Minimal Introns Are Not "Junk"

Jun Yu,[1,2,3,5,6] Zhiyong Yang,[4,5] Miho Kibukawa,[1] Marcia Paddock,[1] Douglas A. Passey,[1] and Gane Ka-Shu Wong[1,2,3]

[1]*University of Washington Genome Center, Department of Medicine, Seattle, Washington 98195, USA;* [2]*Hangzhou Genomics Institute, Institute of Bioinformatics of Zhejiang University, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China;* [3]*Beijing Genomics Institute, Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China;* [4]*Walter and Eliza Hall Institute of Medical Research, Royal Melbourne Hospital, Victoria 3050, Australia.*

Intron-size distributions for most multicellular (and some unicellular) eukaryotes have a sharp peak at their "minimal intron" size. Across the human population, these minimal introns exhibit an abundance of insertion-deletion polymorphisms, the effect of which is to maintain their optimal size. We argue that minimal introns affect function by enhancing the rate at which mRNA is exported from the cell nucleus.

Decades of research on the mechanisms of pre-mRNA splicing have revealed a remarkably intricate process. These complexities include exon versus intron recognition (Berget 1995), cotranscriptional splicing (Goldstrohm et al. 2001), alternative splicing (Dredge et al. 2001; Grabowski and Black 2001), exonic splicing enhancers (Blencowe 2000; Nissim-Rafinia and Kerem 2002), intronic splicing enhancers (McCullough and Berget 2000), and tight coupling between splicing and efficient mRNA export from the nucleus (Luo and Reed 1999; Zhou, et al. 2000). Given such complexity, it is not hard to imagine that different introns are processed differently, not only between species but also within species. That being the case, can we segregate intron sequences according to differences in how they are processed? If so, might these differences be reflected in the nature of the sequence polymorphisms that are found in the population? We will argue that the answer to both questions is yes.

On the surface, this is an implausible idea, because intron sequences are poorly conserved. Known splicing motifs (e.g., GT-AG, branch points) are only a few bases in size, whereas intron lengths can be hundreds of kilobases. In the transition between cold-blooded to warm-blooded vertebrates, many introns experienced a twofold increase in GC content (Bernardi 2000). The fact that intron sequence contents are so pliable is the reason why introns are often considered "junk." On the other hand, the enzymatic degradation of the excised introns must be a significant biochemical burden for the cell, especially if most of the human genome is transcribed (Wong et al 2000, 2001). Why would the cell go to so much trouble? Why not just get rid of the introns? People who do experiments on transgenic mice have an answer, for they have long known that some introns are essential for a high level of expression (Choi et al. 1991; Palmiter et al. 1991). If introns can influence expression levels, they are certainly not junk. Where might one go to find such introns?

One of the most conspicuous features of eukaryotic genomes is that a significant fraction of the introns are often clustered around a species-specific peak at the low end of the size distribution. We call them "minimal" introns because there are none smaller. Our objective is to show that the evolutionary persistence of such an optimal intron size is owing to functional constraints. However, there are many practical difficulties. Minimal intron sequence contents are degenerate (Lim and Burge 2001). Not every intron is size constrained, and any benefits to having an optimal intron size are likely to be marginal. We reasoned that our chances of success would be best in a genome in which large introns are prevalent, like human, because evolution would have already selected those introns that need to remain small from those that do not. Because of the recent expansion in the human population (Harpending and Rogers 2000), even a slight benefit, as might be expected from a small change in the intron size, would have a high probability of being fixed in the population. One might thus expect to see an abundance of minor alleles that embody the process of intron size optimization.

Specifically, we present resequencing data on a collection of 93 minimal introns sampled across a diverse human population. The data reveal an abundance of insertion-deletion (indel) polymorphisms that are clearly trying to maintain the optimal intron size. From an analysis of the yeast expression data, we will show that minimal introns can enhance mRNA synthesis rates. In essence, we present an example of selection based on conservation of intron size, as opposed to conservation of sequence content. In fact, studies of recombination rates in *Drosophila melanogaster* have indicated that there are selective pressures on intron size (Carvalho and Clark 1999). Perhaps the perception that introns are junk is an artifact of an overly narrow focus on conservation of sequence content as the only signature of selection.

