter ratio is ~38% in human, ~42% in rodents, ~50% in *Drosophila*, ~60% in plants, and ~53% in fungi (Rogozin et al. 2001), implying that a substantial fraction of mutant alleles with spurious PSCs are selected against. There is, however, a strong distance-dependent gradient of the deficit of PSCs, with the bias being negligible >500 bp upstream of the translation-start site (fig. 2). That translation-related selection is the primary cause of this bias is evident from the incidence of AUGs in external introns located within the gradient (fig. 3). Although there is a slight depression in intronic AUGs in the 5′ UTR, there is no discernable gradient to the pattern. In prokaryotes, such deficits only extend 10–30 bp upstream of the translation-start point, i.e., just up to the location of the Shine-Delgarno sequence (Saito and Tomita 1999).

A Null Model for 5′-UTR Lengths

The preceding results highlight three general observations about eukaryotic 5′ UTRs that require explanation: (1) the universal use of short TISs; (2) the L-shaped distributions of 5′-UTR lengths; and (3) the phylogenetic constancy in the average lengths of 5′ UTRs. To evaluate the potential influence of the recurrent mutational origin of PSCs on these features, we will consider a simple null model for the evolution of 5′-UTR lengths. This model

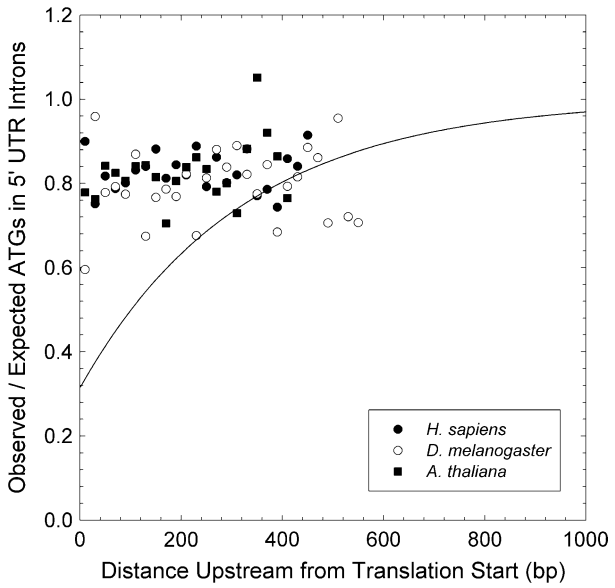

FIG. 3.—The fraction of observed AUG triplets relative to the random expectation in 20-bp intervals within external introns increasingly upstream of the true translation-start site. Sample sizes for all data points are at least 25, with the locations of individual introns being defined with respect to the position within the processed UTR. The solid line is the within-exons reference from figure 2. Intron locations and sequences were determined via alignment of full-length cDNA transcripts (Stapleton et al. 2002; Strausberg et al. 2002; Castelli et al. 2004) with complete genomic sequences using BLAT (Kent 2002) and custom software tools.

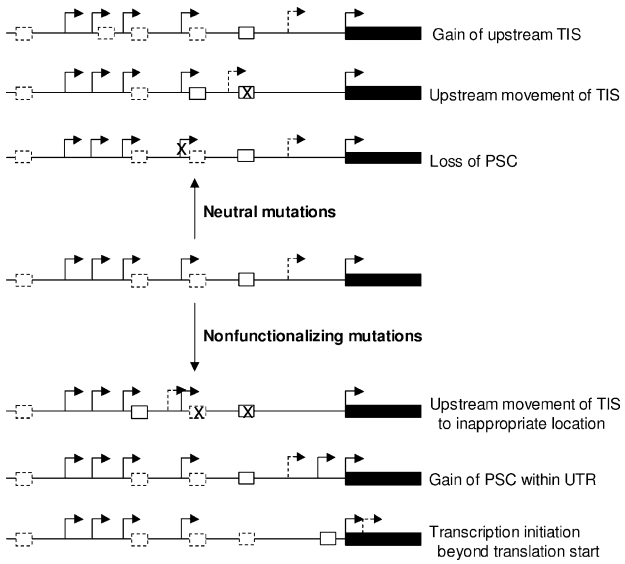

FIG. 4.—Schematic for some alternative allelic types associated with mutations in the 5′ UTR. Solid arrows denote potential translation-initiation sites; those at the beginning of the solid box denote the true translation-start codon, with upstream arrows denoting latent PSCs. Small open boxes denote TISs (the dashed boxes denoting latent TISs, and solid boxes denote currently used TISs). Dashed arrows denote active transcription-initiation sites.

assumes that the transcription apparatus employs as a TIS the nearest appropriate sequence upstream of the translation-initiation site and allows potential PSCs and TISs to appear and disappear stochastically by mutation. Flexibility in the precise locations of TISs is supported by the empirical observation that the elimination of a functional TIS can be compensated by alternative matching sequences in locations sufficient to sustain transcription initiation (Hahn, Hoar, and Guarente 1985; Nagawa and Fink 1985). Under this model, functional alleles require the absence of a harmful PSC within the UTR, but PSCs are free to accumulate in the genomic DNA upstream of the currently utilized transcription-initiation site. Such behavior results in a gradient of increasing density of PSCs upstream of a gene's translation-initiation site, thereby producing a barrier to the upstream movement of a TIS by mutational processes while allowing downstream movement. The recurrent gain and loss of PSCs and TISs by mutational processes yields a steady-state stochastic distribution of 5′-UTR lengths.

