Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans

Benjamin P. Lewis*†‡, Richard E. Green*‡§, and Steven E. Brenner*†§¶

Departments of *Plant and Microbial Biology, §Molecular and Cell Biology, and †Biophysics Graduate Group, University of California, Berkeley, CA 94720

Communicated by Sydney Kustu, University of California, Berkeley, CA, November 6, 2002 (received for review August 1, 2002)

To better understand the role of alternative splicing, we conducted a large-scale analysis of reliable alternative isoforms of known human genes. Each isoform was classified according to its splice pattern and supporting evidence. We found that one-third of the alternative transcripts examined contain premature termination codons, and most persist even after rigorous filtering by multiple methods. These transcripts are apparent targets of nonsense-mediated mRNA decay (NMD), a surveillance mechanism that selectively degrades nonsense mRNAs. Several of these transcripts are from genes for which alternative splicing is known to regulate protein expression by generating alternate isoforms that are differentially subjected to NMD. We propose that regulated unproductive splicing and translation (RUST), through the coupling of alternative splicing and NMD, may be a pervasive, underappreciated means of regulating protein expression.

regulation | EST | RefSeq | human genome | regulated unproductive splicing and translation

Alternative splicing plays a major role in modulating gene function by expanding the diversity of expressed mRNA transcripts (1–4). An extreme example in *Drosophila* is the alternative splicing of the *Dscam* gene, which may generate >38,000 distinct mRNA isoforms (5), more than twice the number of predicted genes in the entire genome (6), to mediate the formation of neuronal cell–cell contacts. Moreover, alternative splicing of genes with just a few isoforms may nonetheless yield profound regulatory effects. This finding is exemplified by human Bcl-x, whose products include two isoforms with markedly different activities. Bcl-x(L) is an antiapoptotic factor, whereas Bcl-x(S) can induce apoptosis (7). Seeking to understand alternative splicing and the protein repertoire encoded by the human genome, many groups have undertaken studies to infer and enumerate alternative mRNA isoforms (2, 8–12).

Standard analyses, however, may not provide a full appreciation of how alternative splicing modulates gene function. Because of the limitations of the ESTs from which alternative splicing information is commonly derived (13), researchers sometimes cautiously restrict their analyses to exon skipping and mutually exclusive exon usage (2, 12). Similarly, researchers commonly dismiss alternative transcripts that code for apparent early translational termination, because those mRNAs are deemed incapable of generating a functional product. A more complete understanding of alternative splicing requires an unbiased consideration of all reliable alternative mRNA isoforms.

Alternative Isoform Inference

We examined the alternative mRNAs suggested by EST alignments, using a protocol designed to comprehensively identify maximally reliable sequences that are alternatively spliced (Fig. 1*a*). To exclude errors from genome sequencing and assembly, and to simplify the task of determining reading frame for each transcript, our analysis used 16,163 well-characterized human mRNAs from RefSeq and LocusLink (14). This set excludes the computational genome annotation RefSeq category, as well as 617 mRNAs containing premature termination codons (see *Analysis of Premature Termination Codons in RefSeq mRNAs*). First, we mapped the mRNAs to the human genome, requiring that an mRNA align to the genomic sequence over the full length of the coding sequence, without gaps in the exons. We further required 98% identity between the coding sequences, favoring the RefSeq sequence in cases of nucleotide mismatch. When multiple RefSeq mRNAs aligned to the same region of the genomic sequence, we used only the mRNA containing the largest number of exons. To detect alternative isoforms, we aligned 4.6 million EST sequences from dbEST (15) to the genomic sequence and used TAP (8) to infer alternative mRNA splice forms from these alignments (Fig. 1*c*). Because we used known genes, the reading frame of each canonical mRNA isoform (i.e., the RefSeq mRNA) was known. To ensure that the reading frame could be determined for all EST-suggested alternative isoforms, we excluded any EST whose 5′ end aligned to regions of the genomic sequence that did not correspond to coding exons of the RefSeq mRNA. We also excluded cases of intron retention, because these are indistinguishable from incompletely processed transcripts, a common dbEST contaminant. After applying these filters for reliability, this protocol identified 3,127 RefSeq mRNAs, whose genes undergo alternative splicing to generate 8,820 distinct mRNAs. Within this set, we have higher confidence in splicing events with coverage by multiple ESTs, because these are less likely to result from experimental artifacts in dbEST. The overall process involved the following steps.