RESULTS

Minimal Introns Are Found in Most Multicellular Eukaryotes

The distributions for intron size in *Homo sapiens*, *Arabidopis thaliana*, *D. melanogaster*, and *Caenorhabditis elegans* are displayed in Figure 1. All of these data are based on cDNA-to-genomic alignments, not gene-prediction programs. The mean number of introns per gene is 12.1, 6.2, 4.7, and 7.7, respectively. A significant fraction of the introns is always clustered about a species-specific minimum size, reflected by the sharp "spike" in the distribution centered around a mean ($\pm$SD) intron size of 92 ± 14, 89 ± 12, 61 ± 10, and 48 ± 9 bp, respectively. The idea that introns might have a minimum

[5]**These authors contributed equally to this work.**
[6]**Corresponding author.**
E-MAIL junyu@u.washington.edu; FAX (206) 685-7344.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.224602.

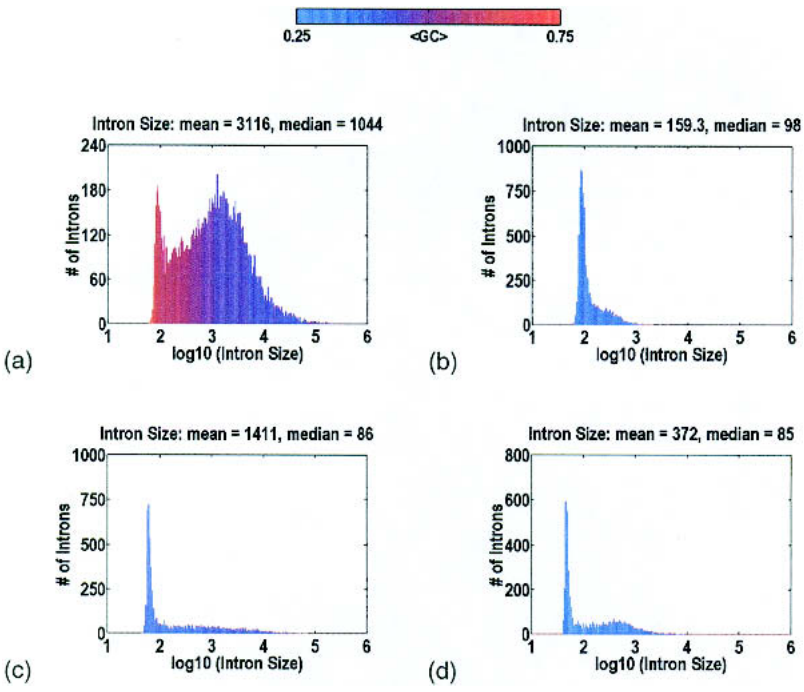

Figure 1 Intron size for *Homo sapiens* (*a*), *Arabidopis thaliana* (*b*), *Drosophila melanogaster* (*c*), and *Caenorhabditis elegans* (*d*). There is always a species-specific minimum intron size, at which a significant fraction of the introns tend to cluster. This "spike" in the distribution is centered around the mean (±SD) intron sizes of 92±14, 89±12, 61±10, and 48±9 bp, respectively. For larger introns, the size distribution is highly species-specific. In the extreme case, *H. sapiens*, there is a broad "hump" attributable to transposon insertions inside the introns. Color indicates GC content: Red is GC-rich; blue, AT-rich.

size, independent of sequence content, is not new (Wieringa et al. 1984). Presumably, it is a reflection of the physical constraints imposed by the cellular machinery, and the dimensions of this machinery are species specific. Minimal introns are also observed in *Mus musculus*, *Gallus gallus*, *Xenopus laevis*, *Fugu rubripes*, and *Oryza sativa*, albeit with annotation data parsed from GenBank. Yeast also has minimal introns, at 92±20 and 49±11 bp, for *Saccharomyces cerevisiae* (Ares et al. 1999; Spingola et al. 1999) and *Schizosaccharomyces pombe* (Wood et al. 2002), respectively.

