Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms, and RNA editing

Supplementary text 1 - Filtering mRNA sequences

We started with a total of 34,431 sequences based on downloads from RefSeq and MGC on May, 16th, 2003. When we downloaded updated sequences on August 1st, 2003, 277 sequences in this original set are absent, some due to our analysis in the intervening time. From this set, we further remove non-human sequences that were identified using nucleotide BLAST (Altschul et al., 1990) against the nr database.

We loosely define chimeric mRNA clones as those for which there is strong evidence that the sequence does not originate from a traditional set of adjacent and consecutively ordered group of exons. Primarily, this consists of sequences that align to two or more distinct locations in the genome, usually on separate chromosomes. These most likely are an artifact of the cloning process in which two pieces of unrelated transcribed RNA become joined before being inserted into the cloning vector (J. Schmutz, personal communication). We are aware that some of these discarded sequences may be examples of trans-splicing, multi-locus transcription, or other special situations (see (Romani et al., 2003) and references therein). It is our contention that the vast majority of chimeric clones we identified are artifacts.

Based on our identification and analysis of non-human and chimeric sequences, hundreds of RefSeq entries have been updated or discontinued over the past year. The MGC has also utilized our analysis in evaluating clones from their collection. In many cases, corrections were made before we downloaded the final sequences minimizing the number of sequences discarded. This analysis prompted the removal of 156 sequences.

In addition, many RefSeq mRNA sequences are based partially or wholly upon a sequence in MGC. There are 3,166 cases where a single MGC clone is completely contained within the RefSeq sequence, and that is the only

MGC clone used in the RefSeq sequence. For these, the MGC sequence was removed to eliminate redundancy.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.

Romani, A., Guerra, E., Trerotola, M., and Alberti, S., 2003. Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res.* **31**(4):e17.