Mapping RefSeq mRNAs to the Human Genome. Annotations from the August 2002 version of LocusLink (14) were used to associate 16,163 human mRNAs from the August 2002 version of RefSeq (14) with contig sequences from the National Center for Biotechnology Information (NCBI) human genome build 30 (16). The coding regions of the RefSeq mRNAs were aligned against the corresponding contig sequences with the mRNA alignment tool SPIDEY (ref. 17; Fig. 1*a*). Because the untranslated regions of the RefSeq mRNAs often aligned poorly to the genomic sequence, we constructed alignments for only the coding portions of the RefSeq mRNAs. Cases where alternative splicing affects the untranslated regions of RefSeq-coding genes (e.g., in SC35; ref. 18) were thus excluded (Fig. 1*a*).

Aligning EST Sequences to Genomic Sequences. Repetitive elements in the genomic template sequences were masked with REPEATMASKER [A. F. A. Smit and P. Green (1996–2001) http://ftp.genome.washington.edu/RM/RepeatMasker.html]. Using

Abbreviation: NMD, nonsense-mediated mRNA decay; RUST, regulated unproductive splicing and translation.

‡B.P.L. and R.E.G. contributed equally to this work.

¶To whom correspondence should be addressed at: Department of Plant and Microbial Biology, 111 Koshland Hall, #3102, University of California, Berkeley, CA 94720-3102. E-mail: brenner@compbio.berkeley.edu.