Because of the multisite nature of the problem, arriving at an analytical solution for the steady-state distribution is difficult, but the entire process is readily modeled by computer simulation. Each generation, an allele is subject to several types of mutations in its upstream region (fig. 4). First, PSCs that are neutrally maintained upstream of the current transcription-initiation site are each subject to elimination by mutation at rate 3μ, where μ is the per-nucleotide mutation rate and the 3 accounts for the 3 nt in ATG. Second, all potential TISs are subject to mutational loss at rate $n\mu$, where n is the number of nucleotides in the TIS (e.g., for TATA, $n = 4$). All such mutant alleles are fully viable, unless the mutation involves the currently utilized TIS. In the latter case, because the UTR expands up to the transcription-initiation site associated with the nearest upstream alternative TIS, a nonfunctional allele will result if the extended region contains a harmful PSC. Third, PSCs are mutationally acquired at the rate $3\mu/64$ per nucleotide site in the upstream region of a gene (the 3 accounting for all three reading frames, and 1/64 being the fraction of random triplets that are ATG). Such newly arisen PSCs are neutral unless they fall within the current UTR, in which case they have the potential to yield a nonfunctional allele (below). Fourth, new potential TISs arise at the rate $n\mu/(4^n)$ per nucleotide site. Such mutant alleles are always fully viable, unless the TIS is so close to the translation-start site that transcription initiates downstream of the translation-start codon. Should a new TIS arise sufficiently upstream of the translation-start site but downstream of the previously employed TIS, the UTR will take on a shorter length.

Under the assumption that all viable alleles have equal fitness and that defective alleles are rapidly eliminated by selection, the steady-state stochastic distribution of UTR lengths can be obtained by simply following the evolution of a single allelic lineage over time. This distribution can be viewed as either the long-term distribution of UTR lengths of an individual gene or the expected snapshot genomic distribution for all genes under similar constraints. The following simulation results are generally based upon at least 4×10^{11} generations with a recording made every 10^6 generations, with $\mu = 10^{-7}$ per base pair per generation, a 30-bp distance from the TIS to the site of transcription initiation (as in TATA), and an empirical function for the probability that a PSC is potentially harmful based on the data in figure

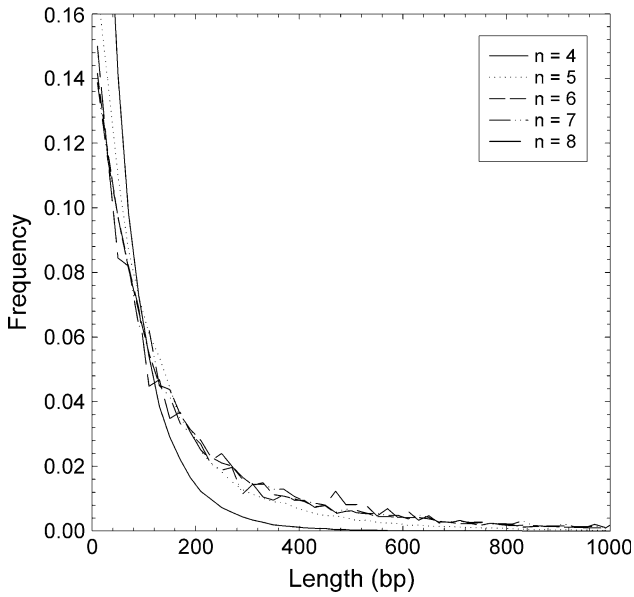

FIG. 5.—Expected steady-state frequency distributions for 5′-UTR lengths under a model in which TISs and PSCs randomly appear and disappear by mutational processes, as described in the text. Results are given for TISs of various lengths in base pair (denoted by *n*). Frequencies are given for bin widths of 20 bp.

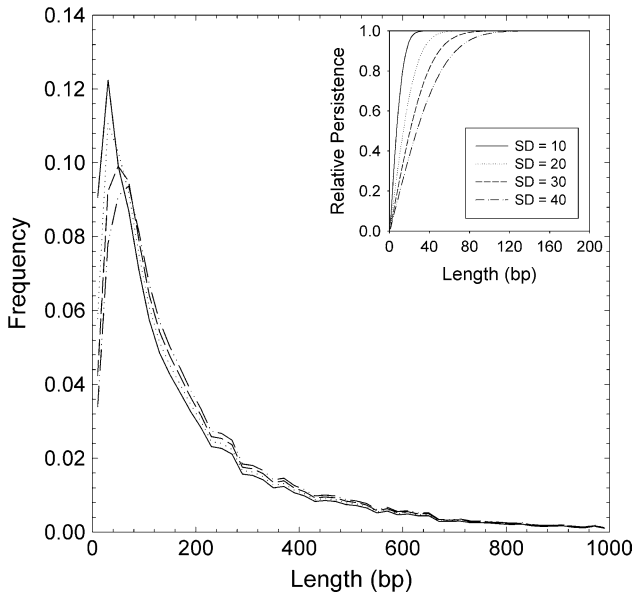

FIG. 6.—Expected steady-state frequency distributions for 5′-UTR lengths under a model in which TISs and PSCs randomly appear and disappear by mutational processes, with selection operating against overly short UTRs. The curves in the main graph are obtained by weighting the results in figure 5 for $n = 4$–7 by the persistence functions provided in the inset. The latter functions are scaled from the areas of the left tails of normal distributions, as described in the text.

