



Fig. 2 Cytogenetic coordinate system for chromosomes 7 and 22 and localization of mapped BAC clones. To convert the cytogenetic position to a coordinate system, conventional chromosome bands were divided into arbitrarily defined intervals. This coordinate system allows for delineation of clone order even in the absence of other cytogenetic landmarks.

ping efforts of B. Trask and U.-J. Kim as well as the parallel and complementary efforts of the laboratory of J. Korenberg⁸.

Ilan R. Kirsch¹, Eric D. Green², Raluca Yonescu¹, Robert Strausberg³, Nigel Carter⁶, David Bentley⁶, Margaret A. Leversha⁶, Ian Dunham⁶, Valerie V. Braden², Eva Hilgenfeld¹, Greg Schuler⁴, Alex E. Lash⁴, Grace L. Shen⁵, Maria Martelli¹, W. Michael Kuehl¹, Richard D. Klausner³ & Thomas Ried¹

¹Genetics Department, Medicine Branch, National Cancer Institute; ²Genome Technology Branch, National Human Genome Research Institute; ³Office of the Director, National Cancer Institute; ⁴National Center for Biotechnology Information, National Library of Medicine; and ⁵Cancer Genetics Branch, National Cancer Institute, NIH, Bethesda, Maryland, USA. ⁶Sanger Centre, Cambridge, UK. Correspondence should be addressed to I.R.K. (e-mail: kirsch@exchange.nih.gov).

1. Bouffard, G.G. *et al. Genome Res.* **7**, 673–692 (1997).
2. Collins, J.E. *et al. Nature* **377**, 367–379 (1995).
3. Yunis, J.J. & Chandler, M.E. *Prog. Clin. Pathol.* **7**, 267–288 (1978).
4. Ried, T., Baldini, A., Rand, T.C. & Ward, D.C. *Proc. Natl Acad. Sci. USA* **89**, 1388–1392 (1992).
5. *ISCN 1995 An International System for Human Cytogenetic Nomenclature* (ed. Mitelman, F.) (Karger, Basel, 1994).
6. Jang, W., Chen, H.C., Sicotte, H. & Schuler, G.D. *Trends Genet.* **15**, 284–286 (1999).
7. Dunham, I. *et al. Nature* **402**, 489–495 (1999).
8. Korenberg, J.R. *et al. Genome Res.* **9**, 994–1001 (1999).

in the process of mapping clone sets for chromosomes 3 and 5 (N. Nowak), chromosome 14 (V. Cheung) and chromosome

12 (R. Kucherlapati). The use and power of this resource will certainly be aided by the related BAC development and map-

ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome

In recent years, the discovery of many regulatory elements within introns, the recognition of the regulatory potential of intronic and other non-protein coding RNAs, and the concept of a cellular 'ribotype' resulting from differences in RNA processing in different cells and tissues have led to increasing interest in the role of introns in enhancing eukaryotic genetic complexity, via alternative splicing and as both the recipient and donor of *cis*-acting and *trans*-acting elements^{1–4}.

To explore the evolution and function of introns in eukaryotes, we have developed an intron sequence information system (ISIS; <http://isis.bit.uq.edu.au/>) which

contains information on over 170,000 spliceosomal introns. Data in ISIS version 1 is based on intron-containing sequences from GenBank release 111. ISIS contains phylogenetic and protein homology categories, information about individual sequences and various bioinformatic analyses of taxonomical groupings of sequences using non-redundant subsets of the data. The database is searchable by Blast, GenBank attributes and elements that we have annotated within introns, and gives graphical views of gene structure and elements such as alternative coding regions, EST matches and repetitive sequences.

During analysis of this database, we found many EST matches within sequences annotated as introns, indicating that there are many previously unrecognized alternatively spliced exons, especially as many of these exons are conserved between species. Alternative splicing was first predicted by Walter Gilbert⁵, and subsequently verified by the discovery of cDNA isoforms exhibiting the addition or exclusion of whole or partial exons^{6,7}, although identification of such splice variants has largely occurred on an ad hoc basis. The development of large human EST (partial cDNA) sequence libraries over recent years, however, provides an opportunity to examine the incidence of alternative splicing globally by searching these libraries for exon skipping, exon truncation or inclusion of sequences currently described as intronic.