GENETICS

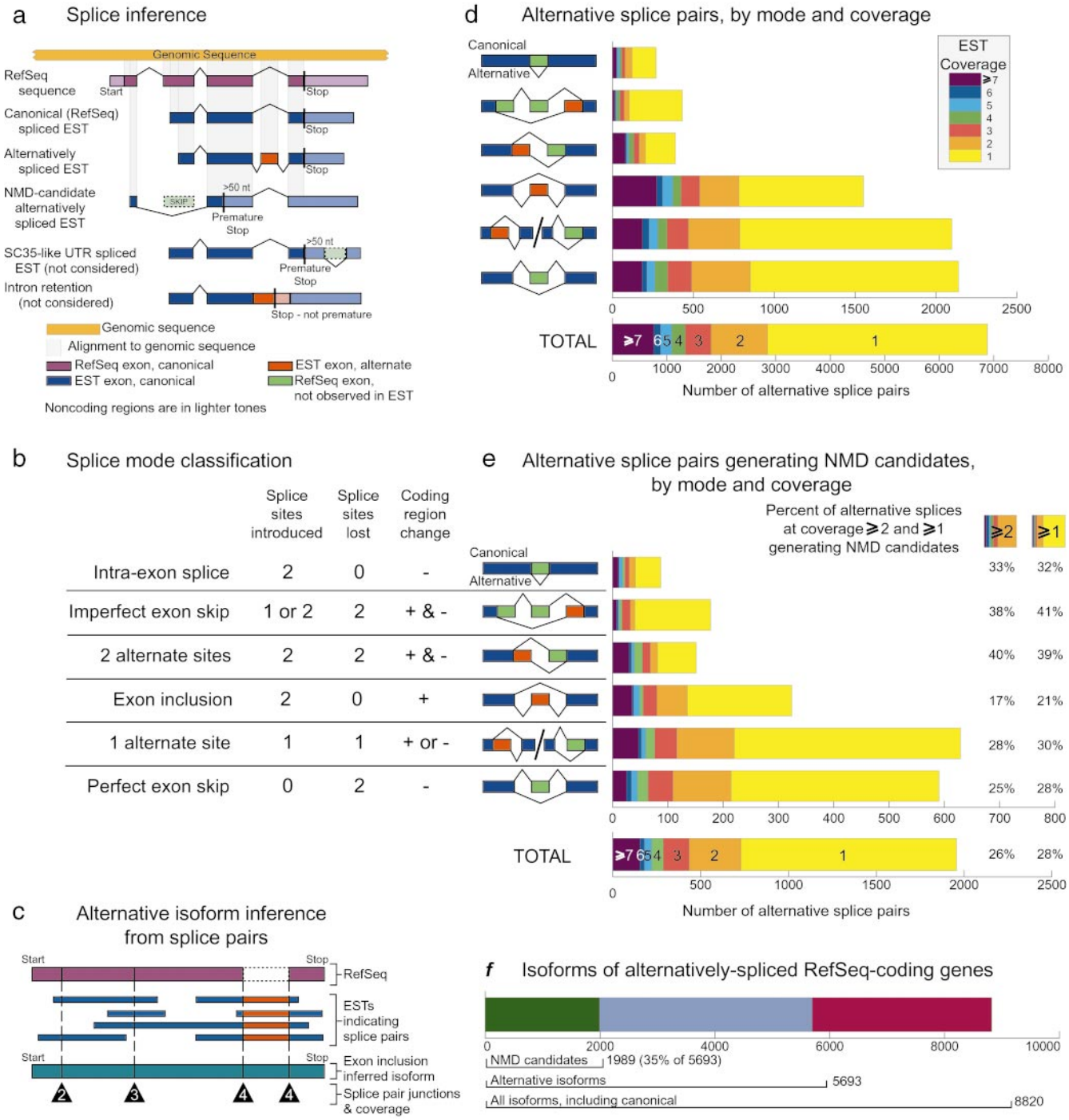

Fig. 1. Alternative splice detection and classification. (*a*) Splice inference. Coding regions of RefSeq mRNAs were aligned to the genomic sequence to determine canonical splicing patterns. EST alignments to the genomic sequence confirmed the canonical splices and indicated alternative splices. Canonical (RefSeq) splices are indicated above the exons, whereas alternative splices are indicated below the exons. When an alternative splice introduced a stop codon >50 nucleotides upstream of the final exon–exon splice junction of an inferred mRNA isoform, the stop codon was classified as a premature termination codon and the corresponding mRNA isoform was labeled a NMD candidate. In the NMD-candidate example shown, an exon skip caused a frameshift, resulting in the introduction of a premature termination codon. Restricting the analysis to coding regions assured high alignment quality, but this excluded alternative splicing in noncoding regions, such as that which occurs with splicing factor SC35. Intron retentions were also excluded because ESTs indicating intron retention are indistinguishable from incompletely processed transcripts, a common dbEST contaminant. (*b*) Splice mode classification. Alternative splices were categorized according to splice site usage and effects on the coding sequence. ''Splice sites introduced'' shows the number of splice donor/acceptor sites that were observed in the alternative splice but were not included in the canonical splice. ''Splice sites lost'' shows the number of splice donor/acceptor sites that were included in the canonical splice but were absent in the alternative splice. ''Coding region change'' indicates whether an alternative splice added (red) or subtracted (green) coding sequence to the alternative isoform relative to the canonical isoform. By our method, mutually exclusive exon usage appears as exon inclusion. Our analysis excluded intron retentions, which would be classified as zero splice sites introduced, two sites lost, and addition of coding sequence. (*c*) Alternative isoform inference from splice pairs. Splice pairs are splice donor/acceptor sites (▲) inferred from the alignments. Alternative splice pairs are those indicated by ESTs, but not by a RefSeq mRNA. The exon composition of an isoform was determined from EST-demonstrated splice pairs, which may be covered by multiple ESTs. Coverage of splice pairs is indicated in each ▲. Coverage for a complete isoform is not meaningful because of the variability in coverage of its splice pairs.

WU-BLASTN 2.0MP-WASHU [07-Jun-2002] [W. R. Gish (1996–2002) (Washington University, St. Louis) http://blast.wustl.edu], we searched the 4.6 million EST sequences from dbEST (15) version 280802 for matches to the coding exons of the RefSeq mRNA, and the intervening intron sequences in the human genome. The EST sequences with P value $<10^{-30}$ were aligned to the genomic sequences by using SIM4 1.4 (19). Only EST alignments with >92% identity were used.

Alternative Isoform Inference. We used TAP (8) to infer alternative mRNA splice forms from the EST alignments.

Alternative Isoform Analysis

Alternative isoforms were inferred, analyzed, and further filtered as follows.

Analysis of Canonical and Alternative Splice Pairs. Alternative splice pairs are defined as EST-inferred splice junction donor and acceptor sites that differ from those in the canonical RefSeq mRNAs (Fig. 1*a*). To avoid erroneous alternative splice pair predictions resulting from ambiguity in the alignments surrounding splice junctions, we rejected putative alternative splice pairs found <7 bp from a canonical splice pair. Each aligned EST may indicate multiple alternative and canonical splice pairs. Alternative splice pairs within the same mRNA isoform may have various levels of EST coverage (Fig. 1*c*). Whenever a splice in an alternative isoform was not covered by ESTs, it was assumed to be canonical.

Classification of Alternative Splice Pairs. Each EST-inferred alternative splice pair was classified according to EST coverage (Fig. 1*c*), its effect on the coding region of the underlying genomic sequence, and exon and splice site usage (Fig. 1*d*). By this method, mutually exclusive exon usage appeared as exon inclusion. Note that two alternative splice pairs are associated with a single exon inclusion event. Also, exon inclusion may be viewed as exon skipping from the perspective of the alternative isoform.