2. Because eukaryotic genes can usually employ the two alternative TISs, TATA and Inr, the birth rate of TISs is assumed to be twice that noted above.

The amount of information necessary for localizing the point of transcription initiation is not entirely clear. However, given the widespread use of TATA and Inr elements, *n* is at least 4, and the presence of putative conserved accessory elements implies that *n* could be as large as 10 in some cases (Singh et al. 1997; Soppa 1999*a*, 1999*b*; Gourse, Ross, and Gaal 2000; Ohler et al. 2002). Fortunately, the predictions of the null model are quite robust to uncertainty in *n*. The length distributions of 5′ UTRs obtained under this model are strongly L-shaped in all cases, with means of 66 and 128 and CVs (coefficients of variation, ratios of standard deviations [SD] to means) of 1.08 and 1.18 for $n = 4$ and 5, respectively (fig. 5). For $n = 6$ and higher, the steady-state length distribution is nearly independent of *n*, with an overall mean of ~190 bp and CV of ~1.27. Thus, the first two moments generated by this simple model are qualitatively consistent with observations in multiple genomes—for vertebrates, invertebrates, plants, and unicellular species, 5′-UTR lengths have means of 127, 115, 106, and 147 and CVs of 1.13, 1.15, 1.22, and 1.34, respectively. Moreover, as can be seen by comparing figures 1 and 5, the right sides of the observed and expected distributions are very similar.

The primary distinction between the predictions of the null model and observed 5′-UTR length distributions involves the deficit of observed UTRs shorter than 50 bp. Although this discrepancy might be a result of the selective elimination of inefficiently translated alleles with 5′ UTRs shorter than a few dozen base pairs, at least two other factors may also contribute. First, if the point of transcription initiation is not at a fixed position downstream of the TIS, some transcripts would inadvertently initiate downstream of the translation-start site. Although few attempts have been made to quantify fine-scale differences in the features of transcript leaders, noisy transcription-start sites are known to occur in eubacteria (Nicolaides et al. 1995), plant mitochondria (Lizama, Holuigue, and Jordana 1994; Carrodeguas and Vallejo 1997), and nuclear genes of metazoans (Bergsma et al. 1996; Yu et al. 1996; Kaji et al. 1998; Weiner, Herrmann, and Browning 2000; Suzuki et al. 2001*b*) and unicellular species (Hahn, Hoar, and Guarente 1985; Nagawa and Fink 1985; Watanabe et al. 2002). Ranges of transcription-start sites on the order of 10–100 bp are not uncommon, and in humans, where the data are most extensive, many genes have a continuum of transcription-initiation sites, with an average range of 62 bp and SD of 20 bp (Suzuki et al. 2001*b*).

Provided the sites of transcription initiation were normally distributed with a SD of 20 bp, 16% of transcripts from a TIS positioned to yield an average 20-bp 5′ UTR would be unproductive and the same would be true for any TIS with a CV of actual UTR lengths equal to 1. Similarly, for any TIS with a CV = 0.5, 3.3% of transcripts would initiate downstream of the translation-start site. As a rough approximation of the influence of selection on the distribution of 5′-UTR lengths, we will consider the left tail of a normal distribution to be a measure of the disadvantage of an overly short UTR, letting the relative persistence time of an allele scale linearly with the area of the right (positive) side of the distribution, with areas of 0.5 and 1.0 equating to relative persistences of 0.0 and 1.0, respectively. As shown in figure 6, such weighting transforms the left side of the length distribution to one fairly

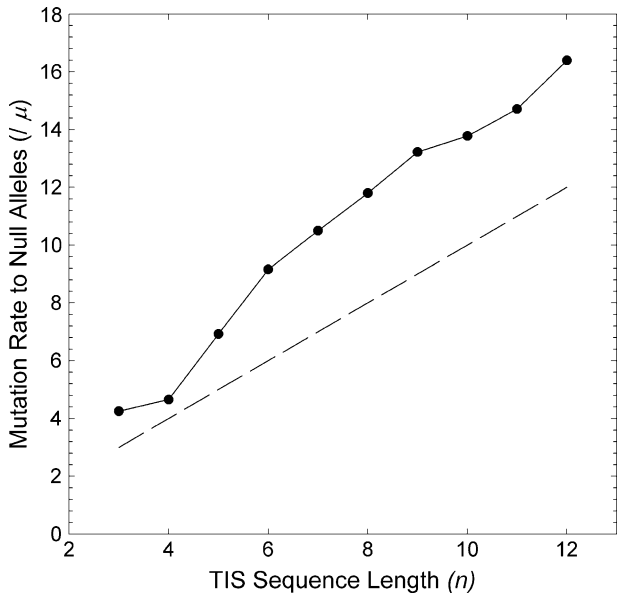

FIG. 7.—The relationship between the steady-state mutation rate to null alleles and the complexity of the TIS, *n*. The actual mutation rate can be obtained by multiplying the plotted values by the mutation rate per nucleotide site, μ. The dashed line denotes the mutation rate to nulls with a 5′ UTR of length zero, assuming no sloppy transcription-initiation sites. The horizontal distance between the two functions is equal to the excess mutation rate of a lineage of alleles with stochastically evolving 5′-UTR lengths relative to those with zero length.

similar to that actually observed, while leaving the right side largely unaltered.

