Distinguishing protein-coding and noncoding genes in the human genome

Michele Clamp*†, Ben Fry*, Mike Kamal*, Xiaohui Xie*, James Cuff*, Michael F. Lin‡, Manolis Kellis*‡, Kerstin Lindblad-Toh*, and Eric S. Lander*†§¶‖

*Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, MA 02142; ¶Department of Biology and ‡Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; §Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142; and ‖Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, October 3, 2007 (sent for review August 1, 2007)

Although the Human Genome Project was completed 4 years ago, the catalog of human protein-coding genes remains a matter of controversy. Current catalogs list a total of ≈24,500 putative protein-coding genes. It is broadly suspected that a large fraction of these entries are functionally meaningless ORFs present by chance in RNA transcripts, because they show no evidence of evolutionary conservation with mouse or dog. However, there is currently no scientific justification for excluding ORFs simply because they fail to show evolutionary conservation: the alternative hypothesis is that most of these ORFs are actually valid human genes that reflect gene innovation in the primate lineage or gene loss in the other lineages. Here, we reject this hypothesis by carefully analyzing the nonconserved ORFs—specifically, their properties in other primates. We show that the vast majority of these ORFs are random occurrences. The analysis yields, as a by-product, a major revision of the current human catalogs, cutting the number of protein-coding genes to ≈20,500. Specifically, it suggests that nonconserved ORFs should be added to the human gene catalog only if there is clear evidence of an encoded protein. It also provides a principled methodology for evaluating future proposed additions to the human gene catalog. Finally, the results indicate that there has been relatively little true innovation in mammalian protein-coding genes.

comparative genomics

An accurate catalog of the protein-coding genes encoded in the human genome is fundamental to the study of human biology and medicine. Yet, despite its importance, the human gene catalog has remained an elusive target. The twofold challenge is to ensure that the catalog includes all valid protein-coding genes and excludes putative entries that are not valid protein-coding genes. The latter issue has proven surprisingly difficult. It is the focus of this article.

Putative protein-coding genes are identified based on computational analysis of genomic data—typically, by the presence of an open-reading frame (ORF) exceeding ≈300 bp in a cDNA sequence. The underlying premise, however, is shaky. Recent studies have made clear that the human genome encodes an abundance of non-protein-coding transcripts (1–3). Simply by chance, noncoding transcripts may contain long ORFs. This is particularly so because noncoding transcripts are often GC-rich, whereas stop codons are AT-rich. Indeed, a random GC-rich sequence (50% GC) of 2 kb has a ≈50% chance of harboring an ORF ≈400 bases long [supporting information (SI) Fig. 4].

Once a putative protein-coding gene has been entered into the human gene catalogs, there has been no principled way to remove it. Experimental evidence is of no utility in this regard. Although one can demonstrate the validity of protein-coding gene by direct mass-spectrometric evidence of the encoded protein, one cannot prove the invalidity of a putative protein-coding gene by failing to detect the putative protein (which might be expressed at low abundance or in different tissues or at different developmental stages).

The lack of a reliable way to recognize valid protein-coding transcripts has created a serious problem, which is only growing as large-scale cDNA sequencing projects yield ever-larger numbers of transcripts (2). The three most widely used human gene catalogs [Ensembl (4), RefSeq (5), and Vega (6)] together contain a total of ≈24,500 protein-coding genes. It is broadly suspected that a large fraction of these entries is simply spurious ORFs, because they show no evidence of evolutionary conservation. [Recent studies indicate that only ≈20,000 show evolutionary conservation with dog (7).] However, there is currently no scientific justification for excluding ORFs simply because they fail to show evolutionary conservation; the alternative hypothesis is that these ORFs are valid human genes that reflect gene innovation in the primate lineage or gene loss in other lineages. As a result, the human gene catalog has remained in considerable doubt. The resulting uncertainty hampers biomedical projects, such as systematic sequencing of all human genes to discover those involved in disease.

The situation also complicates studies of comparative genomics and evolution. Current catalogs of protein-coding genes vary widely among mammals, with a recent analysis of the dog genome (8) reporting ≈19,000 genes and a recent article on the mouse genome (2) reporting at least 33,000 genes. The difference is attributable to nonconserved ORFs identified in cDNA sequencing projects. It is currently unclear whether it reflects meaningful evolutionary differences among species or simply varying numbers of spurious ORFs in species with more cDNAs in current databases. In addition, the confusion about protein-coding genes clearly complicates efforts to create accurate catalogs of non-protein-coding transcripts.

The purpose of this article is to test whether the nonconserved human ORFs represent bona fide human protein-coding genes or whether they are simply spurious occurrences in cDNAs. Although it is broadly accepted that ORFs with strong cross-species conservation to mouse or dog are valid protein-coding genes (7), no work has addressed the crucial issue of whether nonconserved human ORFs are invalid. Specifically, one must reject the alternative hypothesis that the nonconserved ORFs represent (*i*) ancestral genes that are present in our common mammalian ancestor but were lost in mouse and dog or (*ii*) novel genes that arose in the human lineage after divergence from mouse and dog.

