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.

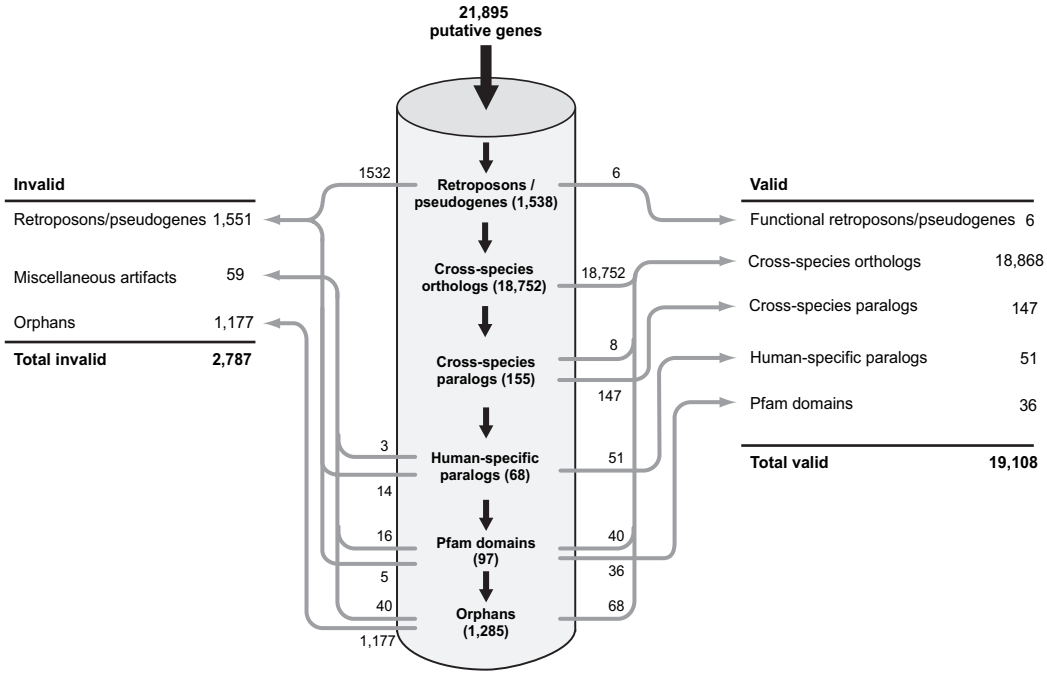


Fig. 1. Flowchart of the analysis. The central pipeline illustrates the computational analysis of 21,895 putative genes in the Ensembl catalog (v35). We then performed manual inspection of 1,178 cases to obtain the tables of likely valid and invalid genes. See text for details.

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-

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

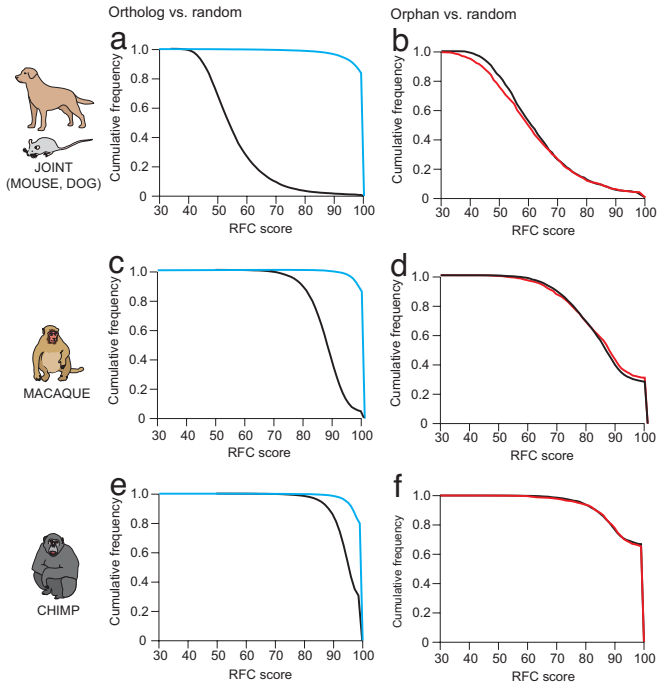


Fig. 2. Cumulative distributions of RFC score. (*Left*) Human genes with cross-species orthologs (blue) versus matched random controls (black). (*Right*) Human orphans (red) versus matched random controls (black). RFC scores are calculated relative to mouse and dog together (*Top*), macaque (*Middle*) 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

