

Protein Diversity as a Result of Transcript Isoform Variation in Expressed Genes

Winston Hide, Vladimir N. Babenko, Peter A. van Heusden, and Janet F. Kelso

South African National Bioinformatics Institute, University of the Western Cape, private bag x17, Bellville, South Africa, 7535. [winhide\(AT\)sanbi.ac.za](mailto:winhide(AT)sanbi.ac.za)

Completion of the human genome sequence has provided evidence for a gene count with a lower bound of 27 500 – 40 000. Significant protein complexity encoded from this gene set must derive in part from multiple transcript isoforms. Recent studies utilizing ESTs and reference mRNA data have revealed that alternate transcription, including alternative splicing, polyadenylation and transcription start sites, occurs within at least 30-40 % of human genes. It is likely that this is an underestimate as EST sampling has been used to derive the figures. Transcript form surveys have yet to integrate the genomic context, expression, and contribution to protein diversity of isoform variation. Exhaustive manual confirmation of genome sequence annotation, coupled with comparison to available expressed sequence data has been used here to accurately associate isoforms showing exon skipping with genomic sequence to reveal potential protein coding alteration. In addition, relative expression levels of transcripts have been estimated from EST representation in the public databases. Our rigorous method has been applied to 545 described genes in the first intensive study of exon skipping based on chromosome 22 genome sequence and matched human transcripts. The study has led to the discovery of 62 exon skipping events in 52 genes, with 57 exon skips altering the protein coding region. A single gene, (FBXO7) expresses an exon repetition. EST sampling analysis indicates that 58.8% of highly represented multi-exon genes are likely to express exon-skipped isoforms in ratios that vary from 1:1 to 1:>100. Comparisons with mouse show a similar overall level of skipping, although not at the same exon boundaries/genes. Analysis of cancer genes show that aberrant forms of skipping may segregate with cancer expression libraries.