We examined the incidence of unrecognized exons in introns in 2,698 non-redundant human genes in ISIS which contain at least one complete sequenced intron and two flanking exons. We

excluded hypervariable immune-related genes and any genes with previously annotated alternative transcripts in GenBank. After removal of known repetitive sequences in the introns, we identified 3,119 EST clusters from 1,122 genes (42%) containing sequences previously annotated as intronic. The presence of such sequences in the EST databases may be the result of genuine alternative splicing events or pre-mRNA contamination of preparations used to construct libraries. We therefore discarded any cases of whole-intron retention (although some may represent genuine splice variants^{8,9}), as well as all other ambiguous cases in which the EST sequences were indistinguishable from the genomic sequence.

The remaining 209 clusters (186 genes) had unequivocal alternative splicing events, with cryptic exons, 5' exon extensions and 3' exon extensions occurring in roughly equal proportions.

We also examined the frequency of exon skipping and length variation by analysing EST sequences that crossed exon/exon boundaries, identifying 507 genes in which known exons were absent or altered in length relative to the GenBank annotation. Combining these two sets, we identified 582 different human genes showing unequivocal alternative splicing via exon insertion, extension, truncation or deletion. This represents 22% of genes analysed, many of which exhibit multiple alternative splicing events. This reveals an unexpectedly high

frequency of alternative splicing in genes not previously known to be alternatively spliced. Moreover, this analysis provides a very conservative estimate, given the fragmentary EST coverage, the 3' bias of most EST libraries and the removal of all ambiguous or indeterminate cases. Furthermore, our analysis did not detect any new exons or exon truncations/extensions at the 5' and 3' ends of transcripts, which are common^{10,11}.

A recent comparison of 475 disease-associated human protein sequences with the human EST database¹² suggested that as many as one in three may be alternatively spliced, although because this study was not based on a data set of whole genomic DNA for each gene, the background of incomplete splicing/pre-mRNA contamination could not be fully assessed. Similar estimates were obtained for another sample of 392 genes¹⁰. Given the incomplete EST and intron sequence coverage in the databases, we anticipate that the real frequency of alternative splicing in human genes will be much greater than we have been able to measure, suggesting that there may be several hundred-thousand different mRNAs and protein isoforms produced in different cells and tissues at various stages of development and under different physiological conditions.

This complexity also suggests microarray systems for analysing gene expression will need to be expanded beyond a single coding sequence per gene to include all possible

alternative exons on the array, if we are to properly understand the genetic output and cellular circuitry of higher organisms.

Full details of the experimental methods and results are available (http://isis.bit.uq.edu.au/a_splicers.html).

Acknowledgements

We thank B. Huang and D. Kennedy for helpful comments and discussions.

Larry Croft^{1,2*}, Soeren Schandorff^{3*}, Francis Clark^{1,2*}, Kevin Burrage², Peter Arctander³ & John S. Mattick¹

*These authors contributed equally to this work.

¹ARC Special Research Centre for Molecular and Cellular Biology, ²Department of Mathematics, University of Queensland, Brisbane, Queensland, Australia. ³Department of Evolutionary Biology, University of Copenhagen, Copenhagen, Denmark. Correspondence should be addressed to J.S.M. (e-mail: j.mattick@cmcb.uq.edu.au).

1. Mattick, J.S. *Curr. Opin. Genet. Dev.* **4**, 823–831 (1994).
2. Herbert, A. & Rich, A. *Nature Genet.* **21**, 265–269 (1999).
3. Fire, A. *Trends Genet.* **15**, 358–363 (1999).
4. Chabot, B. *Trends Genet.* **12**, 472–478 (1996).
5. Gilbert, W. *Nature* **271**, 501 (1978).
6. Breitbart, R., Andreadis, A. & Nadal-Ginard, B. *Annu. Rev. Biochem.* **56**, 467–495 (1987).
7. Adams, M., Rudner, D. & Rio, D. *Curr. Opin. Cell Biol.* **8**, 331–339 (1996).
8. McKeown, M. *Annu. Rev. Cell Biol.* **8**, 133–155 (1992).
9. Smith, C.W.J., Patton, J.G. & Nadal-Ginard, B. *Annu. Rev. Genet.* **23**, 527–577 (1989).
10. Mironov, A., Fickett, J. & Gelfand, M. *Genome Res.* **9**, 1288–1293 (1999).
11. Gautheret, D., Poirot, O., Lopez, F., Audic, S. & Claverie, J.M. *Genome Res.* **8**, 524–530 (1998).
12. Hanke, J. et al. *Trends Genet.* **15**, 389–390 (1999).