Here, we provide strong evidence to show that the vast majority of the nonconserved ORFs are spurious. The analysis begins with a thorough reevaluation of a current gene catalog to identify conserved protein-coding genes and eliminate many putative genes resulting from clear artifacts. We then study the remaining set of nonconserved ORFs. By studying their properties in primates, we

Author contributions: M.C. and E.S.L. designed research; M.C., B.F., M. Kamal, X.X., J.C., M.F.L., M. Kellis, K.L.-T., and E.S.L. performed research; M.C., B.F., M. Kamal, X.X., J.C., M.F.L., M. Kellis, K.L.-T., and E.S.L. analyzed data; and M.C. and E.S.L. wrote the paper.

The authors declare no conflict of interest.

†To whom correspondence may be addressed. E-mail: mclamp@broad.mit.edu or lander@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709013104/DC1.

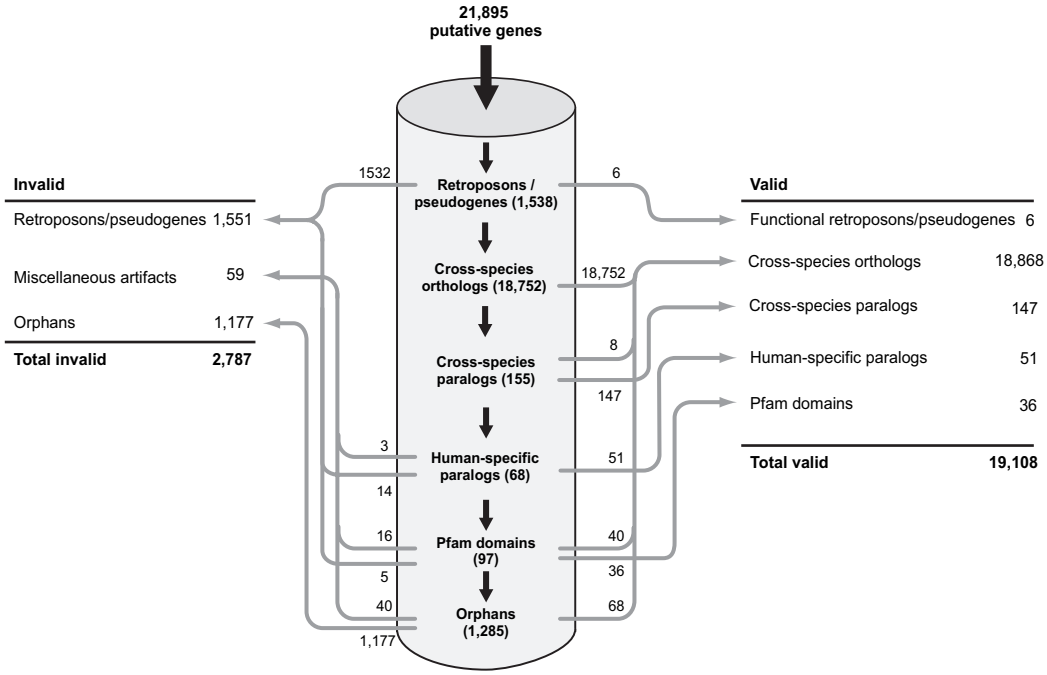

Fig. 1. Flowchart of the analysis. The central pipeline illustrates the computational analysis of 21,895 putative genes in the Ensembl catalog (v35). We then performed manual inspection of 1,178 cases to obtain the tables of likely valid and invalid genes. See text for details.

show that the vast majority are neither (*i*) ancestral genes lost in mouse and dog nor (*ii*) novel genes that arose after divergence from mouse or dog.

The results have three important consequences. First, the analysis yields as a by-product a major revision to the human gene catalog, cutting the number of genes from ≈24,500 to ≈20,500. The revision eliminates few valid protein-coding genes while dramatically increasing specificity. Second, the analysis provides a scientifically valid methodology for evaluating future proposed additions to the human gene catalog. Third, the analysis implies that the mammalian protein-coding genes have been largely stable, with relatively little invention of truly novel genes.

Results

Identifying Orphans. Our analysis requires studying the properties of human ORFs that lack cross-species counterparts, which we term "orphans." Such study requires carefully filtering the human gene catalogs, to identify genes with counterparts and to eliminate a wide range of artifacts that would interfere with analysis of the orphans. For this reason, we undertook a thorough reanalysis of the human gene catalogs.