A second complication not directly accounted for by the preceding model concerns the presence of (external) introns in the 5′ UTR. By definition, a TIS must lie upstream of all introns, but because introns are removed prior to translation, they are free to accumulate ATG triplets, as noted above. In principle, a 5′-UTR intron is free to be lost, as it contributes nothing to the mature mRNA. However, the potential inclusion of ATGs within such an intron may inhibit its loss in at least two ways. First, if a new TIS arises within a 5′-UTR intron, the previous 5′ intronic splice site will no longer be contained within the pre-mRNA, which will result in the incorporation of the remaining downstream intronic sequence (between the new TIS and the previous 3′ splice site) into the 5′ UTR of the mature mRNA. A nonfunctional allele would then result should the added region contain a harmful PSC. Second, any direct mutational loss of either intronic splice site will result in the incorporation of the full intron into the 5′ UTR, yielding an especially high likelihood of introduction of a PSC. In this sense, once established, external introns are expected to be exceptionally stable. This argument is supported by the fact that per physical distance, introns are about three times more abundant in 5′ than in 3′ UTRs in a wide variety of species (Hawkins 1988; Pesole et al. 2001). The preceding model assumes that such reinforcement effects are strong enough to reduce the origin of new TISs within external introns to a negligible level.

The simulation results also suggest a reason for the universal use of relatively short TISs. Due to the stochastic expansion and contraction of UTR lengths, the total mutation rates within individual UTRs can be expected to vary through time, but in the long run, an allelic lineage with a specific level of complexity for the TIS can be characterized by an average mutation rate to null alleles. In particular, a greater n is expected to magnify the mutation rate to null alleles by increasing the size of the TIS itself and also by increasing the average interval between alternative TISs. The exact relationship between the null mutation rate and n is a function of the degree to which PSCs and alternative TISs build up in a particular genetic background, but the results suggest an approximately linear scaling (fig. 7).

To this point, we have assumed that the length distribution of 5′ UTRs for an allelic lineage is simply determined by the stochastic mutational turnover of TISs and PSCs, which ignores the possibility of selection promoting alleles with TISs closer to the translation-initiation site on the basis of their reduced mutation rate. To examine this issue, it is instructive to consider the extreme contrast between an allele with the TIS at the average random distance from the translation-initiation site and an allele for which the two sites coincide exactly (i.e., a 5′-UTR length of zero). In the latter case, the mutation rate to nulls associated with the 5′ UTR is simply $n\mu$, the mutation rate of the TIS (ignoring any sloppiness in the point of transcription initiation). Drawing from the results in the preceding paragraph, an allele with zero-length 5′ UTR would have an advantage in terms of a reduction in the mutation rate (s) of just 0.6μ when $n = 4$, 3.2μ when $n = 6$, and 3.8μ when $n = 8$. Because weak differential mutation rates to nulls operate like selection with additive allelic effects on fitness (Lynch 2002), the diffusion approximation for the behavior of a newly arisen mutation (Kimura 1962) can be used to evaluate this problem. For a randomly mating diploid population, the probability of fixation of an advantageous allele relative to the neutral expectation is $\Theta \simeq 4N_e s/(1 - \exp[-4N_e s])$, where N_e is the effective population size, which yields $\Theta = 1.02$, 1.21, and 4.07 for $N_e s = 0.01$, 0.10, and 1.00, respectively. Thus, as a first approximation, for selection to have a significant effect on the distribution of 5′-UTR lengths, $N_e\mu$ must exceed 0.053 if $n = 5$, 0.031 if $n = 6$, and 0.026 if $n = 8$.

Estimates of $N_e\mu$ derived from measures of segregating nucleotide variation at silent sites of protein-coding genes (Lynch and Conery 2003) provide insight into these issues. For all multicellular species of eukaryotes for which sufficient data are available, estimates of $N_e\mu$ are smaller than 0.005, whereas a few unicellular species have $N_e\mu$ in the range of 0.01–0.02. In contrast, most well-characterized prokaryotic species have $N_e\mu \geq 0.01$, with several having $N_e\mu$ in the range of 0.04–0.15, and a few even higher. These observations suggest that the effective population sizes of most multicellular eukaryotes are far too small to enable selection to have a significant influence on the distribution of 5′-UTR lengths. However, because the estimates of $N_e\mu$ for unicellular species are likely to be downwardly biased by selection operating on silent sites, prokaryotes may commonly experience weak-enough levels of random genetic drift to allow the selective promotion of alleles with UTRs shorter than those predicted by the null model.

