SUPPLEMENTARY METHODS

Genomic alignments of human mRNAs. We used mRNA alignments to examine the correspondence of UCEs with genes[tr] and exons as well as the conservation of human exons. We complemented the dominant splice variants in Refseq with expressed sequence tags in UniGene. To determine the corresponding chromosomal loci, we aligned human Refseq mRNAs and UniGene cluster representatives to the genome with Mega BLAST (options: -p 90, -W 20, -F "m D"); UniGene representatives served as genomic anchors for subsequent alignments of their entire clusters (below). Sequences were aligned to the entire genome when the chromosome was not annotated. As is expected for any mRNA aligned to an entire chromosome, this step typically produced multiple invalid alignments per transcript, which we resolved subsequently. To obtain precise genomic alignments, we then used Splign to realign the transcripts to the chromosomal loci obtained from Mega BLAST, with allowable boundaries extended by 50 Kbp on each side. We then examined splice signals in order to identify transcripts sequenced from the reverse strand as well as invalid alignments. We determined the strand of the alignment by assigning a score of +1 for GT-AG, GC-AG and AT-AC splice signals, -1 for their reverse complements, otherwise +0.5 for a GT donor and -0.5 for an AC acceptor (which is likely to be a reverse-complemented GT donor). We summed these scores for each alignment, reverse-complemented alignments with a negative sum, then split alignments at junctions with signals other than GT-AG, GC-AG and AT-AC.

For each transcript, we determined the best alignment as that with the greatest fraction of matching bases. Because multiple alignments to distinct loci may be equally valid, we retained alignments whose fraction of matching bases was at least 99% of the maximum. We used Splign to align UniGene clusters to the loci of their representatives, with boundaries extended by 100 Kbp, and repeated the procedures for correction of splice signals and extraction of best alignments. We then excluded NCBI gene^{tr} models (identified as XM_ or XR_ followed by a number), which are predicted RNAs, and retained only intron-spanning alignments within boundaries of Refseq RNAs.

Analysis of exonic and intronic sequences. Using genomic alignments of human Refseq and UniGene mRNAs (see above), we examined the overlap of UCEs with exons. For all intron-spanning aligned mRNAs, we considered initial and terminal exons (i.e., aligned fragments) of at least 225 bp to be UTRs, since that is the point at which the fraction of internal exon lengths falls below those of first and last exons (not shown). For exonic fragments of UCEs and for all exons, we examined the significance of overlaps of different exon types with random trials (one million iterations for each combination).

Identification of best matches for UCEs. The best match for each UCE within the human genome was determined by aligning each UCE to the genome with Mega BLAST using the minimum word length (12) and with filtering disabled, discarding matches overlapping the UCE itself, calculating % identity as the fraction of matching bases in each match, and keeping the match with the highest identity.

Comparison of UCE overlaps with similarly annotated genes[tr]. In order to ascertain whether the depletion of UCEs among SDs or CNVs was a property of the UCEs themselves or of the types of genes[tr] in or near which they occur, we conducted random trials of overlap wherein random sequences were drawn from corresponding locations in or near genes[tr] with functional annotations akin to those containing or flanking UCEs.

Functional annotations for Refseq mRNAs were obtained from the Gene Ontology project (www.geneontology.org). For the three sets of genes[tr] overlapping exonic UCEs, overlapping intronic UCEs, or flanking intergenic[tr] UCEs on either side, all functional annotations were tallied, and Fisher's exact test was used to calculate the enrichment of each functional category relative to the background set of all genes[tr]. In cases where genes[tr] overlapped, only a single instance of each functional annotation was retained. For intergenic[tr] UCEs, the functional enrichment was calculated after including both genes[tr] flanking each UCE, without regard to preferential proximity to either gene[tr]. Annotations with $P < 0.001$ were retained, yielding results similar to those reported by Bejerano et al.[1], such as transcription and related functions for intronic and intergenic[tr] UCEs, and mRNA binding and splicing for exonic UCEs (data not shown).