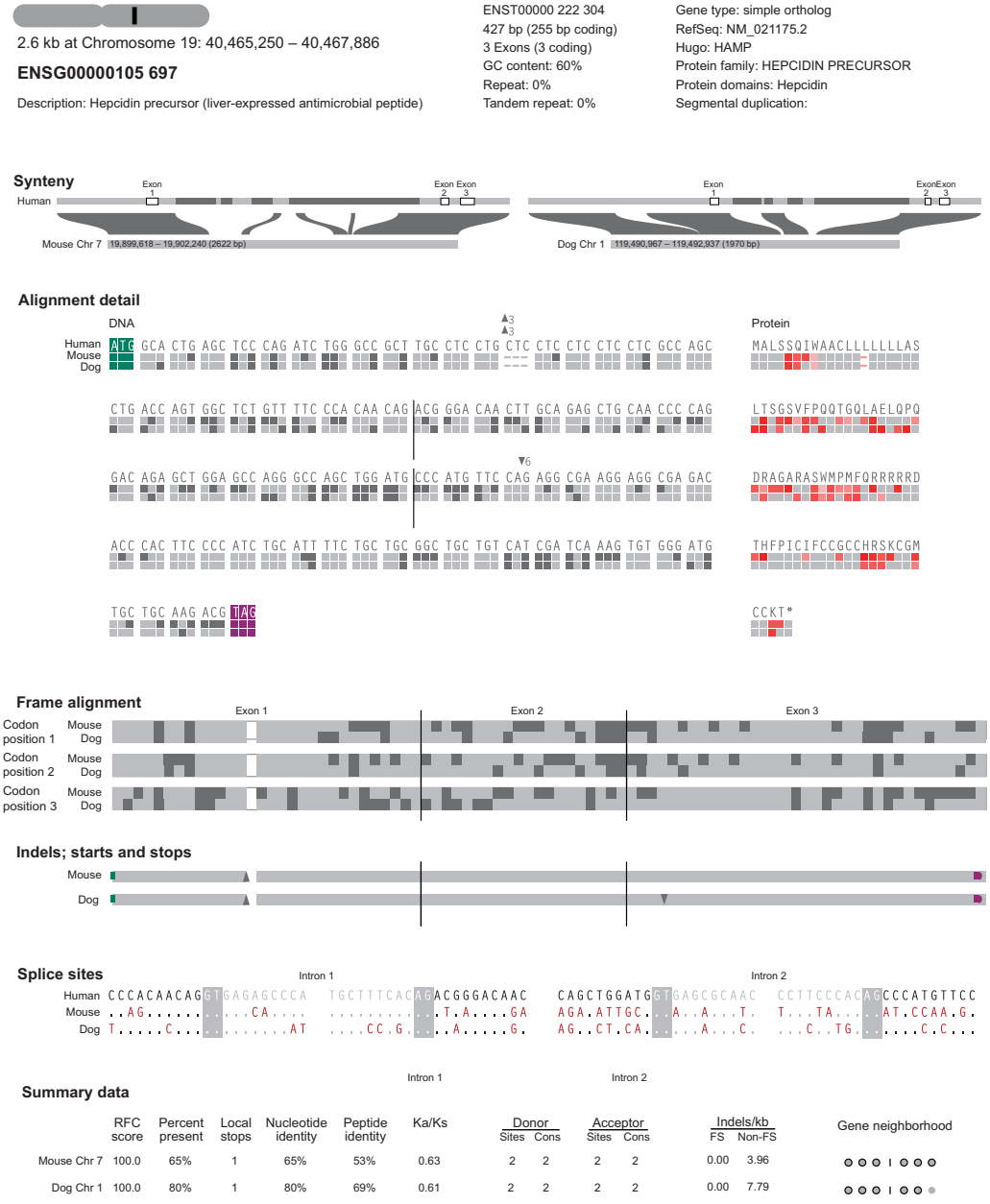


Fig. 3. An example gene report card for a small gene, HAMP, on chromosome 19. Report cards for all 22,218 putative genes in Ensembl v35 are available at www.broad.mit.edu/mammals/alpheus. The report cards provide a visual framework for studying cross-species conservation and for spotting possible problems 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.

I. Methods

M1. Retroposons. We identified retroposons in genes by searching for overlaps with the repeat annotation in the Ensembl human genome sequence and discarding those repeats labeled as low complexity or tandem repeats that overlap with >80% of the gene. The Ensembl genome annotation identifies the repeat content in the genome by using the programs RepeatMasker[1], dust[2] and trf[3] (See S2 for more details).

M2. Pseudogenes. Processed pseudogenes originate from a real ('parent') gene. Therefore, our process for identifying pseudogenes first attempted to find a parent for each input gene. We first compared all human proteins against each other using blastp. We then filtered the raw blastp[4] output to eliminate spurious low complexity matches using the MSPcrunch[5] program. Given an input gene, we identified a 'possible parent gene' as the highest scoring gene with an alignment spanning at least 80% of both itself and the input gene. We then looked for evidence of RNA processing, by checking for the absence of introns present in the parent gene.

We first considered input genes with at least three exons and parent genes that had at least 2 additional exons. We labeled the input gene as a pseudogene if more than two aligned exon boundaries in the parent gene did not correspond to exon boundaries in the input gene.

The above method is suitable for multi-exon parent genes, but does not identify pseudogenes derived from single-exon genes. For this purpose, we took a different approach based on syntenic alignments to the surrounding mouse and dog genomic sequence. If the processed pseudogene is more recent than the human-mouse divergence, the genome alignment to mouse or dog should be interrupted almost precisely at the boundaries of the pseudogene and the pseudogene itself will align elsewhere in the other species' genome. We used the UCSC net alignments to identify the highest scoring region in mouse or dog for each human gene. The net alignments are generated by first aligning one genome to another using blastz[6]. The alignments are then placed on the human genome with the highest scoring one first and the subsequent ones trimmed to fit into regions that aren't already covered. We identified the net alignment for each human gene and found how far up and downstream the net alignment extended. If the alignment extended less than 2000bp and the gene itself had at most 2 exons, we flagged the input gene as a pseudogene. (See S3 for refinements)

M3. Well-studied genes. The RefSeq gene catalog is a set of ~18,000 human genes which have undergone manual inspection. Each entry has one (or more) reference cDNAs and also a list of publications that reference this gene. We took all the Ensembl genes that also had a RefSeq identifier and counted up how many publications were listed in the RefSeq entry. We defined a gene as well-studied if it was associated with > 5 publications. We found 5,985 well-studied genes (see S7 for more details and for description of the random controls).

M4. Orthologs. For each human gene we used 3 methods to assign an ortholog in mouse and dog. The first method used a combination of synteny and protein similarity. We compared all human, mouse and dog proteins to each other using blastp and the results were filtered for low complexity matches using MSPcrunch. A second filter only considered similarities if they lay in syntenic positions (Mike Kamal pers. comm.) and a final filter only retained similarities if the two genes were aligned across 80% of both their lengths. We labeled a gene a 1:1 (simple) ortholog if a human gene was only similar to a single mouse and/or dog gene and a many:many ortholog if a gene from any of the 3 species was similar to multiple genes in another.

For the remaining genes without an ortholog we used whole genome alignments (UCSC net alignments) to identify the orthologous region in mouse and dog. If the mouse or dog regions had been annotated with an Ensembl gene we assigned that gene as an ortholog.

Finally if a human gene still had no ortholog in mouse or dog we attempted to identify an orthology – that is, to define an annotation in the other species. We again identified the orthologous region using whole genome alignments and concatenated the sequences corresponding exactly to the human exons to create a putative coding sequence. After finding the longest ORF in the putative coding sequence we aligned the peptide to human using both clustalw[7] and genewise[8] . If either peptide alignment had at least 50% to mouse or 60% identity to dog and at least 80% of the human peptide was contained in the alignment we assigned an ortholog.

To determine the type of ortholog we combined the last two categories with the first. We labeled genes many:many (complex orthology) if a similar gene was present within 1Mb. All remaining genes were labeled 1:1 (simple orthology).

M5. Paralogs. We compared all human genes against each other to find probable paralogs. We first used blastp to compare all human peptides against each other. After filtering for low complexity matches using MSPcrunch we assigned a paralogous relationship if at least 80% of one gene was similar to at least 80% of another. We identified the most similar gene as the one that had the longest alignment. In the event that more than one gene with the same length alignment we took the gene with the highest percent identity as the best hit.

M6. Pfam domains. Ensembl labels Interpro protein domains in all its annotations. We used only the pfam annotations (as they have fewer false positive matches). For Vega and Refseq genes we found pfam[9] domains using a Hidden Markov Model (pfam_scan.pl which uses hmmpfam from the HMMER[10] package) to search Pfam-A v21.

M7. Combined RFC score. We initially calculated the RFC score for each gene independently for mouse and dog. If the gene contained multiple transcripts the highest scoring transcript was chosen to represent that gene. We then created a combined RFC score which took into account the number of orthologous bases in the other species (a shorter alignment is more likely to give a high RFC score). Specifically, we used the highest RFC score of both species as the combined score unless the following conditions were met: if the orthologous alignment for one species aligned to less than 80% of the

human gene and the alignment for the other aligned to more than 80% the score from the species with the longer alignment was used.

M8. Ka/Ks and percent identity values. The program codeml from the paml[11] package was used to calculate Ka/Ks values. The percent identity was calculated as the percentage of aligned bases that were identical.