Discussion

These results support the idea that the features of 5′ UTRs in most multicellular, and probably a wide range of unicellular, eukaryotes are largely dictated by random genetic drift and mutational processes that cause stochastic turnover in transcription-initiation sites and premature start codons. Under the simplest model that we present, natural selection only indirectly influences the lengths of UTRs through the mutational origin of premature initiation codons within the UTR. If this hypothesis is correct, selection for gene-specific regulatory features need not be invoked to explain the 1,000-fold range of 5′-UTR lengths among genes within species. The broad distribution of UTR lengths with a long tail to the right (fig. 1) is expected to arise via mutational processes alone, and the observed within-species CVs in UTR lengths are also consistent with the robust theoretical prediction of 1.2–1.3.

An attractive feature of the proposed theory is the insensitivity of the models predictions to the actual length of TISs, which might vary among species. Most notably, once n exceeds 5, there is a near invariance in the steady-state distribution of 5′-UTR lengths predicted by the model, despite the fact that the average distance between random TISs is an increasing function of n. This asymptotic behavior results from the strong barrier to upstream movement of TISs imposed by the neutral accumulation of potentially harmful upstream PSCs as well as by the vulnerability of overly long UTRs to the mutational elimination by the appearance of PSCs within their confines.

The one incompatibility between the data and the predicted 5′-UTR length distribution is the shift of the observed distributions to the right by ~30–50 nt, which suggests that the very shortest size classes may be selected against by forces other than PSCs. As discussed above, some such selection is expected to result from stochastic variation in points of transcription initiation operating at the cellular level, which would have deleterious consequences for genes with TISs close enough to the translation-start codon to occasionally initiate transcription beyond the translation-start site.

A second potential complication not incorporated into the model concerns the presence of (external) introns in the 5′ UTR, which may inhibit the mutational production of viable alleles with short UTRs. As noted above, the accumulation of PSCs may render such introns exceptionally stable. In the event that a mutational event produces a TIS within a 5′-UTR intron that does not contain a harmful downstream PSC, a new successful allele will be produced, with the upstream portion of the 5′ UTR of the ancestral gene being eliminated and the downstream portion of the intron acquiring the new TIS being incorporated. Given the average sizes of eukaryotic 5′ UTRs (this paper) and the average sizes of introns (Lynch and Conery 2003), the net effect of these potential changes will often be an overall increase in the length of the UTR. The mere presence of introns may also inhibit the evolution of short 5′ UTRs for purely structural reasons, as there appears to be a minimal exon size essential for efficient splicing (Sterner, Carlo, and Berget 1996).

Because short TISs magnify the chances of the transcriptional apparatus being subverted to an inappropriate (false positive) site, TISs of at least moderate complexity would seem to be required for efficient transcription. For both eubacterial and archaeal genomes, the efficiency of natural selection is sufficient to maintain the number of spurious core promoters at levels below random expectations (Hahn, Stajich, and Wray 2003). That inappropriate utilization of random TISs actually occurs in eukaryotes is suggested by the fact that ~25% of human cDNAs contain no obvious open reading frame and are largely derived from AT-rich genomic regions likely to harbor spurious TATA sequences (Ota et al. 2004). Nearly 10 times more genomic DNA is transcribed than can be accounted for by known exons (Kapranov et al. 2002), and similar observations have been made in *Giardia* (Elmendorf, Singer, and Nash 2001). Many of these noncoding RNAs are from the antisense strand, overlap the exons of coding DNA (Cawley et al. 2004), and could have a function, but that remains to be determined.

This being said, n need not be very large to minimize the chances of false positives. Under the assumption of equal nucleotide frequencies, the expected distance between random sequences of n nucleotides is 4^n bp, so a genome 10^8 bp in length (e.g., an average invertebrate) would contain ~24,400 TISs of length $n = 6$, ~1,530 of length $n = 8$, and only ~95 of length $n = 10$. Thus, because the number of genes per eukaryotic genome usually exceeds 10^4, the length of a TIS need not be much greater than eight to insure that nearly all such sequences are actively maintained by selection in the vicinity of functional genes. Any further increase in n would impose a higher mutation rate to defective alleles without providing any obvious benefits in terms of transcriptional efficiency. With μ being $\sim 10^8$ per generation (Denver et al. 2004), and ~15,000 genes in a typical eukaryotic genome, the preceding results suggest that just ~0.2% of gametes would carry a new 5′-UTR associated null allele at some locus if $n = 8$. This level of null mutation is easily accommodated by existing estimates of the gametic lethal mutation rate of ~1.0% (Lynch and Walsh 1998; Fry et al. 1999).