We focused on the Ensembl catalog (version 35), which lists 22,218 protein-coding genes with a total of 239,250 exons. Our analysis considered only the 21,895 genes on the human genome reference sequence of chromosomes 1–22 and X. (We thus omitted the mitochondrial chromosome, chromosome Y, and "unplaced contigs," which involve special considerations; see below.)

We developed a computational protocol by which the putative genes are classified based on comparison with the human, mouse, and dog genomes (Fig. 1; see *Materials and Methods*). The mouse and dog genomes were used, because high-quality genomic sequence is available (7, 8), and the extent of sequence divergence is well suited for gene identification. The nucleotide substitution rate relative to human is ≈0.50 per base for mouse and ≈0.35 for dog, with insertion and deletion (indel) events occurring at a frequency that is ≈10-fold lower (8, 9). These rates are low enough to allow reliable sequence alignment but high enough to reveal the differential mutation patterns expected in coding and noncoding regions.

After the computational pipeline, we undertook visual inspection of ≈1,200 cases to detect instances misclassifications due to limitations of the algorithms or apparent errors in reported human gene annotations; this process revised the classification of 417 cases. We briefly summarize the results.

Class 0: Transposons, pseudogenes, and other artifacts. Some of the putative genes consist of transposable elements or processed pseudogenes that slipped through the process used to construct the Ensembl catalog. Using a more stringent filter, we identified 1,538 such cases. These were 487 cases consisting of transposon-derived sequence, 483 processed pseudogenes derived from a multiexon parent gene (recognizable because the introns had been eliminated by splicing), and 568 processed pseudogenes derived from a single-exon parent gene (recognizable because the pseudogene sequence almost precisely interrupts the aligned orthologous sequence of human with mouse or dog).

Class 1: Genes with cross-species orthologs. We next identified putative genes with a corresponding gene in the syntenic region of mouse or dog. We examined the orthologous DNA sequence in each species, checking whether an orthologous gene was already annotated in current gene catalogs for mouse or dog and, if not, whether we could identify an orthologous gene. Such cases are referred to as "simple orthology" (or 1:1 orthology). We then expanded the search to a surrounding region of 1 Mb in mouse and dog to allow for cases of local gene family expansion. Such cases are referred to as "complex orthology" (or "coorthology"). In both circumstances, the orthologous gene was required to have an ORF that aligns to a substantial portion (≥80%) of the human gene and have substantial peptide identity (≥50% for mouse, ≥60% for dog). Orthologous genes were identified for 18,752 of the putative human genes, with 16,210 involving simple orthology and 2,542 involving coorthology.

Class 2: Genes with cross-species paralogs. The pipeline then identified 155 cases of putative human genes that have a paralog within the human genome, that, in turn, has an ortholog in mouse or dog. These genes largely represent nonlocal duplications in the human lineage (three-quarters lie in segmental duplications) or possibly gene losses in the other lineages. Among these genes, close inspec-

GENETICS

tion revealed eight cases in which a small change to the human annotation allowed the identification of a clear human ortholog.

Class 3: Genes with human-only paralogs. The pipeline identified 68 cases of putative human genes that have one or more paralogs within the human genome, but with none of these paralogs having orthologs in mouse or dog. Close inspection eliminated 17 cases as additional retroposons or other artifacts (see *SI Appendix*). The remaining 51 cases appear to be valid genes, with 15 belonging to three known families of primate-specific genes (DUF1220, NPIP, and CDRT15 families) and the others occurring in smaller paralogous groups (two to eight members) that may also represent primate-specific families.

Class 4: Genes with Pfam domains. The pipeline identified 97 cases of putative genes with homology to a known protein domain in the Pfam collection (10). Close inspection eliminated 21 cases as additional retroposons or other artifacts (see *SI Appendix*) and 40 cases in which a small change to the human annotation allowed the identification of a clear human ortholog. The remaining 36 genes appear to be valid genes, with 10 containing known primate-specific domains and 26 containing domains common to many species.

Class 5: Orphans. A total of 1,285 putative genes remained after the above procedure. Close inspection identified 40 cases that were clear artifacts (long tandem repeats that happen to lack a stop codon) and 68 cases in which a cross-species ortholog could be assigned after a small change correction to the human gene annotation. The remaining 1,177 cases were declared to be orphans, because they lack orthology, paralogy, or homology to known genes and are not obvious artifacts. We note that the careful review of the genes was essential to obtaining a "clean" set of orphans for subsequent analysis.

Characterizing the Orphans. We characterized the properties of the orphans to see whether they resemble those seen for protein-coding genes or expected for randoms ORFs arising in noncoding transcripts.

ORF lengths. The orphans have a GC content of 55%, which is much higher than the average for the human genome (39%) and similar to that seen in protein-coding genes with cross-species counterparts (53%). The high-GC content reflects the orphans' tendency to occur in gene-rich regions.