For the three lists of enriched functional annotations, all genes[tr] with those annotations (i.e., throughout the genome) were collected, leading to three sets of genes[tr]: one for exonic UCEs (GO exonic gene set), another for intronic UCEs (GO intronic gene set), and a third for intergenic[tr] UCEs (GO intergenic[tr] gene set); genes[tr] containing exonic UCEs, containing intronic UCEs, or flanking intergenic[tr] UCEs were removed from these

sets, respectively. New random trials of overlap were conducted where, for exonic (intronic) UCEs, random sequences were drawn from the conserved portions of exons (introns) in the GO exonic (intronic) gene^{tr} set. For intergenic^{tr} UCEs, random sequences were obtained from the conserved portions (see below) of intergenic^{tr} regions flanking the GO intergenic^{tr} genes^{tr}. One million random trials were run for each set of UCEs with SDs then with CNVs, and observed overlaps were then compared with the distributions of random overlaps. Conserved sequences were those contained in 200-mers at least 1% conserved (see below) in mouse and rat, mouse and dog, or chicken, spanning ~1 Gbp and approximately equivalent to human sequences conserved in the mouse genome.

Restriction of overlap to non-repetitive genomic regions. After noting that UCEs consist almost entirely (99.85%) of non-repetitive sequence, we conducted tests of overlap wherein randomly chosen sequences were taken only from non-repetitive stretches of the genome. Since we required chosen sequences to lie entirely within non-repetitive fragments, we deleted all fragments < 200 bp in length, leaving 1.3 Gbp. For each sequence, we then chose a random location between 0 and 1.3 billion minus the length of the sequence, found the chromosome and non-repetitive fragment to which that location corresponded, confirmed that the sequence would fit entirely within the selected fragment, found the corresponding position in the whole genome, calculated the overlap with the dataset in question (SDs, etc.), and summed the overlaps for the entire set of randomly chosen sequences corresponding to UCEs. This procedure was conducted one million times for each dataset.

Clustering of UCEs. Because UCEs often lie close to each other in what appear to be clusters, we repeated the random trials of overlap while taking clustering into consideration. We first chose a distance threshold based on the distribution of inter-UCE distances, then created clusters by joining UCEs within this set distance of each other. By this definition, clusters often contained more than two UCEs and extended beyond the chosen threshold distance. Each cluster was treated as a unit, with its inter-UCE distances maintained and other sequences and clusters disallowed from overlapping it.

Identification of imperfectly conserved sequences. We wondered to what extent the depletions we observed could be attributed to the strict requirement of 100% conservation for UCEs. To address this question, we sought imperfectly conserved sequences defined as similarly as possible to UCEs. Compared to the selection of perfectly conserved sequences, the selection of imperfectly conserved sequences is complex. For example, a long fragment with 98% conservation contains shorter fragments showing 100% conservation. We therefore analyzed the human genome as a series of 200-bp fragments, tiled every bp, starting at the first position of every orthologous block in mouse, rat, dog and chicken.

Since sequences in orthologous blocks were already aligned to each other, we calculated the conservation of each fragment with a simple scoring scheme: +1 for each match, 0 for each mismatch (including deletions in the ortholog), and -1 for each one-bp deletion in human. We then divided the total by 200. For human sequences not containing gaps, this score was equal to the fraction of matching bases. We then determined the final level of

conservation for each fragment in a manner akin to that with which we defined the combined set of 896 UCEs; that is, we used the maximum level of human-mouse-rat (HMR), human-dog-mouse (HDM) and human-chicken (HC) conservation. HMR conservation was the minimum of human-mouse and human-rat conservation, and HDM conservation was the minimum of human-dog and human-mouse conservation. We then joined overlapping fragments with equal conservation. For comparison, this procedure yielded a set of perfectly conserved tiled 200-mers that is 98.6% concordant with the combined set of 896 UCEs.

Finally, because we wished to conduct analyses with imperfectly conserved sequences while gradually decreasing conservation in 1% decrements, we subtracted fragments with 100% conservation from those with ≥ 99% conservation, retained only those of at least 200 bp for an analysis of depletion, and then subtracted these from the set with ≥ 98% conservation for the subsequent analysis of depletion, and so forth.

As was done with UCEs, the overlaps of these sets of imperfectly conserved sequences with the pooled SDs of the Scherer[12] and Eichler[14] groups, the pooled CNVs, as described in the text, and the pooled DELs were then compared with the overlaps obtained using sets of sequences that were chosen at random from the conserved portion of the human genome and that also matched our sets of imperfectly conserved elements in number and length, with 1,000 random trials conducted for each conservation threshold and dataset. Note that our selection of random sequences from just the 35% of the genome that is

conserved (i.e., at least 1% by the methods described here) greatly increased the stringency of this analysis.