In contrast, there is an enormous variability from species to species in the distributions for larger introns. In the most extreme case, *H. sapiens*, there is a broad "hump" in the intron size distribution, extending out to hundreds of kilobases. Sequence content analysis with `RepeatMasker` (http://ftp.genome.washington.edu/RM/RepeatMasker.html) reveals a gradual transition, from introns with no detectible transposons, at <1 Kb, to introns with one or more transposons, at >1Kb. It is thus reasonable to make a distinction between "minor" humps owing to introns <1Kb, and "major" humps owing to introns >1Kb. Given that the absence of a major hump can be owing to acquisition biases against large sequence contigs, the only statement that we are comfortable with is that there is a major hump in *H. sapiens*, *M. musculus*, *G. gallus*, and *X. laevis*.

Minimal Introns Are Not Randomly Distributed Among Genes

Considering that so many *H. sapiens* introns have been expanded by transposon insertions, one has to wonder why the minimal intron peak persists. Are there benefits to the organism for maintaining some introns at an optimal size? Given the predominance of the neutral theory of evolution (Kimura 1983), we must first eliminate any neutral or semineutral explanations. Perhaps some introns were never bombarded by transposons. Even if an intron was bombarded, it could have deleted back to the minimum because of the mutational bias for deletions over insertions, which is well known from comparisons of processed pseudogenes with their functional paralogs (Ophir and Graur 1997; Petrov 2001). Combined with the selection against introns too small for the splicing machinery, it would appear that persistence of minimal introns can be explained, without using any special functional constraints. However, something else must also be going on, as there is a glaring inconsistency in the data.

A distinguishing characteristic of the above-mentioned processes is that they do not favor any particular intron or gene. Therefore, if minimal introns are a fraction f_m of the total, and if there are R introns per gene, the probability that a gene has a minimal intron would be $1-(1-f_m)^R$. More precisely, we define a minimal intron as anything that lies within three standard deviations of the optimum, which in *H. sapiens* amounts to 13.6% of the introns. Because f_m varies with GC content, we compute f_m in four groups (similar results are obtained with eight groups) based on GC content in 10-Kb windows at each end of the gene. We then integrate over all observed Rs. The computation is performed on 882 genes from a previous analysis (Wong et al. 2000), containing every gene with a cDNA sequence that could be aligned in its entirety to finished genomic sequence. The neutral expectation is that minimal introns will be found in 56.2% of the genes, but the reality is 44.4%. The difference is statistically significant, P(binomial)$=5\times10^{-13}$, which means that minimal introns tend to cluster in certain genes.

In our four primary data sets—*H. sapiens*, *A. thaliana*, *D. melanogaster*, and *C. elegans*—the magnitude of the difference between the observed and expected number of genes with at least one minimal intron is −11.7%, −2.0%, −6.5%, and −4.7%, respectively. Evidently, the nonrandom distribution of minimal introns among genes is most readily observable in those species with a greater number of extremely large introns. Assuming that these introns are the result of transposon bombardment, this implies that transposon activity acts as a probe of how sensitive each gene is to the presence of minimal introns. Without significant transposon activity over the evolutionary history of a species, minimal introns remain randomly distributed. Thus, the genome we should resequence is *H. sapiens*, because evolution has already separated those introns that need to remain small from those that do not.

Minimal Introns Are Full of Indel Polymorphisms

According to Kimura (1993), "polymorphism is just a tran-

sient phase of molecular evolution." We reasoned that, if evolution is really trying to maintain some species-specific minimal intron size, it might be possible to catch this process in action from an analysis of minimal intron polymorphisms across a diverse population. Without further justification, we will let the data speak for themselves.

Our resequencing efforts were focused on introns with sizes close to the human optimum of 92 ± 14 bp. We resequenced 93 of these introns in a population of diverse ethnicity (Collins et al. 1998), over an average of 45.7 individuals (91.4 chromosomes). These introns were small, so there were no transposons in them. Their mean (±SD) size was 94 ± 14 bp. To minimize the potential biases arising from differences in mutation and recombination rates, most of which are correlated with local GC content, we selected introns that span the full range of GC content, as depicted in Figure 2. We identified 42 polymorphic sites in all, 30 single-base substitutions and 12 indels, with the indels in nine different introns. To compare our results with the published data, we adjusted for variations in sample depths and sequenced lengths. If K polymorphic sites are found in a region of length L after sequencing n chromosomes, the commonly used population genetics parameter (Cargill et al. 1999; Halushka et al. 1999) is the normalized number of variant sites,

$$\theta = K \Big/ \sum_{i=1}^{n-1} \frac{L}{i}.$$

We decompose θ into separate components, θ(subst) = 6.75×10^{-4} for substitutions, and θ(indel) = 2.70×10^{-4} for indels. Strikingly, this substitution rate is not significantly different from the substitution rate of 7.51×10^{-4} reported for the human genome; Sachidanandam et al. 2001). It means that intron sequence content is not what was being conserved.