Although the variation among phylogenetic groups in the average lengths of 5′-UTR lengths is exceptionally small relative to other genomic attributes, some significant lineage-specific differences may exist (it should be noted though that the standard errors in table 1 are not corrected for phylogenetic nonindependence). A number of second-order effects could be responsible for such differences. First, in deriving the theoretical expectations, it was assumed that all 4 nt are equally likely, whereas most eukaryotic genomes deviate somewhat from these conditions. However, although the average lengths of 5′ UTRs may be expected to scale negatively with the random expected frequencies of TATA and ATG sequences within UTR regions, no such relationship exists in the observed data ($r^2 = 0.08$ and 0.01, respectively). Second, for reasons discussed above, lineage-specific variation in the frequency and/or numbers of external introns could impose different constraints on the indirect selective pressures toward shorter UTRs. Third, the model developed above considers only nucleotide-substitution mutations, whereas significant interspecific differences may exist in rates of deletion and/or insertion (Petrov et al. 2000).

In summary, our empirical and theoretical results support the idea that the reduction in N_e that accompanied the evolution of eukaryotes, particularly multicellular species, produced a population-genetic environment conducive to the movement of TISs to random positions, subject only to the constraint imposed by the stochastic mutational production of premature start codons. Some microbial eukaryotes may have large-enough effective population sizes to selectively maintain average 5′-UTR lengths below the expectations under effective neutrality (Ghosh et al. 1994; Singh et al. 1997; Liston and Johnson 1999; Yee et al. 2000; Adam 2001), with virtually all such species exhibiting other genomic hallmarks of large N_e, including small sizes and numbers of introns and a low incidence of mobile genetic elements (Lynch and Conery 2003). However, essentially all multicellular species may have an N_e insufficiently large to prevent the physical expansion of 5′ UTRs by nonadaptive mechanisms. If this hypothesis is correct, selection for gene-specific regulatory features need not be invoked to explain the expansion of eukaryotic 5′ UTRs relative to the situation in prokaryotes. Nevertheless, once permanently established, expanded 5′ UTRs may have provided novel substrate for the evolution of mechanisms for posttranscriptional regulation of eukaryotic gene expression, providing still another example of how a reduction in N_e can passively promote the evolution of novel forms of gene architecture that ultimately facilitate the evolution of organismal complexity (Lynch et al. 2001; Lynch 2002; Lynch and Conery 2003).

Even here, there is room for caution. Although many structural features of 5′ UTRs (including their lengths) are known to influence the rate of protein synthesis by modifying the efficiency of translation, prior to accepting natural selection as the underlying explanation for such features, further consideration of semineutral processes may prove worthwhile. For example, upstream open-reading frames (uORFs) can slow the rate of translation by causing the ribosome to terminate and/or reinitiate, and secondary UTR structure and/or internal ribosome entry sites may have similar indirect roles. A number of authors have suggested that uORFs serve an adaptive function (Morris and Geballe 2000; Meijer and Thomas 2002; Vilela and McCarthy 2003). However, uORFs are generally on the order of 20 codons in length, approximately what is expected by chance, and transcripts from some uORF-containing genes are subject to degradation by the nonsense-mediated decay pathway (Ruiz-Echevarria and Peltz 2000). In principle, a number of uORFs may simply exist because their stop codons have neutralized the effects of a PSC, enabling their carrier alleles to perform at normal levels.

Acknowledgments

This work was supported by NSF grant MCB-0342431 to M.L. and NSF grant DBI-0434671 to D.G.S.

Literature Cited

Adam, R. D. 2001. Biology of *Giardia lamblia*. Clin. Microbiol. Rev. **14**:447–475.

Bell, S. D., C. P. Magill, and S. P. Jackson. 2001. Basal and regulated transcription in Archaea. Biochem. Soc. Trans. **29**: 392–395.

Bergsma, D. J., Y. Ai, W. R. Skach, K. Nesburn, E. Anoia, S. Van Horn, and D. Stambolian. 1996. Fine structure of the human galactokinase GALK1 gene. Genome Res. **6**:980–985.

Carrodeguas, J. A., and C. G. Vallejo. 1997. Mitochondrial transcription initiation in the crustacean *Artemia franciscana*. Eur. J. Biochem. **250**:514–523.

Carroll, S. B., J. K. Grenier, and S. D. Weatherbee. 2002. From DNA to diversity. Blackwell Science, Malden, Mass.

Castelli, V. et al. 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.

Cawley, S. et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell **116**:499–509.

Choi, W. S., M. Yan, D. Nusinow, and J. D. Gralla. 2002. *In vitro* transcription and start site selection in *Schizosaccharomyces pombe*. J. Mol. Biol. **319**:1005–1013.

Davidson, E. H. 2001. Genomic regulatory systems. Academic Press, San Diego, Calif.

Denver, D. R., K. Morris, M. Lynch, and W. K. Thomas. 2004. High mutation rate and insertion predominance in the *Caenorhabditis elegans* genome. Nature **430**:679–682.

Elmendorf, H. G., S. M. Singer, and T. E. Nash. 2001. The abundance of sterile transcripts in *Giardia lamblia*. Nucleic Acids Res. **29**:4674–4683.

Fry, J. D., P. D. Keightley, S. L. Heinsohn, and S. V. Nuzhdin. 1999. New estimates of the rates and effects of mildly deleterious mutation in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **96**:574–579.

Ghosh, S., J. W. Jaraczewski, L. A. Klobutcher, and C. L. Jahn. 1994. Characterization of transcription initiation, translation initiation, and poly(A) addition sites in the gene-sized macronuclear DNA molecules of *Euplotes*. Nucleic Acids Res. **22**:214–221.