We examined the ORF lengths of the orphans, relative to their GC-content. The orphans have relatively small ORFs (median = 393 bp), and the distribution of ORF lengths closely resembles the mathematical expectation for the longest ORF that would arise by chance in a transcript-derived form human genomic DNA with the observed GC-content (SI Fig. 4).

Conservation properties. We then focused on cross-species conservation properties. To assess the sensitivity of various measures, we examined a set of 5,985 "well studied" genes defined by the criterion that they are discussed in more than five published articles. For each well studied gene, we selected a matched random control sequence from the human genome, having a similar number of "exons" with similar lengths, a similar proportion of repeat sequence and a similar proportion of cross-species alignment, but not overlapping with any putative genes.

The well studied genes and matched random controls differ with respect to all conservation properties studied (SI Fig. 5 and SI Table 1). The nucleotide identity and Ka/Ks ratio clearly differ, but the distributions are wide and have substantial overlap. The indel density has a tighter distribution: 97.3% of well studied genes, but only 2.8% of random controls, have an indel density of <10 per kb. The sharpest distinctions, however, were found for two measures that reflect the distinctive evolution of protein-coding genes: the reading frame conservation (RFC) score and the codon substitution frequency (CSF) score.

Reading frame conservation. The RFC score reflects the percentage of nucleotides (ranging from 0% to 100%) whose reading frame is conserved across species (SI Fig. 6). The RFC score is determined

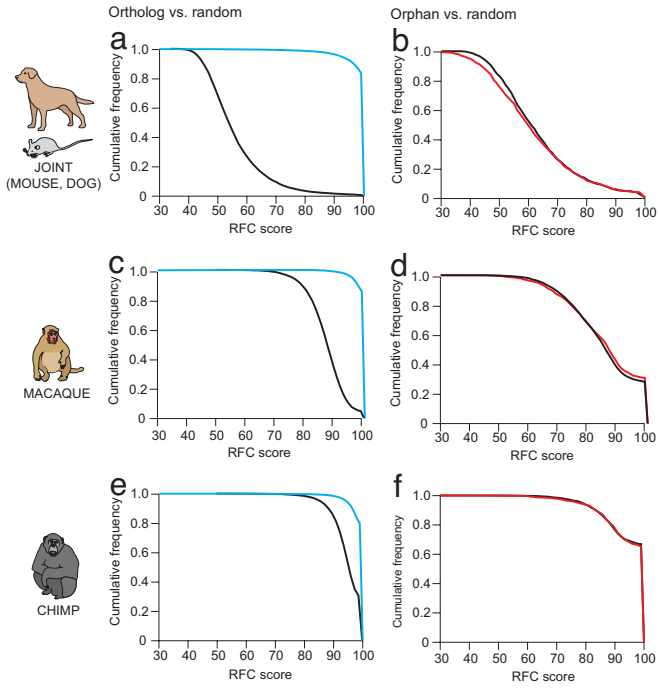

Fig. 2. Cumulative distributions of RFC score. (*Left*) Human genes with cross-species orthologs (blue) versus matched random controls (black). (*Right*) Human orphans (red) versus matched random controls (black). RFC scores are calculated relative to mouse and dog together (*Top*), macaque (*Middle*) and chimpanzee (*Bottom*). In all cases, the orthologs are strikingly different from their matched random controls, whereas the orphans are essentially indistinguishable from their matched random controls.

by aligning the human sequence to its cross-species ortholog and calculating the maximum percentage of nucleotides with conserved reading frame, across the three possible reading frames for the ortholog. The results are averaged across sliding windows of 100 bases to limit propagation of local effects due to errors in sequence alignment and gene boundary annotation. We calculated separate RFC scores relative to both the mouse and dog genomes and focused on a joint RFC score, defined as the larger of two scores. The RFC score was originally described in our work on yeast, but has been adapted to accommodate the frequent presence of introns in human sequence (see *SI Appendix*).

The RFC score shows virtually no overlap between the well studied genes and the random controls (SI Fig. 5). Only 1% of the random controls exceed the threshold of RFC >90, whereas 98.2% of the well studied genes exceed this threshold. The situation is similar for the full set of 18,752 genes with cross-species counterparts, with 97% exceeding the threshold (Fig. 2*a*). The RFC score is slightly lower for more rapidly evolving genes, but the RFC distribution for even the top 1% of rapidly evolving genes is sharply separated from the random controls (SI Fig. 5).

By contrast, the orphans show a completely different picture. They are essentially indistinguishable from matched random controls (Fig. 2*b*) and do not resemble even the most rapidly evolving subset of the 18,572 genes with cross-species counterparts. In short, the set of orphans shows no tendency whatsoever to conserve reading frame.