However, 28.6% of our minimal introns polymorphisms were indels, which is significantly more than usual. To make this point, we need a background rate for human indels. Such a rate is not readily available, as most large-scale polymorphism discovery projects are focused on the easier-to-genotype substitution polymorphisms. Many of the putative indels are in poly-N tracts, where N is any nucleotide, and these are usually caused by sequencing errors. For example, 13% of chromosome 22 polymorphisms were indels, but this ratio was only 4% when poly-N tracts were ignored (Mullikin et al. 2000). We note that in our minimal intron indels, the longest poly-N tract was a run of only nine Gs, as shown in Table 1. For a background rate, we used the data from the Environmental Genome Project (http://www.genome.washington.edu/projects/egpsnps). These data focus on introns of every size but are restricted to the first few hundred bases flanking the exons, much like our minimal intron data, which also never stray far from the exons. Averaged over 90 genes, 7.9% of 392 intron polymorphisms were indels. Taking 7.9% as the null hypothesis, the observation of 28.6% indel polymorphisms in minimal introns is statistically significant, with P(binomial) = 6×10^{-5}.

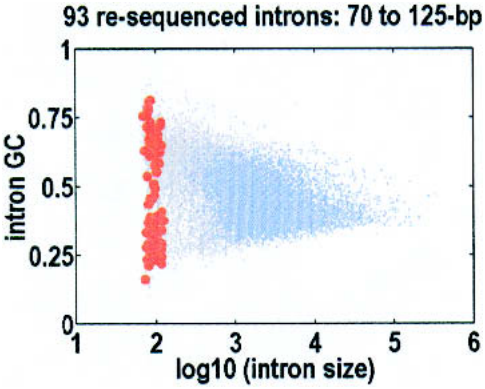

Figure 2 Resequenced introns, shown against the joint distribution for intron size and GC content. Introns with a detectible transposon are colored blue; introns with no detectible transposons, gray. The 93 introns that we resequenced are red. They are selected to sample the full range of GC contents, and their mean size is 94 ± 14 bp.

The observed indels lie in two distinct clusters. There are 10 rare indels of minor allele frequency $f < 0.06$, plus two common indels of minor allele frequency $f > 0.35$. The direction of the intron size change, relative to the major allele, is shown in Figure 3. All the rare indels drive the introns back toward their optimal size of 92 bp. The exceptions are the two common indels, which likely arose from different population dynamics. We further note that the probability of 10 indels in a row with the correct sign, under a null hypothesis that the sign is random, is 1×10^{-3}. To confirm that the major allele is the ancestral allele, we resequenced these indel-containing introns in a panel of 10 primates, ranging from chimpanzees to lemurs. Our polymerase chain reaction primers failed on the three most GC-rich introns, but in the other introns, the major allele agreed with the primate orthologs. The sole exception was in the most AT-rich intron, in which both human alleles were observed in different primates.

Minimal Introns Can Enhance the Export of Spliced mRNAs

These data indicate that at least for some genes, the presence of a minimal intron can be beneficial. Transgenic mice experiments (Choi et al. 1991; Palmiter et al. 1991) have long shown that some introns can affect expression levels. The most current explanation (Luo and Reed 1999; Zhou et al. 2000) is that "splicing generates a specific isolable complex that promotes rapid and efficient mRNA export." Thus, our conjecture is that minimal introns can affect mRNA maturation by coupling more efficiently to the biochemically linked machineries of splicing and export, thereby increasing the rate at which mRNA is exported from the cell nucleus.

