Supplementary Methods

Identification of the longest common substrings

To determine the sequence common to all transcripts we have developed an approach based on suffix arrays. Suffix arrays were developed for rapid sequence searches, and involve building all possible suffixes of a given string, sorting them and then performing a binary search to determine if a particular sequence occurs or not. We build the suffixes of all the transcript sequences, whereby keeping track of the identity of each suffix by an identifier, which related it to its parent sequence. The suffixes were then sorted lexically, which leads to common substrings occurring next to each other in the sorted list. We scanned the sorted suffixes in a window size equal to the number of transcripts of a respective gene with a step-size of one. For each window analyzed, suffixes were checked for their parent identity and only those common substrings were considered, whose members belonged to a different parent sequence. For genes with highly variable transcript sequences that showed less than 300bp overlap, we extended this approach to combinatorially eliminate transcripts that did not show sufficiently long similarity to the remaining set of transcripts of the respective gene. Transcripts with insufficiently long overlaps were treated individually in esiRNA and primer design.

Two-step PCR approach for the generation of templates for esiRNA production

Transcript-specific primers appended 5’ with TCACTATAGGGAGAG (forward primer)
or TCACTATAGGGAGAC (reverse primer) were synthesized by MWG Biotech.

A first 25 µl PCR reaction containing 1 X NH₄ reaction buffer, 1.6 mM MgCl₂, 0.8 mM dNTPs (all from Bioline), 0.4 µM of each transcript-specific forward and reverse primer, 70 ng cDNA (PCR amplified from the MegaMan™ human transcriptome library from Statagene), and 1 unit BIOTAQ Red polymerase (Bioline) was initially denatured for 2 minutes at 94°C, annealed for 30 seconds at 60°C, and extended for 30 seconds at 72°C, followed by 6 cycles of 2 minutes at 94°C, 30 seconds at 60°C, and 30 seconds at 72°C, followed by 6 cycles of 2 minutes at 94°C, 30 seconds at 62°C, and 30 seconds at 72°C, followed by 21 cycles of of 2 minutes at 94°C, 30 seconds at 65°C, and 30 seconds at 72°C, and a final elongation step of 5 minutes at 72°C.

The full-length T7 promoter sequence was appended by a second 50 µl PCR reaction containing 1 X NH₄ reaction buffer, 2.0 mM MgCl₂, 0.8 mM dNTPs, 0.2 µM of the universal primers T7-G (GCTAATACGACTCACTATAGGGAGAG) and T7-C (GCTAATACGACTCACTATAGGGAGAc), 2 µl of the first PCR reaction, and 2 units BIOTAQ Red polymerase. The PCR reaction was initially denatured for 2 minutes at 94°C, annealed for 30 seconds at 60°C, and extended for 30 seconds at 72°C, followed by 39 cycles of 30 seconds at 94°C, 30 seconds at 60°C, and 30 seconds at 72°C, and a final elongation step of 5 minutes at 72°C.

After the second PCR 3 µl PCR product was analyzed on a 2% PCR agarose gel. The PCR products were arrayed in a manner that resulted in an alternating size pattern of the fragments to enable the judgment of the correct size.

**Hexamer complementarity**
The enrichment for transcripts that contain a given hexamer within their 3’ UTR was carried out by first tabulating all the human genes with LocusLink IDs that contained the hexamer within their annotated 3’ UTR, according to Unigene 161 Hs.seq.uniq. The probability of overlap of the microarray signature with a hexamer set was calculated with the hypergeometric cumulative distribution function using a background set of all the genes on the array that contained LocusLink annotated 3’ UTRs.

References