Classification of Alternative Splicing Modes. Alternative splices were categorized according to splice site usage and effects on the coding sequence (Fig. 1*b*), as described in the legend to Fig. 1.

Identification of Premature Termination Codons. Premature termination codons are stop codons that occur >50 nucleotides upstream of the final splice junction (20–27). When an inferred mRNA isoform was found to contain a premature termination codon, that isoform was labeled as a NMD candidate. The tendency for alternative splicing to introduce premature termination codons may be viewed at the level of alternative splice pairs (Fig. 1*e*) or alternative mRNA isoforms (Fig. 1*f*).

Analysis of Polyadenylation Signals. POLYADQ (28) was used to search the alternative mRNAs for polyadenylation sites. On average, a predicted polyadenylation signal occurred once every 2,646 nucleotides in the coding exons of the RefSeq mRNAs and the intervening introns. Regions spanning from a premature termination codon to the first splice junction >50 nucleotides downstream contained predicted polyadenylation signals once every 3,187 nucleotides.

Analysis of Premature Termination Codons in RefSeq mRNAs. To determine whether premature termination codons exist in experimentally identified mRNA transcripts, we examined the occurrence of premature termination codons in the set of reviewed RefSeq mRNAs from the August 2002 version of RefSeq (14). All RefSeq mRNAs that are identified as reviewed RefSeq records have been individually examined by NCBI staff. Thus, these sequences represent the most reliable segment of RefSeq. The position of the termination codon in each reviewed RefSeq mRNA was taken from the RefSeq annotation. The position of the final splice junction was determined by using SPIDEY (17) to align the mRNA to an NCBI human genome build 30 contig sequence that had been associated by using LocusLink (14). If the stop codon of the RefSeq mRNA was found >50 nucleotides upstream of the final splice junction, the stop codon was then identified as a premature termination codon.

Selection of Nonnormalized, Nondiseased-Cell EST Libraries. We used UNILIB library annotations to construct a restricted set of EST libraries [National Center for Biotechnology Information (NCBI), www.ncbi.nlm.nih.gov/UniLib]. The keyword "protocol," type "nonnormalized," was used to search the classification hierarchy for nonnormalized libraries. The keyword "histology," type "normal," was used to identify libraries constructed by sequencing nondiseased tissue. We took ESTs in the intersection of these two subsets as being from nonnormalized, nondiseased-cell libraries.

Results and Discussion

Among the RefSeq mRNAs in our analysis, 3,127 were found to have 6,884 alternative splice pairs and 5,693 alternative mRNA isoforms. We categorized the alternative mRNAs according to exon and splice site usage (Fig. 1 *b* and *d*). Each canonical and alternative isoform is described in Table 1, which is published as supporting information on the PNAS web site, www.pnas.org.

We found that many alternative mRNA isoforms have premature termination codons that render them apparent targets for NMD. Recent work has elucidated the following model for mammalian NMD (24, 25, 29, 30). During mRNA processing, exon–exon splice junctions are marked with exon junction complexes that serve the dual purpose of facilitating export to the cytoplasm and remembering gene structure (20). As translation occurs, the ribosome displaces all exon junction complexes in its path. If a complex remains after a pioneering round of translation (21), a series of reactions ensue, leading to transcript degradation. Thus, transcripts that contain premature termination codons, that is, termination codons >50 nucleotides 5′ of the final exon (20–27), are candidates for NMD. As Wagner and Lykke-Anderson (27) report, "NMD is a critical process in normal cellular development." NMD has been shown to occur in all eukaryotes tested and, although it has variable efficiency (31), eukaryotic mRNAs containing premature termination codons are almost always degraded rapidly (26). Further supporting this idea, we observed that only 4.3% of mRNAs from the reviewed category of RefSeq are NMD candidates, with stop codons

(*d*) Alternative splice pairs by mode and coverage. The total number of alternative splice pairs associated with each splicing mode is shown at various levels of EST coverage. The distance from the *y* axis to the right edge of each box corresponds to the total number of splice pairs with coverage greater than or equal to the number indicated. Note that each exon inclusion event involves two splice pairs. (*e*) Alternative splice pairs generating NMD candidates by mode and coverage. The panel shows the subset of alternative splice pairs that produce premature termination codons. These splice pairs are involved in generating NMD-candidate mRNA isoforms. The numbers of splice pairs are displayed as in *d*. Also shown are the NMD-candidate splice pairs at coverage ≥1 and ≥2 as a percentage of all alternative splice pairs for each splicing mode. (*f*) Isoforms of alternatively spliced RefSeq-coding genes. Shown are the total numbers of isoforms of the RefSeq-coding genes for which alternative isoforms were found. These are subdivided into the following categories: all isoforms including canonical, alternative isoforms (i.e., all isoforms excluding canonical), and NMD candidates.

located >50 nucleotides upstream of the final exon. In contrast, we discovered that in 34% of these sequences, the start codon occurred downstream of the first exon.