Support for this conjecture can be found in the yeast expression data (Holstege et al. 1998). One must be careful not to mix up *S. cerevisiae* and *S. pombe*, because their minimal intron peaks are at very different sizes, and their genomes contain different sets of splicing-related proteins (Kaufer and Potashkin 2000). For this purpose, *S. cerevisiae* is more appealing because, with the 229 of the 6188 genes that do have introns, there is generally only one intron per gene. Moreover, there is a striking dichotomy in the types of introns found in different types of genes. Ribosomal-protein genes have nonminimal introns, but nonribosomal genes have minimal introns. As we show in Table 2, mRNA synthesis rates were 3.4 times higher in nonribosomal genes with minimal introns than in nonribosomal genes without introns. Ribosomal-protein genes with and without nonminimal introns showed no such differences in mRNA synthesis rates. Although there may be other explanations for this observation besides mRNA export, these data are not inconsistent with our conjecture.

Table 1. Insertion-deletion (indel) polymorphism summary

L(intron)	dL	f(minor)	GC(gene)	GC(intron)	Primates	intron sequence
73	1	0.033	0.387	0.164	95–99%	gtaatttaaaacttaaaattatttatttgattgtatttttattcatgtgctt aaagaattttTctttttttgtag
80	1	0.054	0.595	0.787	N/A	gtgagtcccagggtggggctggggaccgtgggacGgggggggtcccagccc tgccctcacgcccaccccaccgcccccag
81	1	0.034	0.595	0.778	N/A	gtggggcggggcccaggcggggaGgggggcccacgcagcggagcagccccaa catcccgcggccatctcccacccccaacag
90	1	0.022	0.595	0.700	N/A	gtgagtgggcaggacaggggcctggggtaggggacagcaagtgaCccccccc tccacagcccagtctgacccacccctccgtggccgcag
101	−3	0.011	0.361	0.317	88–100%	gtaagaaagcaggtgtctgcaaaaagtcatgtatcgatttattgtttgtaat gatacAGTagtatagcagataactaagacatattttcttgaatttgcag
120	−3 −4 −5 −1	0.021 0.010 0.010 0.010	0.516	0.292	100%	gtattttgtcactcttgaaagtttttattgggtaagaggttcatgccctttg tcctcattttTTCttcttgttattttatcTTTAtttaCTTTTtccacttca tgTttttttttcctttag
125	−1	0.021	0.411	0.376	92–98%	gtaaatgTtctcctctttgttcaactcttaagtttcacatccagaagtttcat acactgacaagttgtggctttgatctggtttttgcgtaaccttaaatatga ctttttttttccccaccccag
79	−2	0.354	0.374	0.234	100%	gtagtaaattacttaaattcaatttttccttgaaatAAgtgtgattagtaac ccattattatttccttttttattttcag
81	−1	0.375	0.516	0.370	100%	gtaggaagagtgggagttttgcaaatggacaacTtaaagatggggaagaga atcaaactacacttttttccttttttctag

From left to right, we list the intron size, the size change with respect to the major allele, the minor allele frequency, the gene level GC content, the intron level GC content, the degree of conservation relative to their primate orthologs, and lastly, the intron sequence itself with the indel in uppercase. A total of 42 polymorphic sites were identified. There were 30 single-base substitutions and 12 indels, at normalized rates of $\theta(\text{subst}) = 6.75 \times 10^{-4}$ and $\theta(\text{indel}) = 2.70 \times 10^{-4}$.

Additional support for our conjecture is observed in Drosophila populations with a 66-bp intron presence-absence polymorphism in the jingwei (jgw) gene. Absence of this minimal intron reduces the expression level by almost a factor of two (Llopart et al. 2002).

DISCUSSION

The general understanding is that many mRNAs are actively transported out of the nucleus, not passively diffused. Incompletely processed mRNAs are poor substrates for this export machinery, indicating that mRNA export is a type of quality control to ensure that only functional mRNAs reach the cytoplasm (Cullen 2000). We envision a competition to get out of the cell nucleus, with at least three different export paths: one each for mRNAs with no introns, mRNAs with minimal introns, and mRNAs with nonminimal introns. Perhaps minimal introns function as "routing" tags that define a more secure export path. Comparisons between species might not be simple. Particular export paths may not be present in some species, and orthologous genes need not use the same export path. For example, in *Encephalitozoon cuniculi* (Katinka et al. 2001) and in the nucleomorph chromosomes of *Guillardia theta* (Douglas et al. 2001), all introns are minimal (23 to 52 bp in *E. cuniculi* and 42 to 52 bp in *G. theta*), but they are in ribosomal-protein genes, the opposite of the situation for *S. cerevisiae*.