Codon substitution frequency. The CSF score provides a complementary test of for the evolutionary pattern of protein-coding genes. Whereas the RFC score is based on indels, the CSF score is based on the different patterns of nucleotide substitution seen in protein-coding vs. random DNA. Recently developed for comparative genomic analysis of *Drosophila* species (11), the method calculates a codon substitution frequency (CSF) score based on alignments

across many species. We applied the CSF approach to alignments of human to nine mammalian species, consisting of high-coverage sequence (≈7×) from mouse, dog, rat, cow, and opossum and low-coverage sequence (≈2×) from rabbit, armadillo, elephant, and tenrec.

The results again showed strong differentiation between genes with cross-species counterparts and orphans. Among 16,210 genes with simple orthology, 99.2% yielded CSF scores consistent with the expected evolution of protein-coding genes. By contrast, the 1,177 orphans include only two cases whose codon evolution pattern indicated a valid gene. Upon inspection, these two cases were clear errors in the human gene annotation; by translating the sequence in a different frame, a clear cross-species orthologs can be identified.

Orphans Do Not Represent Protein-Coding Genes. The results above are consistent with the orphans being simply random ORFs, rather than valid human protein-coding genes. However, consistency does not constitute proof. Rather, we must rigorously reject the alternative hypothesis.

Suppose the orphans represent valid human protein-coding genes that lack corresponding ORFs in mouse and dog. The orphans would fall into two classes: (*i*) some may predate the divergence from mouse and dog—that is, they are ancestral genes that were lost in both mouse and dog, and (*ii*) some may postdate the divergence—that is, they are novel genes that arose in the lineage leading to the human. How can we exclude these possibilities? Our solution was to study two primate relatives: macaque and chimpanzee. We consider the alternatives in turn.

1. Suppose that the orphans are ancestral mammalian genes that were lost in dog and mouse but are retained in the lineage leading to human. If so, they would still be present and functional in macaque and chimpanzee, except in the unlikely event that they also underwent independent loss events in both macaque and chimpanzee lineages.
2. Suppose that the orphans are novel genes that arose in the lineage leading to the human, after the divergence from dog and mouse [≈75 million years ago (Mya)]. Assuming that the generation of new genes is a steady process, the birthdates should be distributed across this period. If so, most of the birthdates will predate the divergence from macaque (≈30 Mya) and nearly all will predate the divergence from chimpanzee (≈6 Mya) (12).

Under either of the above scenarios, the vast majority of the orphans must correspond to functional protein-coding genes in macaque or chimpanzee.

We therefore tested whether the orphans show any evidence of protein-coding conservation relative to either macaque or chimpanzee, using the RFC score. Strikingly, the distribution of RFC scores for the orphans is essentially identical to that for the random controls (Fig. 2 *d* and *f*). The distribution for the orphans does not resemble that seen even for the top 1% of most rapidly evolving genes with cross-species counterparts (SI Figs. 7–9).

The set of orphans thus shows no evidence whatsoever of reading-frame conservation even in our closest primate relatives. (It is of course possible that the orphans include a few valid protein-coding genes, but the proportion must be small enough that it has no discernable effect on the overall RFC distribution.) We conclude that the vast majority of orphans do not correspond to functional protein-coding genes in macaque and chimpanzee, and thus are neither ancestral nor newly arising genes.

If the orphans represent valid human protein-coding genes, we would have to conclude that the vast majority of the orphans were born after the divergence from chimpanzee. Such a model would require a prodigious rate of gene birth in mammalian lineages and a ferocious rate of gene death erasing the huge number of genes born before the divergence from chimpanzee. We reject such a model as wholly implausible. We thus conclude that the vast majority of orphans are simply randomly occurring ORFs that do not represent protein-coding genes.

Finally, we note that the careful filtering of the human gene catalog above was essential to the analysis above, because it eliminated pseudogenes and artifacts that would have prevented accurate analysis of the properties of the orphans.

Experimental Evidence of Encoded Proteins. As an independent check on our conclusion, we reviewed the scientific literature for published articles mentioning the orphans to determine whether there was experimental evidence for encoded proteins. Whereas the vast majority of the well studied genes have been directly shown to encode a protein, we found articles reporting experimental evidence of an encoded protein *in vivo* for only 12 of 1,177 orphans, and some of these reports are equivocal (SI Table 2). The experimental evidence is thus consistent with our conclusion that the vast majority of nonconserved ORFs are not protein-coding. In the handful of cases where experimental evidence exists or is found in the future, the genes can be restored to the catalog on a case-by-case basis.

Revising the Human Gene Catalogs. With strong evidence that the vast majority of orphans are not protein-coding genes, it is possible to revise the human gene catalogs in a principled manner.

Ensembl catalog. Our analysis of the Ensembl (v35) catalog indicates that it contains 19,108 valid protein-coding genes on chromosomes 1–22 and X within the current genome assembly. The remaining 15% of the entries are eliminated as retroposons, artifacts or orphans. Together with the mitochrondrial chromosome [well known to contain 13 protein-coding genes (13)] and chromosome Y [for which careful analysis indicates 78 protein-coding genes (14)], the total reaches 19,199.