M9. Codon substitution frequency (CSF) Score. We assigned a score to each observed codon substitution between human and an aligned informant sequence. This was equal to the log-likelihood-ratio of observing that codon substitution in coding sequence versus noncoding sequence, conditioned on observing a substitution of the human codon. We computed these log-likelihood-ratios from codon distance matrices estimated by counting codon substitutions in alignments of annotated genes and noncoding regions, similar to the BLOSUM amino acid distance matrices estimated from protein alignments (Henikoff and Henikoff 1992). We then summed the scores of all codon substitutions in a candidate region to obtain a score for that region, with the condition that no score was assigned to gapped or perfectly conserved codons. If multiple informant sequences were present, we used the median of the scores of all codon substitutions in each codon column as the score of that column, and the score of each column was summed to score the region. The CSF approach is described more fully in Lin *et al* (submitted).

II. Supplementary Notes

S1 - Genome data sources

S1.1 - Genome assemblies. Different versions of the human, mouse and dog reference genome sequence exist, reflecting additions of data or changes in the genome assembly. The versions used in this work are as follows:

Species	UCSC Version	NCBI version
Human	hg17	35
Mouse	mm4	36
Dog	canFam1	Broad version 1.0
Chimp	panTro1	CGSC Build 1 Version 1
Macaque	rheMac2	Baylor College of Medicine HGSC Mmul_0.1

A document relating the UCSC versions to the NCBI versions can be found here http://genome.ucsc.edu/FAQ/FAQreleases#release1.

S1.2 - Gene annotations.

1.2.1 Human annotations For the initial detailed analysis human gene annotations were taken from Ensembl v35 (November 2005 available from

ftp://ftp.ensembl.org/pub/release-35/homo_sapiens_35_35h). Extra genes were later included from a more recent catalog (Ensembl 38 which is available from ftp://ftp.ensembl.org/pub/release-38/homo_sapiens_38_36). The Ensembl gene building pipeline is described in more detail in Curwen *et al* but we outline the process briefly below.

Ensembl creates annotations in a two-step process. First, all known genes are aligned to the genome. Their protein sequence is aligned to the genome to find the coding exons and then the full cDNA sequence is aligned to the genome to identify the UTR regions. When aligning a sequence to the genome only hits that have > 95% identity across 90% of their length are considered. Of these the highest scoring position is taken along with all other positions within 2% of the top score.

The second step is to use similarity to known genes to identify paralogs. All vertebrate proteins are aligned to the genome and, after masking out any overlap with the known gene positions, all alignments that cover ≥80% of the protein length are annotated as genes. The final gene set is filtered for transposable elements (from a locally maintained list of known transposable element accession numbers).

1.2.2 Mouse annotations Mouse gene annotations were used for steps 1 and 2 of the orthology finding process. We used the Ensembl gene annotation (Ensembl version 31, mouse annotation version 33g). The genome sequence, annotation coordinates and sequences are available from ftp://ftp.ensembl.org/pub/release-31/mouse-31.33g.

1.2.3 Dog annotations Dog gene annotations were used for steps 1 and 2 of the ortholog finding process. We used the Ensembl gene annotation (Ensembl version 31, dog annotation version 1c). The genome sequence, annotation coordinates and sequences are available from ftp://ftp.ensembl.org/pub/release-31/dog-31.1.c.

It may seem odd that we have used different Ensembl version numbers for the human (35) and the mouse (31) and dog (31) annotations. This was necessary as we were constrained by the assembly versions used in the UCSC genome alignments. We consider this to have a negligible effect on the analysis. The differences in the assemblies are too small to have an appreciable effect on the gene content and the annotation process was the same between assembly versions.

1.2.4 Refseq Genes After the detailed analysis of the Ensembl35 gene set we expanded our analysis to include novel genes from other catalogs. Refseq is a widely-used catalog which, although only containing 18,000 genes, aims to provide high quality annotation through manual curation.

Refseq entries are available in different flavors, which differ in the level of experimental support and amount of manual curation. We only considered entries with at least experimental (cDNA) support and discarded all predictions.

In order to be as up to date as) possible we didn't use a release version of Refseq but used the nightly snapshot maintained at UCSC (taken March 27th 2007 from genome.ucsc.edu) of all available RefSeq entries including incremental Genbank entries. This snapshot data is available from http://www.broad.mit.edu/~mclamp/alpheus/data/RefSeq.032707.txt. As there can be

more than one RefSeq entry per gene we took the entry with the longest CDS as a representative.

S1.2.5 Vega genes As well as the RefSeq genes we also included genes from the Vega catalog. Although only 90% of the genome is covered this gene set is important to include for two reasons. Firstly all genes have been manually curated and so represent the 'best' gene structure as judged by human eyes. Secondly, unlike the Ensembl gene set, gene structures have been included which are only supported by EST evidence and so will give a more comprehensive picture of the human gene content. The Vega annotation used is available from ftp://ftp.sanger.ac.uk/pub/vega/human/pep/Homo_sapiens.VEGA.jan.pep.known.fa.gz and ftp://ftp.sanger.ac.uk/pub/vega/human/pep/Homo_sapiens.VEGA.jan.pep.novel.fa.gz

S1.3 - Inter-species genome alignments. The comparative analysis called for the alignment of orthologous regions between human and mouse and human and dog. Pairwise alignments were used (using a local alignment method) that, for each stretch of human genome, identified the most similar region of the mouse (or dog) genome.

These alignments were downloaded from UCSC from the following locations:

Mouse	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsMm4
Dog	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsCanFam1
Chimp	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsPanTro1
Macaque	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsRheMac2

S2 Identification of Special regions

We put a certain number of the 22,218 Ensembl35 genes to one side as they were either not part of the reference genome, were redundant copies, or were in regions that we did not have orthologous alignments for. We thus did not analyse genes on chrY, genes not assigned to a chromosome (identified by having an assigned chromosome identifier containing _NT_), mitochondrial genes (chromosome identifier MT), and genes on a known haplotype (chromosome identifier containing DR). We retained all genes on chromosomes 1-22 and X.

S3 Identification of Transposons

S3.1 - Identification of transposons within the gene set. Part of the Ensembl annotation method is based on finding paralogs to known genes as well as identifying the genomic location of sequenced cDNAs. Some transposons are known to be transcribed and, as many copies are found in the genome, it is possible that non-transcribed transposons will be falsely identified as genes through similarity. The positions of all transposable repeats were taken from the Ensembl 35 annotation database (mysql tables are available from ftp.ensembl.org/pub/release-35/data/mysql/homo_sapiens_core_35_35h). These repeats are identified by Smith-Waterman alignment to the consensus sequence of families of

known repeats (RepeatMasker), low complexity regions (dust) and tandem repeats (trf). We only filtered based on non low-complexity repeats identified by RepeatMasker.

We tagged a gene as being a transposon if ≥80% of a gene's bases were annotated as such. Of the 496 genes identified this way we found 15 cases (listed below) that either had papers associated with them or HGNC names. Further inspection revealed 1 case that was spurious (no cDNA and the HGNC name had been misassigned), 2 cases were chrY satellite repeats, 10 cases were transcribed but had no peptide evidence and 2 cases were transcribed and did have peptide evidence.

