

## **Web Note A. Computational Subtraction (Methods)**

**Sources of sequence databases.** Our analysis was performed using sequence available from the NCBI (<http://ncbi.nlm.nih.gov>), Celera Genomics (<http://www.celera.com>), and the Genetic Information Research Institute (GIRI) (<http://www.girinst.org>). Human EST sequences (dbEST)<sup>1</sup> & library information, the HGP genomic sequences (phases 0-3), the RefSeq human mRNA set, and UniVec vector sequences were downloaded from NCBI on March 6, 2001. The “nt” BLAST databases were downloaded from NCBI on March 23, 2001. The human mitochondrial genome sequence is GenBank accession # NC\_001807. The Celera draft of the human genome and shotgun sequence from the mouse genome were downloaded from their website in January 2001. RepBase 6.2 was downloaded from the GIRI on March 7, 2001. The estimated fold-coverage at time of download was >6.5-fold for the public human genome sequence, ~8-fold for the Celera human genome sequence, and ~3-fold for the Celera mouse genome sequence. Celera estimates ~95% coverage of the euchromatic portion of the human genome<sup>2</sup>, whereas the HGP estimates ~96% coverage of euchromatin and ~94% coverage of the whole genome<sup>3</sup>.

**Algorithms & Computation.** Both the filtering of EST's and the matching of EST's to nucleotide databases was achieved using the MEGABLAST algorithm<sup>4</sup>. We used default BLAST affine gapping penalties in all cases with a word-size of 24, 20, or 16 for nucleotide-nucleotide matching<sup>5</sup>. The MEGABLAST executable is available from the NCBI (<http://ncbi.nlm.nih.gov>). Species were assigned to matching records based on GenBank annotation of sequence records and the NCBI Taxonomy database.

**Data processing.** We began with a set of 3,287,578 human expressed sequence tags (EST's). These EST's were serially compared against the seven filter databases using the MEGABLAST tool with a word-size of 24 (see Fig. 1 in main text). An alignment score of

greater than or equal to 60.0 bits (equivalent to 30 consecutive identical nucleotides, and roughly corresponding to an e-value of  $1e^{-7}$ ) was considered to have matched a given EST to a filter. The filters used, in the order applied, were (1) RefSeq human mRNAs (2) human mitochondrial genome (3) the UniVec vector database (4) the RepBase human repeat database (5) the HGP phase 0-3 human genome sequence (6) the Celera assembled human genome sequence (7) the Celera mouse genomic fragments database. EST's that matched were removed at each step to facilitate the computational tractability of the problem. 137,011 EST's passed all seven filters. To improve the sensitivity of the filtering process, these EST's were re-run against the seven filters at a lower word-size (20) and matching EST's were removed. The process was again repeated with a word-size of 16, leaving 102,009 EST's. This set was further filtered for short (< 150 nucleotides) and ambiguous (> 2.5% N's) sequences, leaving 65,839 EST's that we considered unmatched (see Web Note H for further discussion of poor-quality sequences & quality controls). These EST's were compared to the GenBank "nt" nucleotide database, using the MEGABLAST (word-size of 16) algorithm.

Alignments to the "nt" database were categorized by the species annotation of the GenBank record to which a given EST aligned. Near-exact matches, defined as alignments scoring greater than 250 bits (roughly equivalent to an e-value of  $1e^{-64}$ , or a perfect alignment over 125 bases) were used to construct a searchable database that is available at: <http://research.dfci.harvard.edu/meyersonlab/comsub.html>. Where possible, detailed information on each alignment and the library from which a given EST originated has been made available. The website permits query by species category & minimum alignment score, as well as the flexible application of library quality filters (based on library source & quality, as discussed in Web Note F).