Supplementary materials 2: Antisense ncRNA pipeline

Mapping of human and mouse “mRNA” (cDNA) and EST sequences to respective genome assemblies (builds hg16 and mm4) and GENSCAN exon predictions for the same assemblies were obtained from the UCSC Genome Browser Database (1). The human mapping was dated March 1, 2004 and the mouse mapping May 9, 2004. (Note: we are aware of problems associated with the mm4 mouse assembly, such as local reversal artifacts. Our manual inspection of numerous sense : antisense predictions indicate that those problems only affect a negligible proportion of cases. We are currently repeating the pipeline predictions on the mm5 assembly, which should become available shortly.)

A mapped cDNA sequence was considered a putative ncRNA if it fulfilled the following criteria: (i) it had no protein coding sequence annotation; (ii) it did not map within 1,500bp (on the same strand) of a coding sequence annotated in another cDNA (designed to exclude UTR fragments); (iii) it did not overlap by >200bp any GENSCAN exons; (iv) it did not have an ORF larger than 100aa; and (v) it did not overlap the mapping of any transcript classified as coding by these criteria. The fifth criterion was applied iteratively until the number of ncRNAs did not change. The comparisons with other mappings and GENSCAN exons were made only between elements that mapped onto the same genomic strand. In the ORF search, virtual cDNAs corresponding to all combinations of mismatches between transcript and genome sequence were searched in order to account for sequence errors. To focus on small mismatches, which are more likely to correspond to sequence errors, simultaneous transcript and genome sequence inserts (unaligned regions) were ignored if one insert was three times the size of the other or more. Single-sequence inserts (gaps in one sequence only) larger than 3bp were also ignored. Lastly, we excluded putative internally primed transcripts by searching the genomic region [-10,+14] relative to the 3’-end of mappings for 14-base widows containing 10 or more As. Of 152,507 human and 130,032 mouse cDNAs, 5,161 and 6,811 passed the filters.

Separately, all cDNA and EST mappings were fed into a pipeline designed to refine mappings, filter out low-quality mappings and reliably assign mappings to the correct genomic strand. The steps of this pipeline were as follows:

(1) If a transcript sequence begins or ends a few bases into an exon, those bases are likely to be either unmapped or incorrectly mapped by spliced alignment programs due to sensitivity issues. This problem is augmented by low sequence quality at EST ends and as a result regions larger than 100bp can be unmapped. To minimize this problem, we applied an algorithm that extends mapping ends using information about exon positions from neighboring mappings (P.G.E. and B.L., unpublished).

(2) Mappings with <150 nt and <75% of the transcript sequence mapped were discarded, as were mappings with a percent identity below 98% (for cDNAs) or 97.5% (for ESTs). All mappings were scored according to a scoring scheme designed to select against processed pseudogenes: mapping score = number of matches – number of mismatches + 5 · number of introns – gap penalty, where introns were defined as genome sequence inserts of at least 20 bp with GT-AG, GC-AG or AT-AC junctions and the gap penalty was log2(gap length)+1 summed over all non-intron

gaps. For each transcript sequence, only the best-scoring mappings were retained. Transcript sequences with three or more best-scoring mappings were discarded.

(3) Mappings of sequences annotated with the same cDNA clone id were merged if possible.

(4) Each mapping was assigned to a genomic strand (plus or minus) that should correspond to the sense strand of the gene identified by the mapping, or excluded if strand assignment was not possible. A mapping with introns (defined as in step 2) was oriented according to the first and last dinucleotides of the introns (splice signals). An intronless mapping was oriented according to a combined assessment of poly-A tails, polyadenylation signals and, for EST mappings, annotated read direction. Where read direction was the only information available to orient an EST mapping, we discarded that mapping unless the EST was from a library we found to have reliable directional annotation. Such libraries were identified by comparing splice signals to directional annotation for all spliced ESTs. A library was considered to have reliable directional annotation if it contained at least 100 spliced and directionally annotated ESTs and splice signals and annotation agreed for >99% of those ESTs.

(5) To detect and discard transcript sequences resulting from priming at A-stretches in genomic DNA or upstream of the poly-A-tail in RNA transcripts, the criterion described in the second paragraph was used.

(6) Intronless EST mappings that overlapped no other mapping or just another intronless EST mapping were discarded.

Using the resulting mappings, putative noncoding cis-antisense transcripts were identified by, for each mapping of a putative ncRNA, searching for mappings to the other strand that overlapped the ncRNA mapping by at least 20 bp within exons. Where cis-antisense partners were identified, the relative orientation of transcription was categorized either as convergent, divergent or full overlap and the locations of overlaps were noted.

Finally, to reduce the possibility of contaminating coding transcripts having gone previously undetected by the pipeline, BLASTX searches of repeat-masked transcript sequences against a reference protein set, UniRef90 (2), were performed and any significant hits (E-value <1e-5) removed. This left 668 human and 624 mouse putative antisense ncRNAs forming 579 and 571 distinct transcriptional units, defined according to Okazaki *et al.*, 2002 (3).

References

1. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) *Nucleic Acids Res*, 31, 51-54.
2. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2004) *Nucleic Acids Res*, 32 Database issue, D115-119.
3. Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002) *Nature.*, 420, 563-573.