Genes identified as transposons with papers or HGNC names

Ensembl gene identifier	Description	Classification
ENSG00000197604	27 papers and has a refseq entry. Functional expressed retrovirus.	Functional
ENSG00000176171	18 papers and named BNIP3. Western blot shows evidence of peptide which is located in the nucleus.	Functional
ENSG00000196546	6 papers. This is expressed but there is no cdna	Expressed
ENSG00000185654	2 papers. This is transcribed.	Expressed
ENSG00000183058	2 papers. This is transcribed	Expressed
ENSG00000171567	2 papers and is named TIGD1 Tigger transposable element.	Expressed
ENSG00000166718	2 papers but these are high-throughput sequencing papers and have no discussion of this gene.	Expressed
ENSG00000129586	1 paper. Transcribed Buster 1 transposase	Expressed
ENSG00000164608	1 paper. Transcribed Buster 3	Expressed
ENSG00000198048	1 paper. Tigger4 transcribed repeat.	Expressed
ENSG00000198762	1 paper which is a high-throughput sequencing paper and has no discussion of this gene.	Expressed
ENSG00000196134	Named TTTY12. Chr Y gene – dealt with in the chrY section.	Chr Y
ENSG00000196941	Named TTTY10. Chr Y gene – dealt with in the chrY	Chr Y

	section.	
ENSG00000196413	This doesn't have any evidence of expression.	Not expressed

S3.2 Refinement of transposon list. A total of 12 genes that were identified as transposons had publications associated with them (as listed in the associated refseq entries see S6). Of these, 2 had any evidence of function. These were ENSG00000197604 and ENSG00000176171 and were removed from the transposon list and added to the valid gene list.

Close inspection of the other gene categories revealed that 5 genes with paralogs were actually transposons. These genes were:

ENSG00000183683
ENSG00000186967
ENSG00000188426
ENSG00000197480
ENSG00000186750

S4 - Processed pseudogene identification

Similar to the possibility of transposable repeats being identified as genes processed pseudogenes are also prone to being mistakenly identified as genes. We devised two protocols to identify when this had occurred – one for pseudogenes arising from multi-exon genes and one for those from single-exon genes.

S4.1 - Pseudogenes arising from multi-exon genes. When a processed pseudogene is reinserted into the genome it has, by definition, had its introns spliced out. This lack of introns is most clearly shown when the pseudogene is aligned back to the parent gene's genomic region. However, if the positions of the introns are marked up on the parent gene's cDNA we can also identify if a gene has been processed. Most processed pseudogenes will be one exon (or sometimes two if a repeat has been inserted post-pseudogenization) and none of the intron positions will line up with the parent cDNA. Using this principle the following rule was devised to identify such cases: a pseudogene must

a) Align across 80% of its length to another gene in the genome.
b) Have ≤ 2 exons
c) Align to the parent cDNA across 2 or more intron boundaries.

S4.2 Pseudogenes arising from single-exon genes. To identify pseudogenes derived from single-exon genes, we used the syntenic alignments to the surrounding mouse and dog genomic sequence. If a processed pseudogene is more recent than the human-mouse divergence, the genome alignment to mouse or dog should be interrupted almost precisely at the boundaries of the pseudogene and the pseudogene itself will align elsewhere in the other species' genome. We used the UCSC net alignments to identify

the highest scoring region in mouse or dog for each human gene. The net alignments are generated by first aligning one genome to another using blastz. The alignments are then placed on the human genome with the highest scoring one first and the subsequent ones trimmed to fit into regions that aren't already covered. We identified the net alignment for each human gene and found how far up and downstream the net alignment extended. If the alignment extended less than 2000bp and the gene itself had at most 2 exons, we flagged the input gene as a pseudogene.

In some regions where duplications had happened it was possible that this method could misclassify a real, duplicated gene as a pseudogene. We thus applied a final filter that a gene was only tagged as a pseudogene if it was not similar (blastp aligns over 80% of its length) to a gene up to 1Mb away.

S4.3 Refinement of the pseudogene list. We inspected all pseudogenes that had publications associated with them (according to the publication list in the refseq entry. See S6). This revealed 4 genes that had evidence of function and were counted as valid genes. These genes were:

ENSG00000151846	PABPC3
ENSG00000178363	CALML3
ENSG00000130383	FUT5
ENSG00000162595	NM_004675

From the other gene categories, 14 additional genes were found to be pseudogenes that had either been missed due to their syntenic region begin slightly outside our threshold of 2kb larger than the coding region or that that they had many frameshifts/small introns in their gene structure indicating a non-processed pseudogene.

These 14 genes were:

ENSG00000151846
ENSG00000178363
ENSG00000130383
ENSG00000162595
ENSG00000112790
ENSG00000188327
ENSG00000187826
ENSG00000186678
ENSG00000185682
ENSG00000188841
ENSG00000183279
ENSG00000183247
ENSG00000169362
ENSG00000185671
ENSG00000177219
ENSG00000188075
ENSG00000188459
ENSG00000183935

S5. RFC score definition

An RFC score is a property of a gapped pairwise alignment that shows how biased the indel sizes are towards multiples of 3. For a given alignment the bases in the first sequence are numbered as to their codon position (1,2,3). The bases in the second sequence are given three alternative sets of numberings according to which starting codon position (1,2,3) is used. For each starting codon position in the second sequence, two numbers are found: the number of times the codon position is the same in both species and the total number of bases aligned. The RFC score for each frame is the fraction of the total aligned bases where the codon position is the same. The RFC score is then defined as the maximum score taken over all three frames. Because a single indel can throw off the rest of the alignment, the RFC score for a whole alignment is taken as the average over windows of 100bp taken in steps of 1 base.

The above description is identical to the method used in the yeast analysis (Kellis *et al*). A slight modification was made to deal with intron positions in the alignment - any gaps starting or ending at exon boundaries were ignored i.e. had no effect in changing the frame of the alignment.

The RFC analysis was performed for both the mouse and dog alignments. A final joint RFC score was then taken as the maximum of the scores if either (i) both species had at least 80% aligned coverage of the human gene or (ii) neither species had at least 80% coverage. If only one of the species had ≥80% coverage, the RFC score for that species was taken.

S6. Identification of random control transcripts

S6.1. Mouse and dog random control transcripts. For each Ensembl gene a random 'gene' alignment was generated (one each for mouse and dog) to match the coding length, exon size and length of the aligned orthologous sequence according to the following criteria:

1) A random region was chosen in the genome. One random alignment for each exon was chosen from the same region. Sticking to the same region for all exons ensures we get the correct distribution for GC content and variation in local neutral mutation rate.
2) No random alignment could overlap an exon or be >10% repeat. This is to limit us picking regions of genome that are under selection. The repeat content restriction is there to mimic the repeat content of real genes.
3) Once a random exon had been chosen the pairwise alignment to mouse (or dog) was taken from the UCSC pairwise net alignments (see S1.3).
4) A random exon alignment was only accepted if it passed the following test. The RFC score is extremely sensitive to the length of the alignment (a short alignment is less likely to include indels and so more likely to have a high RFC score). We thus only accepted a pairwise alignment if it contained the same proportion of mouse (or dog) bases to human bases with the caveat that we only counted bases as being 'missing' if they were in gaps of >20bp. If

the randomly picked alignment didn't pass this coverage test we carried on picking random human regions until the alignment did.