Gourse, R. L., W. Ross, and T. Gaal. 2000. UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. Mol. Microbiol. **37**:687–695.

Hahn, M. W., J. E. Stajich, and G. A. Wray. 2003. The effects of selection against spurious transcription factor binding sites. Mol. Biol. Evol. **20**:901–906.

Hahn, S., E. T. Hoar, and L. Guarente. 1985. Each of three ''ATA elements'' specifies a subset of the transcription initiation sites at the CYC-1 promoter of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **82**:8562–8566.

Hawkins, J. D. 1988. A survey on intron and exon lengths. Nucleic Acids Res. **16**:9893–9908.

Hori, R., and R. A. Firtel. 1994. Identification and characterization of multiple A/T-rich cis-acting elements that control expression from *Dictyostelium* actin promoters: the *Dictyostelium* actin upstream activating sequence confers growth phase expression and has enhancer-like properties. Nucleic Acids Res. **22**:5099–5011.

Hughes, M. J., and D. W. Andrews. 1997. A single nucleotide is a sufficient 5′ untranslated region for translation in an eukaryotic in vitro system. FEBS Lett. **414**:19–22.

Iwabe, N., and T. Miyata. 2001. Overlapping genes in parasitic protist *Giardia lamblia*. Gene **280**:163–167.

Kaji, H., S. Tai, Y. Okimura, G. Iguchi, Y. Takahashi, H. Abe, and K. Chihara. 1998. Cloning and characterization of the 5′-flanking region of the human growth hormone secretagogue receptor gene. J. Biol. Chem. **273**:33885–33888.

Kapranov, P., S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. Fodor, and T. R. Gingeras. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science **296**:916–919.

Kent, W. J. 2002. BLAT—the BLAST-like alignment tool. Genome Res. **12**:656–664.
Kimura, M. 1962. On the probability of fixation of mutant genes in populations. Genetics **47**:713–719.
Kozak, M. 1987. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J. Mol. Biol. **196**:947–950.
———. 1994. Determinants of translational fidelity and efficiency in vertebrate mRNAs. Biochimie **76**:815–821.
———. 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene **299**:1–34.
Kutach, A. K., and J. T. Kadonaga. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. Mol. Cell. Biol. **20**:4754–4764.
Li, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.
Liston, D. R., and P. J. Johnson. 1999. Analysis of a ubiquitous promoter element in a primitive eukaryote: early evolution of the initiator element. Mol. Cell. Biol. **19**:2380–2388.
Lizama, L., L. Holuigue, and X. Jordana. 1994. Transcription initiation sites for the potato mitochondrial gene coding for subunit 9 of ATP synthase (atp9). FEBS Lett. **349**:243–248.
Lo, K., and S. T. Smale. 1996. Generality of a functional initiator consensus sequence. Gene **182**:13–22.
Lynch, M. 2002. Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. USA **99**:6118–6123.
Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. Science **302**:1401–1404.
Lynch, M., M. O'Hely, B. Walsh, and A. Force. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics **159**:1789–1804.
Lynch, M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, Mass.
Maicas, E., M. Shago, and J. D. Friesen. 1990. Translation of the *Saccharomyces cerevisiae* tcm1 gene in the absence of a 5′-untranslated leader. Nucleic Acids Res. **18**:5823–5828.
Mauro, V. P., and G. M. Edelman. 2002. The ribosome filter hypothesis. Proc. Natl. Acad. Sci. USA **99**:12031–12036.
Meijer, H. A., and A. A. Thomas. 2002. Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. Biochem. J. **367**:1–11.
Moll, I., S. Grill, C. O. Gualerzi, and U. Bläsi. 2002. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. Mol. Microbiol. **43**:239–246.
Morris, D. R., and A. P. Geballe. 2000. Upstream open reading frames as regulators of mRNA translation. Mol. Cell. Biol. **20**:8635–8642.
Nagawa, F., and G. R. Fink. 1985. The relationship between the ''TATA'' sequence and transcription initiation sites at the HIS4 gene of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA **82**:8557–8561.
Nakaar, V., D. Bermudes, K. R. Peck, and K. A. Joiner. 1998. Upstream elements required for expression of nucleoside triphosphate hydrolase genes of *Toxoplasma gondii*. Mol. Biochem. Parasitol. **92**:229–239.
Nicolaides, N. C., K. W. Kinzler, and B. Vogelstein. 1995. Analysis of the 5′ region of PMS2 reveals heterogeneous transcripts and a novel overlapping gene. Genomics **29**:329–334.
Niimura, Y., M. Terabe, T. Gojobori, and K. Miura. 2003. Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. Nucleic Acids Res. **31**:5195–5201.
Ohler, U., G. C. Liao, H. Niemann, and G. M. Rubin. 2002. Computational analysis of core promoters in the *Drosophila* genome. Genome Biol. **3**:research0087.1–0087.12.
Ota, T. et al. 2004. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat. Genet. **36**:40–45.
Peri, S., and A. Pandey. 2001. A reassessment of the translation initiation codon in vertebrates. Trends Genet. **17**: 685–687.
Pesole, G., G. Grillo, A. Larizza, and S. Liuni. 2000. The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis. Brief. Bioinform. **1**:236–249.
Pesole, G., S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi, and C. Saccone. 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., F. Mignone, C. Gissi, G. Grillo, F. Licciulli, and S. Liuni. 2001. Structural and functional features of eukaryotic mRNA untranslated regions. Gene **276**:73–81.
Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw. 2000. Evidence for DNA loss as a determinant of genome size. Science **287**:1060–1062.
Ptashne, M., and A. Gann. 2002. Genes and signals. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
Radford, A., and J. H. Parish. 1997. The genome and genes of *Neurospora crassa*. Fungal Genet. Biol. **21**:258–266.
Rogozin, I. B., A. V. Kochetov, F. A. Kondrashov, E. V. Koonin, and L. Milanesi. 2001. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics **17**:890–900.
Ruiz-Echevarria, M. J., and S. W. Peltz. 2000. The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell **101**:741–751.
Saito, R., and M. Tomita. 1999. On negative selection against ATG triplets near start codons in eukaryotic and prokaryotic genomes. J. Mol. Evol. **48**:213–217.
Shine, J., and L. Dalgarno. 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA **17**:1342–1346.
Singh, U., J. B. Rogers, B. J. Mann, and W. A. Petri Jr. 1997. Transcription initiation is controlled by three core promoter elements in the hgl5 gene of the protozoan parasite *Entamoeba histolytica*. Proc. Natl. Acad. Sci. USA **94**:8812–8817.
Slupska, M. M., A. G. King, S. Fitz-Gibbon, J. Besemer, M. Borodovsky, and J. H. Miller. 2001. Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. J. Mol. Biol. **309**:347–360.
Smale, S. T., and J. T. Kadonaga. 2003. The RNA polymerase II core promoter. Annu. Rev. Biochem. **72**:449–479.
Soppa, J. 1999*a*. Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. Mol. Microbiol. **31**:1589–1592.
———. 1999*b*. Transcription initiation in Archaea: facts, factors and future aspects. Mol. Microbiol. **31**:1295–1305.
Stapleton, M. et al. 2002. A *Drosophila* full-length cDNA resource. Genome Biol. **3**:research0080.1–0080.8.
Sterner, D. A., T. Carlo, and S. M. Berget. 1996. Architectural limits on split genes. Proc. Natl. Acad. Sci. USA **93**:15081–15085.
Strausberg, R. L. et al. 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.
Struhl, K. 1989. Molecular mechanisms of transcriptional regulation in yeast. Annu. Rev. Biochem. **58**:1051–1077.
Suzuki, Y. et al. 2000. Statistical analysis of the 5′ untranslated region of human mRNA using ''Oligo-Capped'' cDNA libraries. Genomics **64**:286–297.