We extended the analysis to the Ensembl (v38) catalog, in which 2,212 putative genes were added and many previous entries were revised or deleted. Our computational pipeline found 598 additional valid protein-coding genes based on cross-species counterparts, 1,135 retroposons, and 479 orphans. The RFC curves for the orphans again closely matched the expectation for random DNA.

Other catalogs. We applied the same approach to the Vega (v34) and RefSeq (March 2007) catalog. Both catalogs contain a substantial proportion of entries that appear not to be valid protein-coding genes (16% and 10%, respectively), based on the lack of a cross-species counterpart (see SI Fig. 10 and *SI Appendix*). If we restrict the RefSeq entries to those with the highest confidence (with the caveat that this set contains many fewer genes), only 1% appear invalid. Together, these two catalogs add an additional 673 protein-coding genes.

Combined analysis. Combining the analysis of the three major gene catalogs, we find that only 20,470 of the 24,551 entries appear to be valid protein-coding genes.

Limitations on the Analysis. Our analysis of the current gene catalogs has certain limitations that should be noted.

First, we eliminated all pseudogenes and orphans. We found six reported cases in which a processed pseudogene or transposon underwent exaptation to produce a functional gene (SI Tables 1 and 3) and 12 reported cases of orphans with experimental evidence for an encoded protein. These 18 cases can be readily restored to the catalog (raising the count to 20,488). There are additional cases of potentially functional retroposons that are not present in the current gene catalogs (15). If any are found to produce protein, they should also be included.

Second, we have not considered the 197 putative genes that lie in the "unmapped contigs." These regions are sequences that were omitted from the finished assembly of the human genome. They largely consist of segmental duplications, and most of the genes are highly similar to others in the assembly. Many of the sequence may