S6.2 - Chimp and macaque random control transcripts. This was done in the same way as for mouse/dog but with no matching for orthologous base coverage (step 4). Because chimp and macaque are so close to human and relatively little insertion/deletion has occurred, this was not deemed necessary.

S7. Well studied genes

S7.1 Definition. The RefSeq gene catalog is a set of ~18,000 human genes which have undergone manual inspection. Each entry has one (or more) reference cDNAs and also a list of publications that reference this gene. We took all the Ensembl genes that also had a RefSeq identifier and counted up how many publications were listed in the RefSeq entry. We defined a gene as well-studied if it was associated with > 5 publications. We found 5,985 well-studied genes of which 102 (1.7%) had RFC scores < 90.

S7.2 Well studied genes with fixed annotation. Closer examination of the alignment properties for these genes shows that 38 have a low RFC due to misannotation of the gene structure and these are listed below.

Ensembl id	Annotation problem
ENSG00000166548	bad exon
ENSG00000171611	bad exon
ENSG00000143842	bad exon
ENSG00000125813	bad exon
ENSG00000126549	bad exon
ENSG00000182346	bad exon
ENSG00000169877	bad region
ENSG00000136732	bad region
ENSG00000065054	bad region
ENSG00000189052	bad region
ENSG00000113520	bad region
ENSG00000134443	single indel
ENSG00000102539	single indel
ENSG00000188705	single indel
ENSG00000105398	single indel
ENSG00000188257	single indel
ENSG00000181374	single indel (fixed)
ENSG00000185918	single indel (fixed)
ENSG00000157576	wrong frame
ENSG00000168993	wrong frame
ENSG00000170290	wrong orf
ENSG00000148400	wrong orf
ENSG00000128394	wrong splice
ENSG00000159958	wrong splice

ENSG00000162493	wrong start
ENSG00000102309	wrong start
ENSG00000102805	wrong start
ENSG00000130288	wrong start
ENSG00000134248	wrong start
ENSG00000138685	wrong start
ENSG00000177047	wrong start
ENSG00000145824	wrong start
ENSG00000146232	wrong start
ENSG00000169194	wrong start
ENSG00000160180	wrong start
ENSG00000162139	wrong start
ENSG00000166301	wrong start
ENSG00000196371	wrong start

The remaining 64 do not have an obvious flaw in their gene structures. All 102 well-studied genes are listed in table S3 along with their RFC score, gene structure change, HGNC name (if any) and manual annotation comments.

S8. Orthologs

S8.1 Ortholog identification. We attempted to assign an orthologous relationship to a mouse and dog gene for each human gene. We took great care to find the correct ortholog and devised a 3 step process:

1. Using the existing Ensembl protein annotations we compared all human proteins to all mouse (and dog) proteins using blastp. We only considered similarities that occurred in syntenic positions (Mike Kamal – pers. comm.) We assigned an orthologous relationship between two genes if they could be aligned across 80% of their length.
2. In order to catch any missed orthologs that weren't detected by the protein alignment method in step two we inferred orthology by using the pairwise genomic alignments (see S1.3). If Ensembl had annotated a gene on the orthologous sequence (see S1.2.2 and S1.2.3) that covered ≥80% of the length of the human gene we assigned an ortholog.
3. Finally, if the first two methods had not called an ortholog we attempted to build an annotation in mouse and dog ourselves. An ortholog was called if the ungapped aligned cross-species sequence translated across ≥80% of its length (with at most one stop codon, to allow for possible errors in the assembled sequence) and the peptide similarity was sufficiently high (≥ 50% for mouse and ≥60% for dog).

Finally an orthologous relationship was designated simple (1:1) if only two genes were similar to each other with no other similar genes within 1Mb. All other genes were designated complex (many:many).

S8.2 Orthologs with RFC<90. Of the 18,752 human genes with orthologs in mouse or dog, a total of 601 (3%) had RFC scores below 90. In 247 (41%) cases, the RFC could be dramatically improved by making a small change to the human gene model (192) or by ignoring a single indel (55).

The small changes to the human gene model included: 66 cases in which the translational start site was not conserved, but there was a nearby conserved start site; 106 cases in which nearly all indels occurred in a very limited part of the gene (the last exon in 32 cases, a single internal exon in 30 cases, and a few adjacent exons in 30 cases), which could be deleted from the gene model; 10 cases in which the wrong splice site had apparently been chosen; 8 cases in which the wrong ORF had apparently been annotated; and 2 cases in which the wrong frame had been chosen.

After making these changes, we were left with 354 genes with RFC<90 (2% of the total orthologs). Roughly 50% of these have RFC>80 (vs random 4% for random ORFs). The ORF length distribution for these genes is non-random with 50% having ORFs of 570bp or more and 30% over 1000bp (vs 15% and 1% for random ORFs).

Of the 103 long ORFs (>1000bp), the vast majority (97) have RFC>70. They are highly enriched in segmental duplications.

S9 - Example gene report cards

A gene report card was produced for each Ensembl transcript, using the pairwise alignments to mouse and dog. These show graphically many properties of the gene and its orthologous alignment to mouse and dog in order for the viewer to easily evaluate the validity (or not) of the gene. Gene report cards for every Ensembl (v35) transcript can be accessed from http://www.broad.mit.edu/~mclamp/alpheus/[+].

S10 – Identification of genes with paralogs within human

S10.1 Identification of paralogs

Human genes were labeled as paralogs if an ortholog could not be found in mouse and dog but the gene was similar to another gene in the Ensembl human catalog. All peptides from human genes were compared to one another using blastp, and only those aligning over ≥80% of both the query and target gene were considered (no score threshold was used). The closest match was identified as the gene with the highest percent identity.

The genes with no orthologs but with paralogs within the human genome were split into two categories: those whose closest paralog had an ortholog in mouse or dog (cross-species paralogs) and (ii) those whose close paralog (or any qualifying paralog) did not have an ortholog in mouse or dog (human specific paralogs).

S10.2 Refinement of the paralog set

[+] This is a temporary address. A permanent address will be available on publication.

S10.2.1 Human-specific paralogs. Of the 68 cases initially identified as human paralogs, 9 were reclassified as pseudogenes, 5 more as transposons and 3 were artifacts in 3'-UTRs of other genes.