Suzuki, Y. et al. 2001*a*. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. **2**:388–393.

———. 2001*b*. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. **11**:677–684.

Thomm, M. 1996. Archaeal transcription factors and their role in transcription initiation. FEMS Microbiol. Rev. **18**:159–171.

Tolstrup, N., C. W. Sensen, R. A. Garrett, and I. G. Clausen. 2000. Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus sol-fataricus*. Extremophiles **4**:175–179.

Tranque, P., M. C. Hu, G. M. Edelman, and V. P. Mauro. 1998. rRNA complementarity within mRNAs: a possible basis for mRNA-ribosome interactions and translational control. Proc. Natl. Acad. Sci. USA **95**:12238–12243.

van den Heuvel, J. J., R. J. Bergkamp, R. J. Planta, and H. A. Raue. 1989. Effect of deletions in the 5′-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. Gene **79**:83–95.

Vilela, C., and J. E. McCarthy. 2003. Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. Mol. Microbiol. **49**:859–867.

Watanabe, J., M. Sasaki, Y. Suzuki, and S. Sugano. 2002. Analysis of transcriptomes of human malaria parasite *Plasmodium falciparum* using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene **291**:105–113.

Weiner, J. III, R. Herrmann, and G. F. Browning. 2000. Transcription in *Mycoplasma pneumoniae*. Nucleic Acids Res. **28**:4488–4496.

Yamauchi, K., M. Mukai, T. Ochiai, and I. Usuki. 1992. Molecular cloning of the cDNA for the major hemoglobin component from *Paramecium caudatum*. Biochem. Biophys. Res. Commun. **182**:195–200.

Yee, J., M. R. Mowatt, P. P. Dennis, and T. E. Nash. 2000. Transcriptional analysis of the glutamate dehydrogenase gene in the primitive eukaryote, *Giardia lamblia*. Identification of a primordial gene promoter. J. Biol. Chem. **275**:11432–11439.

Yu, Y. S., Y. Suzuki, K. Yoshitomo, M. Muramatsu, N. Yamaguchi, and S. Sugano. 1996. The promoter structure of TGF-β type II receptor revealed by ''oligo-capping'' method and deletion analysis. Biochem. Biophys. Res. Comm. **225**: 302–306.

Edward Holmes, Associate Editor

Accepted January 25, 2005