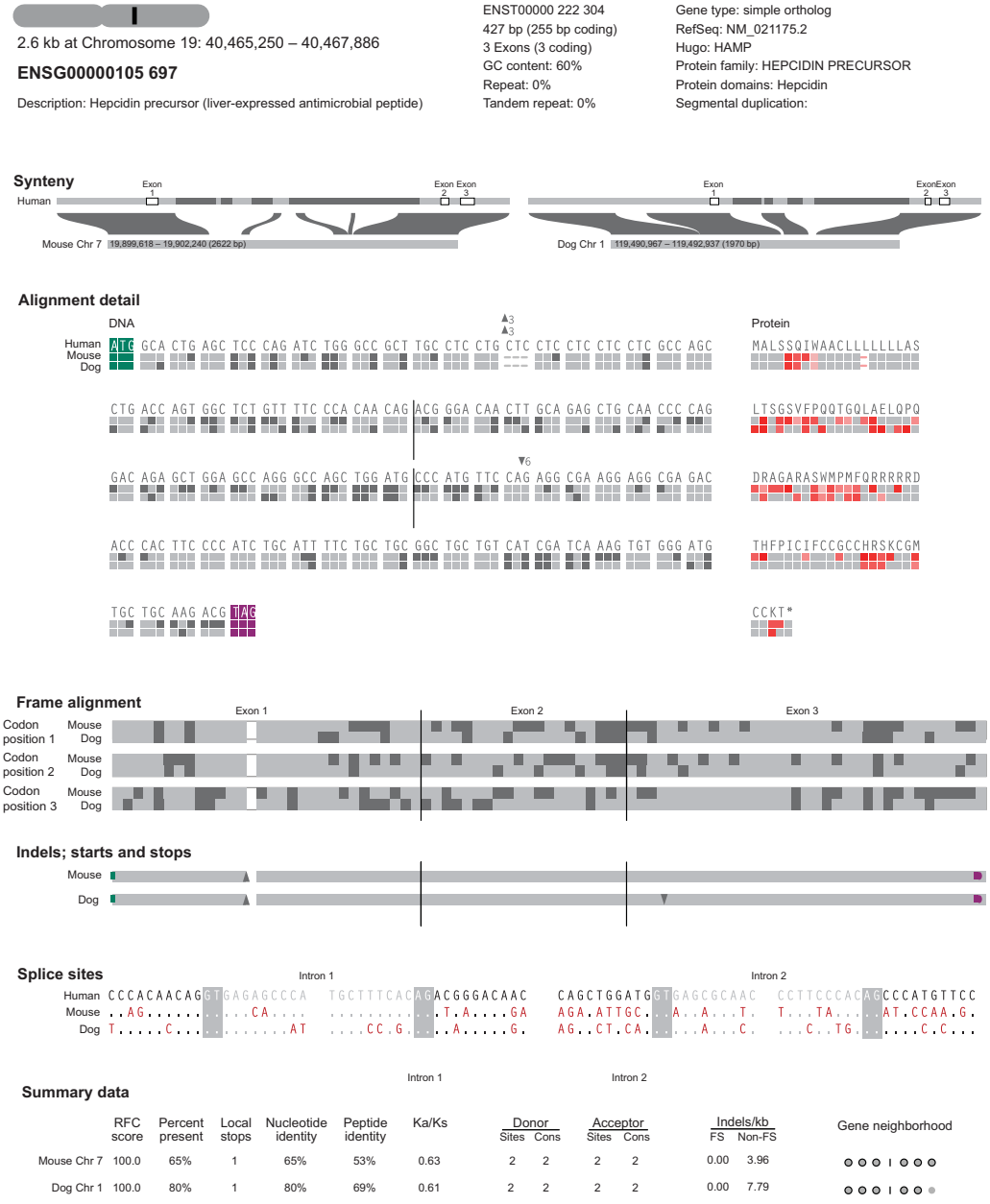

Fig. 3. An example gene report card for a small gene, HAMP, on chromosome 19. Report cards for all 22,218 putative genes in Ensembl v35 are available at www.broad.mit.edu/mammals/alpheus. The report cards provide a visual framework for studying cross-species conservation and for spotting possible problems in the human gene annotation. Information at the top shows chromosomal location; alternative identifiers; and summary information, such as length, number of exons, and repeat content. Various panels below provide graphical views of the alignment of the human gene to the mouse and dog genomes. ''Synteny'' shows the large-scale alignment of genomic sequence, indicating both aligned and unaligned segments. The human sequence is annotated with the exons in white and repetitive sequence in dark gray. ''Alignment detail'' shows the complete DNA sequence alignment and protein alignment. In the DNA alignment, the human sequence is given at the top, bases in the other species are marked as matching (light gray) or nonmatching (dark gray), exon boundaries are marked by vertical lines, indels are marked by small triangles above the sequence (vertex down for insertions, vertex up for deletions, number indicating length in bases), the annotated start codon is in green, and the annotated stop codon is in purple. In the protein alignment, the human amino acid sequence is given at the top, and the sequences in the other species are marked as matching (light gray), similar (pink), or nonmatching (red). ''Frame alignment'' shows the distribution of nucleotide mismatches found in each codon position, with excess mutations expected in the third position. Matching are shown in light gray, and mismatches are shown in dark gray. ''Indels, starts and stops'' provides an overview of key events. Indels are indicated by triangles (vertex down for insertions, vertex up for deletions) and marked as frameshifting (red) or frame-preserving (gray). Start codons are marked in green and stop codons in purple. ''Splice sites'' shows sequence conservation around splice sites, with two-base donor and acceptor sites highlighted in gray and mismatching bases indicated in red. ''Summary data'' lists various conservation statistics relative to mouse and dog, including RFC score, nucleotide identity, number of conserved splice sites, frameshifting and nonframeshifting indel density/kb, and gene neighborhood. The gene neighborhood shows a dot for the three upstream and downstream genes, which is colored gray if synteny is preserved and red otherwise.

represent alternative alleles or misassemblies of the genome. However, regions of segmental duplication are known to be nurseries of evolutionary innovation (16) and may contain some valid genes. They deserve focused attention.

Third and most importantly, the nonconserved ORFs studied here were typically included in current gene catalogs because they have the potential to encode at least 100 amino acids. We thus do not know whether our conclusions would apply to much shorter ORFs. In principle, there exist many additional protein-coding genes that encode short proteins, such as peptide hormones, which are usually translated from much larger precursors and may evolve rapidly. It should be possible to investigate the properties of smaller ORFs by using additional mammalian species beyond mouse and dog.

Improving Gene Annotations. In the course of our work, we generated detailed graphical "report cards" for each of the 22,218 putative genes in Ensembl (v35). The report cards show the gene structure, sequence alignments, measures of evolutionary conservation, and our final classification (Fig. 3).

The report cards are valuable for studying gene evolution and for refining gene annotation. By examining local anomalies by cross-species comparison, we have identified 23 clear errors in gene annotation (including cases in which changing the reading frame or coding strand reveals unambiguous cross-species orthologs) and

332 cases in which cross-species conservation suggests altering the start or stop codon, eliminating an internal exon, or moving a splice site. Of these latter cases, most are likely to be errors in the human gene annotation, although some may represent true cross-species differences. The report cards, together with search tools and summary tables, are available at www.broad.mit.edu/mammals/alpheus.

Discussion

The analysis here addresses an important challenge in genomics—determining whether an ORF truly encodes a protein. We show that the vast majority of ORFs without cross-species counterparts are simply random occurrences. The exceptions appear to represent a sufficiently small fraction that the best course is would be consider such ORFs as noncoding in the absence of direct experimental evidence.