S10.2.2 Cross-species paralogs. The initial count of 155 cross-species paralogs had 11 that were reclassified (8 had fixable annotation resulting in a mouse or dog ortholog

Ensembl id	Refinement class
ENSG00000099804	fixable_paralog
ENSG00000137204	fixable_paralog
ENSG00000136273	fixable_paralog
ENSG00000131864	fixable_paralog
ENSG00000126231	fixable_paralog
ENSG00000125285	fixable_paralog
ENSG00000106404	fixable_paralog
ENSG00000105146	fixable_paralog
ENSG00000196685	paralog_artifact
ENSG00000196563	paralog_artifact
ENSG00000198676	paralog_artifact
ENSG00000183279	paralog_pseudogene
ENSG00000185682	paralog_pseudogene
ENSG00000186678	paralog_pseudogene
ENSG00000187826	paralog_pseudogene
ENSG00000188327	paralog_pseudogene
ENSG00000188841	paralog_pseudogene
ENSG00000169362	paralog_pseudogene
ENSG00000183247	paralog_pseudogene
ENSG00000112790	Paralog_pseudogene

S11 – Identification of genes with known protein domains

We next searched for genes that have a known protein domain. The widely used Pfam database was used as our source of domain information.

S11.1 Refinement of the gene class with known protein domains. Of the 97 genes originally identified as being without ortholog or paralog but containing a pfam domain, the following 61 (16 artifacts, 40 gene structure changes resulting in orthologs in mouse/dog, 5 missed pseudogenes) were reclassified.

ENSG00000143631	pfam_artifact
ENSG00000162621	pfam_artifact
ENSG00000165296	pfam_artifact
ENSG00000159450	pfam_artifact
ENSG00000188553	pfam_artifact
ENSG00000187827	pfam_artifact
ENSG00000189429	pfam_artifact
ENSG00000196293	pfam_artifact
ENSG00000196300	pfam_artifact
ENSG00000196515	pfam_artifact

ENSG00000188211	pfam_artifact
ENSG00000197570	pfam_artifact
ENSG00000198942	pfam_artifact
ENSG00000197778	pfam_artifact
ENSG00000096224	pfam_artifact
ENSG00000198344	pfam_artifact
ENSG00000180233	pfam_ortholog
ENSG00000182064	pfam_ortholog
ENSG00000183666	pfam_ortholog
ENSG00000186704	pfam_ortholog
ENSG00000186629	pfam_ortholog
ENSG00000186470	pfam_ortholog
ENSG00000188506	pfam_ortholog
ENSG00000188613	pfam_ortholog
ENSG00000188642	pfam_ortholog
ENSG00000188818	pfam_ortholog
ENSG00000126545	pfam_ortholog
ENSG00000197990	pfam_ortholog
ENSG00000198904	pfam_ortholog
ENSG00000196502	pfam_ortholog
ENSG00000180081	pfam_ortholog
ENSG00000186144	pfam_ortholog
ENSG00000155130	pfam_ortholog
ENSG00000153468	pfam_ortholog
ENSG00000083544	pfam_ortholog
ENSG00000142408	pfam_ortholog
ENSG00000140839	pfam_ortholog
ENSG00000136205	pfam_ortholog
ENSG00000136098	pfam_ortholog
ENSG00000131165	pfam_ortholog
ENSG00000129038	pfam_ortholog
ENSG00000126264	pfam_ortholog
ENSG00000125787	pfam_ortholog
ENSG00000123576	pfam_ortholog
ENSG00000105499	pfam_ortholog
ENSG00000158683	pfam_ortholog
ENSG00000159710	pfam_ortholog
ENSG00000175658	pfam_ortholog
ENSG00000172689	pfam_ortholog
ENSG00000172687	pfam_ortholog
ENSG00000172458	pfam_ortholog
ENSG00000169887	pfam_ortholog
ENSG00000169876	pfam_ortholog
ENSG00000165949	pfam_ortholog
ENSG00000164877	pfam_ortholog
ENSG00000102043	pfam_ortholog
ENSG00000177219	pfam_pseudogene
ENSG00000185671	pfam_pseudogene
ENSG00000183935	pfam_pseudogene
ENSG00000188075	pfam_pseudogene

ENSG00000188459	pfam_pseudogene

S12. Orphans

S12.1. Orphans with tandem repeats. Tandem repeats are short stretches of DNA that are present in multiple consecutive copies in the genome. If the unit that makes up the repeat does not contain a stop codon and many copies of the repeat are present, then a long ORF can result – that is, one that is much larger than expected by chance in random DNA sequence. To identify such tandem repeats, we again the Ensembl repeat annotation of the human genome. This uses the program trf (tandem repeat finder), which looks for stretches of matching nucleotides (k-tuples) which are separated by a common distance.

S12.2 Refinement of the orphan gene class. Examination of the high rfc (>90) and long ORF (>1000) orphans identified 68 genes that, after small modification of their gene structures would have orthologs in mouse or dog. Most cases involved the removal of a single human exon.

Ensembl id	HGNC name
ENSG00000062582	MRPS24
ENSG00000083093	
ENSG00000086289	
ENSG00000112787	
ENSG00000135185	C7orf23
ENSG00000135272	
ENSG00000137234	
ENSG00000137434	C6orf52
ENSG00000142698	
ENSG00000147174	ACRC
ENSG00000148671	C10orf116
ENSG00000150510	
ENSG00000151773	
ENSG00000152291	TGOLN2
ENSG00000155258	C21orf99
ENSG00000156363	
ENSG00000159182	
ENSG00000160844	
ENSG00000161877	
ENSG00000163800	C2orf16
ENSG00000171388	APLN
ENSG00000174456	
ENSG00000175772	
ENSG00000176075	C1orf46
ENSG00000178475	
ENSG00000178645	C10orf53
ENSG00000178789	
ENSG00000179088	
ENSG00000179954	
ENSG00000180422	
ENSG00000180458	

ENSG00000181995	
ENSG00000182244	
ENSG00000183359	
ENSG00000183562	
ENSG00000183673	
ENSG00000184090	
ENSG00000184206	
ENSG00000184293	
ENSG00000184629	
ENSG00000185065	
ENSG00000186322	
ENSG00000187092	
ENSG00000187172	BAGE2 BAGE5 BAGE4
ENSG00000187257	
ENSG00000188205	
ENSG00000188305	
ENSG00000188385	
ENSG00000188388	
ENSG00000188405	
ENSG00000196243	
ENSG00000196467	
ENSG00000196614	
ENSG00000196624	
ENSG00000196823	
ENSG00000197092	
ENSG00000197520	
ENSG00000197627	
ENSG00000197630	
ENSG00000197648	
ENSG00000197715	
ENSG00000197775	C14orf167
ENSG00000197982	
ENSG00000197983	
ENSG00000198050	
ENSG00000198391	
ENSG00000198426	
ENSG00000198679	

S12.3 Examination of orphans with long ORFs. We found a slight excess of very long ORFs, but close examination revealed that the excess is largely due to 11 putative genes with no cDNA evidence. These cases had been constructed by *de novo* gene prediction programs that concatenate potential exons; because the programs deliberately avoid stop codons, they produce gene predictions with longer ORFs than expected by chance. We conclude that the ORF lengths of the orphans are consistent with a random distribution.