Our conclusion is that those genes that are reliant on the improvement in the rate of mRNA export that having minimal introns provide would be more resistant to intron expansion. Furthermore, any intron that drifts away from this optimum will be returned to it at the first opportunity. Assuming this is the correct explanation, it is difficult to see how such a complicated system of interacting molecules could ever be reconsti-

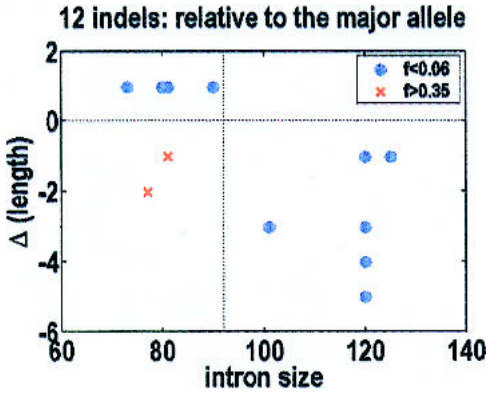

Figure 3 Insertion-deletion (indel) direction, relative to the major allele, as a function of intron size. For the 10 rare indels of minor allele frequency f < 0.06 (blue), the resultant size changes drive the introns back to their optimal size of 92 bp. The two common indels with f > 0.35 (red) are the only exceptions to this rule, presumably because they arose from a different population dynamics.

Table 2. *Saccharomyces cerevisiae* expression data, adapted from Holstege et al. (1998)

	Nonribosomal	Ribosomal-protein
No introns	5.17	91.9
With intron	17.6	91.7
	Minimal intron	Nonminimal intron

We classify the genes in two ways: based (1) on whether they are ribosomal-protein or nonribosomal genes and (2) on whether or not they have any introns. Ribosomal-protein genes tend to have nonminimal introns, and nonribosomal genes tend to have minimal introns. The 2 × 2 grid shows the averaged mRNA synthesis rate for those genes in each of the four categories.

tuted in a typical in vitro experiment. What we did, in effect, was let evolution perform the in vivo experiments for us and then query the results through a statistical analysis of the extant human polymorphisms. The interesting question is whether or not this methodology can be applied to other degenerate sequences with functional significance, such as promoter motifs associated with transcription regulation.

METHODS

We constructed high-quality databases of intron sequences, based exclusively on cDNA-to-genomic sequence alignments (Wong et al. 2000), for all of the multicellular eukaryotes for which a significant fraction of the genome had been finished. Resequencing was performed on the National Human Genome Research Institute (NHGRI)/Coriell Human Diversity Panel, which is a representation of all of the major ethnicities, including Northern European, Chinese, Indo-Pakistani, African American, Middle Eastern, Southwestern American Indian, Japanese, Mexican, and Puerto Rican (Collins et al. 1998). For ancestral alleles, we sequenced a primate panel from Coriell, which has chimpanzee, pigmy chimpanzee, lowland gorilla, orangutan, rhesus macaque, pig-tailed macaque, red-bellied tamarin, woolly monkey, black-handed spider monkey, and ring-tailed lemur. Polymerase chain reaction primers were designed from exon sequences flanking the selected introns. Sequencing was performed with dye-terminator chemistry on capillary sequencers. Every polymorphism, particularly the indels, was confirmed by visual inspection of the sequence traces. SNPs were submitted to GenBank/dbSNP with the handle UWGC (batches 2.12.2001.1 to 2.12.2001.4).

ACKNOWLEDGMENTS

We thank Maynard Olson, Lars Bolund, and Changqing Zeng for comments and suggestions. This analysis was partially supported by a grant from the National Institute of Environmental Health Sciences (1 RO1 ES09909).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ares, Jr., M., Grate, L., and Pauling, M.H. 1999. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5:** 1138–1139.

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270:** 2411–2414.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241:** 3–17.

Blencowe, B.J. 2000. Exonic splicing enhancers: Mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25:** 106–110.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22:** 231–238.