We propose that it is time to undertake a thorough revision of the human gene catalogs by applying this principle to filter the entries. Specifically, we propose that nonconserved ORFs should be included in the human gene catalog if there is clear experimental evidence of an encoded protein. We report here an initial attempt to apply this principle, resulting in a catalog with 20,488 genes.

Our focus has been on excluding putative genes from the human catalogs. We have not explored whether there are additional protein-coding genes that have not yet been included, although it is clear that cross-species analysis can be helpful in identifying such genes. Preliminary analysis from our own group and others suggests that there may be a few hundred additional protein-coding genes to be found but that the final total is likely to remain under ≈21,000. The largest open question concerns very short peptides, which may still be seriously underestimated.

One important biological implication of our results is that truly novel protein-coding genes (encoding at least 100 amino acids) arise only rarely in mammalian lineages. With the current gene catalogs, there are only 168 "human-specific" genes (<1% of the total; only 11 are manually reviewed entries in RefSeq; see SI Table 4). These genes lack clear orthologs or paralogs in mouse and dog, but are recognizable because they belong to small paralogous families within the human genome (2 to 9 members) or contain Pfam domains homologous to other proteins. These paralogous families shows a range of nucleotide identities, consistent with their having arisen over the course of ≈75 million years since the divergence from the mouse lineage. In fact, many of these 168 genes are not entirely novel inventions: One-third show strong similarity to mouse or dog genes across at least 50% of their length; although this falls short of our threshold for declaring orthologs or paralogs (80%), it is nonetheless substantial. Among the orphans, there are only 12 cases with reported experimental evidence of an encoded protein. These cases, which comprise ≈0.06% of the gene catalog, have similar RFC and nucleotide identity scores to neutral sequence and have no similarity with any mouse or dog genes, suggesting these are truly novel inventions. We conclude that mammals thus share largely the same repertoire of protein-coding genes, modified primarily by gene family expansions and contractions.

Finally, the creation of more rigorous catalogs of protein-coding genes for human, mouse, and dog will also aid in the creation of catalogs of noncoding transcripts. This should help propel understanding of these fascinating and potentially important RNAs.

Materials and Methods

All annotations were based on the NCBI35 (hg17) assembly and all genome alignments were taken from the pairwise BLASTZ alignment to mouse assembly NCBI36 (mm4) and dog Broad, Version 1.0 (canFam1; available from http://genome.ucsc.edu). We identified retroposons, using the Ensembl annotation (www.ensembl.org). We then eliminated pseudogenes by identifying transcripts with either retained introns or through interrupted synteny at the boundaries of the transcript. The set of well studied genes were found by using those transcripts whose RefSeq entry contained references to more than five articles. Orthologous genes were identified by using synteny (across >80% of the gene) and peptide identity (>50% for mouse and >60% for dog). The combined RFC score was the highest independent score (taking into account the length of the transcript) for alignments to both mouse and dog. For more details, see *SI Appendix*.

We thank colleagues at the University of California, Santa Cruz, genome browser and the Ensembl genome browser for providing data (BLASTZ alignments, synteny nets, genes, and annotations); L. Gaffney for assistance in preparing the manuscript and figures; S. Fryc and N. Anderson for resequencing data; and a large collection of colleagues around the world for many helpful discussions over the past 3 years that have helped shape and improve this work. This work was supported by the National Institutes of Health National Human Genome Research Institute.

1. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, *et al*. (2005) *Science* 308:1149–1154.
2. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*. (2005) *Science* 309:1559–1563.
3. ENCODE Project Consortium (2007) *Nature* 447:799–816.
4. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*. (2007) *Nucleic Acids Res* 35:D610–D617.
5. Pruitt KD, Tatusova T, Maglott DR (2007) *Nucleic Acids Res* 35:D61–D65.
6. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, Meidl P, Searle SM, Stalker J, Storey R, Trevanion S, *et al*. (2005) *Nucleic Acids Res* 33:D459–D465.
7. Goodstadt L, Ponting CP (2006) *PLoS Comput Biol* 2:e133:1134–1150.
8. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, III, Zody MC, *et al*. (2005) *Nature* 438:803–819.
9. Mouse Genome Sequencing Consortium (2002) *Nature* 420:520–562.
10. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, *et al*. (2006) *Nucleic Acids Res* 34:D247–D251.
11. Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St. Pierre SE, *et al*. (2007) *Genome Res*, 10.1101/gr.6679507.
12. Pilbeam D, Young N (2004) *C R Palevol* 3:305–321.
13. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, *et al*. (1981) *Nature* 290:457–465.
14. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, *et al*. (2003) *Nature* 423:825–837.
15. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) *Proc Natl Acad Sci USA* 103:3220–3225.
16. Eichler EE (2001) *Trends Genet* 17:661–669.