Ensembl id	ORF length	Chr	Chr start	Comment
ENSG00000196820	2601	1	158223760	No full length cDNA support.
ENSG00000171271	1089	10	102872526	This is an oddity. There is no sequenced cDNA but the reference sequence is a 20kb piece of genomic sequence. The reference says this gene was found by looking for large ORFs in the 20kb sequence and confirming them by RTPCR. Only 10% of the cDNA was used for the RT-PCR step and no subsequence EST or mRNA sequencing has confirmed this gene.
ENSG00000196607	1497	11	134110199	No cDNA and 77% tandem repeat.
ENSG00000185823	3471	15	22472108	There is no human cDNA in the UCSC browser but there is a refseq. Examination of this entry references a genbank/embl accession of AC100720 which is a genomic clone. The pubmed reference says this is transcribed only in testis. As there is no submitted cDNA I suspect this not to be coding but it is still highly unlikely to find an ORF in human DNA of this length.
ENSG00000162002	1368	16	695580	No cDNA. This gene is in a region which is the first 2Mb of chr 16. This region was annotated using a combination of transcribed sequences and computational predictions. This is a computational prediction with no full length cDNA support
ENSG00000162012	1920	16	1086812	No cDNA. This gene is in a region which is the first 2Mb of chr 16. This region was annotated using a combination of transcribed sequences and computational predictions. This is a computational prediction with no full length cDNA support
ENSG00000167945	1209	16	795444	No cDNA. This gene is in a region which is the first 2Mb of chr 16. This region was annotated using a combination of transcribed sequences and computational predictions. This is a computational prediction with no full length cDNA support
ENSG00000167949	1308	16	1020255	No cDNA. This gene is in a region which is the first 2Mb of chr 16. This region was annotated using a combination of transcribed sequences and computational

				predictions. This is a computational prediction with no full length cDNA support
ENSG00000167953	1074	16	1280070	No cDNA. This gene is in a region which is the first 2Mb of chr 16. This region was annotated using a combination of transcribed sequences and computational predictions. This is a computational prediction with no full length cDNA support
ENSG00000174109	1071	16	1409945	No cDNA. This gene is in a region which is the first 2Mb of chr 16. This region was annotated using a combination of transcribed sequences and computational predictions. This is a computational prediction with no full length cDNA support
ENSG00000113553	1611	5	133086284	No full length cDNA support.

S12.4 Orphans with published evidence of peptide (12 cases). The 12 orphans with published evidence that they encode a peptide are:

(1) HTN1 (histatin 1 ENSG00000126550). Literature says this is a member of a gene family (other members HTN3, HTN5) and is only expressed in saliva. Similarity is only over the 5' end of the gene. One paper finds peptide but another (mass spec) doesn't.

(2) DCD (Dermicidin ENSG00000161634). Only expressed in sweat glands. Original protein is processed into a 47aa peptide (detemined by mass spec.) which shows antimicrobial activity.

(3) PBOV1 (Also named UROC28 ENSG00000177343). Protein was detected in serum by Western blot. Up-regulated in prostate, breast and bladder cancer.

(4) PRH2 (ENSG00000134551). Proline rich protein. Evidence of yeast 2-hybrid interaction with MUC7 and far Western blots. Expressed in saliva.

(5) LACRT (Lacritin ENSG00000135413). Expressed only in tears and peptide evidence from Western blot. Is definitely transcribed in macaque. The paper says the protein shows 99% identity to mouse over 2/3 of its length but there is no orthologous mouse sequence at all.

6) FMR1NB (ENSG00000176988). Gene is named in the paper NY-SAR-35. SEREX analysis showed testis and tumor specificity. Mouse cDNA exists.

(7,8) FAM9A, FAM9C (ENSG00000177138, ENSG00000187268). These are meant to be part of a gene family (the other one is FAM9B which has a Cor1 pfam domain over 75% of its length). The genes are similar to one another but only in regions.

The papers reports a similarity to SYCP3. FAM9B is similar to this gene in that they contain the same pfam domain but FAM9A and FAM9C show no similarity (using blastp). EGFP fluorescence showed presence in the nucleus and the nucleolus.

(9) IL32 (ENSG00000008517) . This gene is expressed in lymphocytes and expression is increased after activation of T-cells or NK cells. The NK4 protein has been purified using His6 tag affinity, size exclusion and ion exchange chromatography. It has been shown to have typical cytokine activity, induces

TNFalpha and MIP-2 and has been renamed IL-32. This is a 6 exon gene with all splices conserved in dog but only a small amount of orthologous mouse sequence. The peptide identity is very low (31%) but only one of the dog indels put the dog sequence out of frame. In the long, final exon most of the indels are frame preserving suggesting this could be a highly divergent gene.

(10) CROC4 (ENSG00000125462) . Publication reports peptide (Western blot) but the conservation with mouse and dog suggests this is a spurious gene.

(11) Oculomedin (ENSG00000198775) There is no orthologous mouse sequence for this gene and the dog alignment has a few frameshifting indels which introduce stop codons. One publication does report peptide (Western blot) for this gene (names oculomedin) however.

(12) SECTM1 (ENSG00000141574) Western blot shows evidence of peptide

Table S4 lists all orphans with associated publications along with comments on any peptide evidence.

S13 – Updated Ensembl annotation (v38)

The main portion of the paper is based around Ensembl version 35 (November 2005). An updated annotation of the human genome (on a newer assembly - version hg18) had been performed and appeared in Ensembl version 38. To identify any novel genes in this set we mapped all new genes back to the version of the assembly used for Ensembl35 (hg17). As Ensembl tracks genes by identifiers we first identified all new gene identifiers in Ensembl 38. We then took their DNA and aligned them back to the hg17 assembly using a fast, near exact, alignment method (blat). As we expect these genes to be near 100% identical on both assemblies, we filtered the matches to be 95% identical and ≥ 80% of their DNA to align to the genome.

Of the original 23,341 gene in Ensembl (v38), a total of 3,249 had novel identifiers and 2,908 of these mapped to the hg17 assembly. We eliminated 608 that overlapped with existing genes, 472 that were mostly (>80%) repeat sequence, 95 were reused for existing genes and 663 which were pseudogenes. Of the remaining 1,077 genes 516 were retained as they had orthologs (according to our test) in mouse or dog. A further 38 had multi-species paralogs, 22 had human specific paralogs and 22 had pfam domains and were also retained as valid genes. This resulted in 462 genes classed as orphans. (A table of all novel Ensembl38, RefSeq and Vega genes and their assigned classes as well as a list of the novel, deleted and reused genes from Ensembl38 is available in the supplementary data).