Thirty-five percent of the EST-suggested alternative isoforms in our study contain premature termination codons (Fig. 1*f*). For a subset comprising 74% of these NMD-candidate mRNA isoforms, EST alignments cover a premature termination codon and a splice junction >50 nucleotides downstream. In these cases, there is no possibility that additional undetected splicing events might remove 3′ exons, thereby preventing termination from being premature. Furthermore, within this subset of NMD candidates, 83% have premature termination codons occur in all three reading frames, thus precluding the possibility that an upstream splicing event changed the reading frame from that of the canonical form to prevent the incorporation of a premature termination codon. Finally, we found that the distribution of predicted polyadenylation signals in NMD-candidate splices is biased against regions just downstream of premature termination codons, undermining the likelihood that alternative polyadenylation stabilizes many of the NMD-candidate transcripts.

Our analysis identified 1,106 genes that undergo alternative splicing to generate 1,989 alternative mRNA isoforms that are apparent targets for NMD. Such widespread coupling of alternative splicing and NMD may indicate that the cell possesses a large number of irrelevant mRNA isoforms that must be eliminated. A more compelling alternative, which has been investigated in analyses of *smg* mutations in *Caenorhabditis elegans*, is that the deliberate coupling of alternative splicing and NMD plays a functional role in regulating protein expression levels (3, 32, 33). Supporting this view, our analysis turned up several genes known to be regulated by generating isoforms targeted for NMD, including glutaminase (34), and fibroblast growth factor receptor 2 (35). We also found alternatively spliced NMD candidates for six other splicing factors. Besides these, the splicing factor SC35 has been shown to autoregulate its expression through regulated unproductive splicing and translation (RUST) by generating NMD-targeted isoforms (18), although it is excluded from our analysis because its alternative splicing does not affect its coding sequence (Fig. 1*a*).

Additionally, we found that the human genes for 5 translation factors and 11 ribosomal proteins generate NMD-candidate isoforms. Intriguingly, *C. elegans* homologs of three of these ribosomal genes, RP3, RP10a, and RP12, generate splice forms that are cleared by NMD (33), suggesting that this mode of regulating ribosomal protein expression is evolutionarily conserved. Experimental work will be necessary to further characterize the role of coupled alternative splicing and NMD in the expression of the genes we have identified.

Because EST libraries are naturally biased against less stable transcripts, mRNAs subjected to NMD should have lower coverage than stable alternative splice forms of the same gene. Therefore, it is striking that many NMD candidates are indicated by multiple ESTs (Fig. 1*e* and Table 2, which is published as supporting information on the PNAS web site). Within nonnormalized, nondiseased-cell libraries, the fraction of splices that generate NMD candidates with coverage one is slightly reduced, and this fraction drops precipitously at higher coverage (Table 3, which is published as supporting information on the PNAS web site), rendering the quantitation of these data uninterpretable. In light of transcript biases in dbEST and the fact that splicing in the RefSeq 3′ UTR (e.g., in SC35) is excluded from our analysis, we suspect that alternative splicing of NMD-targeted transcripts might be more prevalent than our data suggest.

The coupling of alternative splicing and NMD is easily incorporated into existing models of gene regulation. It allows the use of the intrinsic alternative splicing machinery to regulate protein expression in a developmental stage- and cell-specific manner. Moreover, the transcription of genes that will yield unproductive mRNAs is no more wasteful than the transcription of introns, and particularly for genes that require a long time to be transcribed (e.g., dystrophin, which takes 16 h; ref. 36), post-transcriptional regulation of this sort could provide temporal control unattainable by transcription factors. In light of our findings, we reason that the contribution of alternative splicing to proteome diversity may be balanced by an as yet unappreciated regulatory role in gene expression.