Carvalho, A.B. and Clark, A.G. 1999. Intron size and natural selection. *Nature* **401:** 344.

Choi, T., Huang, M., Gorman, C., and Jaenisch, R. 1991. A generic intron increases gene expression in transgenic mice. *Mol. Cell Biol.* **11:** 3070–3074.

Collins, F.S., Brooks, L.D., and Chakravarti, A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8:** 1229–1231.

Cullen, B.R. 2000. Nuclear RNA export pathways. *Mol. Cell Biol.* **20:** 4181–4187.

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.T., Wu, X., Reith, M., Cavalier-Smith, T., and Maier, U.G. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* **410:** 1091–1096.

Dredge, B.K., Polydorides, A.D., and Darnell, R.B. 2001. The splice of life: Alternative splicing and neurological disease. *Nat. Rev. Neurosci.* **2:** 43–50.

Goldstrohm, A.C., Greenleaf, A.L., and Garcia-Blanco, M.A. 2001. Co-transcriptional splicing of pre-messenger RNAs: Considerations for the mechanism of alternative splicing. *Gene* **277:** 31–47.

Grabowski, P.J. and Black, D.L. 2001. Alternative RNA splicing in the nervous system. *Prog. Neurobiol.* **65:** 289–308.

Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22:** 239–247.

Harpending, H. and Rogers, A. 2000. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* **1:** 361–385.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717–728.

Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., et al. 2001. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414:** 450–453.

Kaufer, N.F. and Potashkin, J. 2000. Analysis of the splicing machinery in fission yeast: A comparison with budding yeast and mammals. *Nucleic Acids Res.* **28:** 3003–3010.

Kimura, M. 1983. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, UK.

Lim, L.P. and Burge, C.B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci.* **98:** 11193–11198.

Llopart, A., Comeron, J.M., Brunet, F.G., Lachaise, D., and Long, M. 2002. Intron presence-absence polymorphism in Drosophila driven by positive Darwinian selection. *Proc. Natl. Acad. Sci.* **99:** 8121–8126.

Luo, M.J. and Reed, R. 1999. Splicing is required for rapid and efficient mRNA export in metazoans. *Proc. Natl. Acad. Sci.* **96:** 14937–14942.

McCullough, A.J. and Berget, S.M. 2000. An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5′ splice sites. *Mol. Cell Biol.* **20:** 9225–9235.

Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., Plumb, R.W., Sims, S.K., et al. 2000. An SNP map of human chromosome 22. *Nature* **407:** 516–520.

Nissim-Rafinia, M. and Kerem, B. 2002. Splicing regulation as a potential genetic modifier. *Trends Genet.* **18:** 123–127.

Ophir, R. and Graur, D. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* **205:** 191–202.

Palmiter, R.D., Sandgren, E.P., Avarbock, M.R., Allen, D.D., and Brinster, R.L. 1991. Heterologous introns can enhance expression of transgenes in mice. *Proc. Natl. Acad. Sci.* **88:** 478–482.

Petrov, D.A. 2001. Evolution of genome size: new approaches to an old problem, *Trends Genet.* **17:** 23–28.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Spingola, M., Grate, L., Haussler, D., and Ares, Jr., M.. 1999. Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* **5:** 221–234.

Wieringa, B., Hofer, E., and Weissmann, C. 1984. A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. *Cell* **37:** 915–925.

Wong, G.K.S., Passey, D.A., Huang, Y.Z., Yang, Z., and Yu, J. 2000. Is "junk" DNA mostly intron DNA? *Genome Res.* **10:** 1672–1678.

Wong, G.K.S., Passey, D.A., and Yu, J. 2001. Most of the human genome is transcribed. *Genome Res.* **11:** 1975–1977.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. 2002. The genome sequence of *Schizosaccharomyces pombe. Nature* **415:** 871–880.

Zhou, Z., Luo, M.J., Strasser, K., Katahira, J., Hurt, E., and Reed, R. 2000. The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature* **407:** 401–405.

WEB SITE REFERENCES

http://ftp.genome.washington.edu/RM/RepeatMasker.html; The `RepeatMasker` program is available at this site.

http://www.genome.washington.edu/projects/egpsnps; Data from the Environmental Genome Project.

Received August 9, 2001; accepted in revised form June 12, 2002.