We are grateful for stimulating discussions with Jasper Rine, Don Rio, and Sydney Kustu. This work was supported by National Institutes of Health Grants K22 HG00056 and T32 HG0047, the Searle Scholars' Program (01-L-116), and an IBM Shared University Research grant. B.P.L. is supported by a Krell Institute Fellowship.

1. Modrek, B. & Lee, C. (2002) *Nat. Genet.* **30,** 13–19.
2. Brett, D., Hanke, J., Lehmann, G., Haase, S., Delbruck, S., Krueger, S., Reich, J. & Bork, P. (2000) *FEBS Lett.* **474,** 83–86.
3. Graveley, B. R. (2001) *Trends Genet.* **17,** 100–107.
4. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. (2002) *Nucleic Acids Res.* **30,** 1083–1090.
5. Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E. & Zipursky, S. L. (2000) *Cell* **101,** 671–684.
6. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000) *Science* **287,** 2185–2195.
7. Boise, L. H., Gonzalez-Garcia, M., Postema, C. E., Ding, L., Lindsten, T., Turka, L. A., Mao, X., Nunez, G. & Thompson, C. B. (1993) *Cell* **74,** 597–608.
8. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. (2001) *Genome Res.* **11,** 889–900.
9. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. (1999) *Genome Res.* **9,** 1288–1293.
10. Huang, Y. H., Chen, Y. T., Lai, J. J., Yang, S. T. & Yang, U. C. (2002) *Nucleic Acids Res.* **30,** 186–190.
11. Burke, J., Wang, H., Hide, W. & Davison, D. B. (1998) *Genome Res.* **8,** 276–290.
12. Hide, W. A., Babenko, V. N., van Heusden, P. A., Seoighe, C. & Kelso, J. F. (2001) *Genome Res.* **11,** 1848–1853.
13. Thanaraj, T. A. (1999) *Nucleic Acids Res.* **27,** 2627–2637.
14. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
15. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. (1993) *Nat. Genet.* **4,** 332–333.
16. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409,** 860–921.
17. Wheelan, S. J., Church, D. M. & Ostell, J. M. (2001) *Genome Res.* **11,** 1952–1957.
18. Sureau, A., Gattoni, R., Dooghe, Y., Stevenin, J. & Soret, J. (2001) *EMBO J.* **20,** 1785–1796.
19. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. (1998) *Genome Res.* **8,** 967–974.
20. Le Hir, H., Izaurralde, E., Maquat, L. E. & Moore, M. J. (2000) *EMBO J.* **19,** 6860–6869.
21. Ishigaki, Y., Li, X. J., Serin, G. & Maquat, L. E. (2001) *Cell* **106,** 607–617.
22. Hilleren, P. & Parker, R. (1999) *Annu. Rev. Genet.* **33,** 229–260.
23. Lykke-Andersen, J., Shu, M. D. & Steitz, J. A. (2000) *Cell* **103,** 1121–1131.
24. Lykke-Andersen, J., Shu, M. D. & Steitz, J. A. (2001) *Science* **293,** 1836–1839.
25. Kim, V. N., Kataoka, N. & Dreyfuss, G. (2001) *Science* **293,** 1832–1836.
26. Nagy, E. & Maquat, L. E. (1998) *Trends Biochem. Sci.* **23,** 198–199.
27. Wagner, E. & Lykke-Andersen, J. (2002) *J. Cell Sci.* **115,** 3033–3038.
28. Tabaska, J. E. & Zhang, M. Q. (1999) *Gene* **231,** 77–86.
29. Mitchell, P. & Tollervey, D. (2001) *Curr. Opin. Cell Biol.* **13,** 320–325.
30. Cartegni, L., Chew, S. L. & Krainer, A. R. (2002) *Nat. Rev. Genet.* **3,** 285–298.
31. Gudikote, J. P. & Wilkinson, M. F. (2002) *EMBO J.* **21,** 125–134.
32. Morrison, M., Harris, K. S. & Roth, M. B. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 9782–9785.
33. Mitrovich, Q. M. & Anderson, P. (2000) *Genes Dev.* **14,** 2173–2184.
34. Labow, B. I., Souba, W. W. & Abcouwer, S. F. (2001) *J. Nutr.* **131,** 2467S–2474S; discussion 2486S–24875S.
35. Jones, R. B., Wang, F., Luo, Y., Yu, C., Jin, C., Suzuki, T., Kan, M. & McKeehan, W. L. (2001) *J. Biol. Chem.* **276,** 4158–4167.
36. Tennyson, C. N., Klamut, H. J. & Worton, R. G. (1995) *Nat. Genet.* **9,** 184–190.