S14 Refseq and Vega gene catalogs

The Refseq catalog contained 18,044 genes after clustering overlapping transcripts. After removing any genes that overlapped with Ensembl v35 and v38 there were 418 remaining genes. 64 of these were eliminated as retroposons or mostly consisting of tandem repeat (11 transposons, 19 tandem repeat, 34 pseudogenes). 248 qualified as orthologs according to our criteria and a further 44 were admitted as genes due to cross-species or human paralogy or possession of a pfam domain (34 cross-species paralogs, 7 human-specific paralogs and 3 pfam domains).

The Vega catalog originally contained ~32,000 distinct gene identifiers for protein-coding genes but, after clustering all overlapping transcripts, and taking a representative transcript, this reduced to only 19,113 non-overlapping genes. There were 2,183 genes that didn't overlap Ensembl v35, v38 or Refseq which, after removal of retroposons and tandem repeats (552 transposons, 154 pseudogenes, 20 tandem repeats) left 1,457 genes. 334 of these qualified as orthologs according to our criteria with a further 48 admitted to the catalog due to paralogy or possession of a pfam domain (21 cross-species paralogs, 19 human-specific paralogs and 8 genes with pfam domains). The remaining 1,095 were labeled orphans.

A table of all novel RefSeq and Vega genes and their classification is available in the supplementary data.

V. Supplementary Data

- Report cards for all genes in Ensembl v35 are available from http://www.broad.mit.edu/~mclamp/alpheus*
- The snapshot of RefSeq taken on March 27th 2007 is available from htt://www.broad.mit.edu/~mclamp/alpheus/data/Refseq.032707.txt
- A table of all Ensembl35 genes, their automatically assigned classes and their class after manual inspection is available from http://www.broad.mit.edu/~mclamp/alpheus/data/Ensembl35.txt.
- A table of all well-studied genes and their properties is available from http://www.broad.mit.eud/~mclamp/alpheus/data/well-studied.txt.
- A table of all novel Ensembl38, RefSeq and Vega genes and their assigned classes is available from http://www.broad.mit.edu/~mclamp/alpheus/data/Ensembl38_RefSeq_Vega.txt.
- A list of the novel, deleted and reused genes from Ensembl38 is available from http://www.broad.mit.edu/~mclamp/alpheus/data/Ensembl38.ids.
- A table of all human specific genes (from Ensembl35, Ensembl38, RefSeq, Vega) is available from http://www.broad.mit.edu/~mclamp/alpheus/data/human_specific_genes.txt.

* This is a temporary address. A permanent address will be supplied.

VI. References cited in supplementary information

1. Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-3.0. http://www.repeatmasker.org (1996-2004).

2. Dust low complexity masking http://blast.wustl.edu/pub/dust/

3. Benson G, Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, **27**, 573-580 (1999).

4. Altschul, SF, Gish, W, Miller, W, Myers, EW, and Lipman, DJ. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403-410 (1990).

5. Sonnhammer, ELL, Durbin, R. A workbench for Large Scale Sequence Homology Analysis. *Comput. Applic. Biosci*. **10**, 301-307 (1994).

6. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human-mouse alignments with BLASTZ. *Genome Res*. **13**,103-107 (2003).

7. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. **31**, 3497-500 (2003).

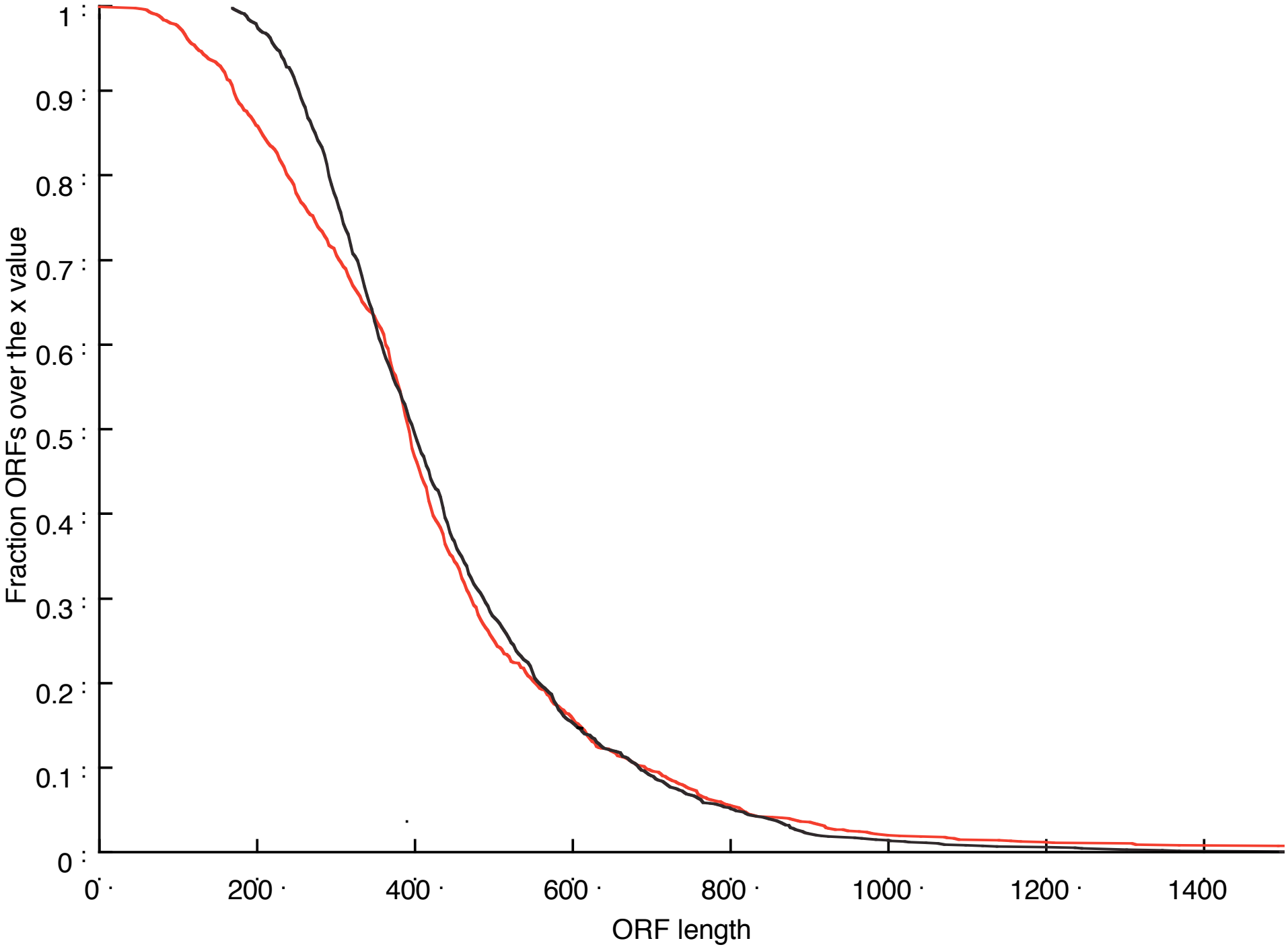
8. Birney E, Clamp M, Durbin R, GeneWise and Genomewise. *Genome Res*. **14**, 988-995 (2004).

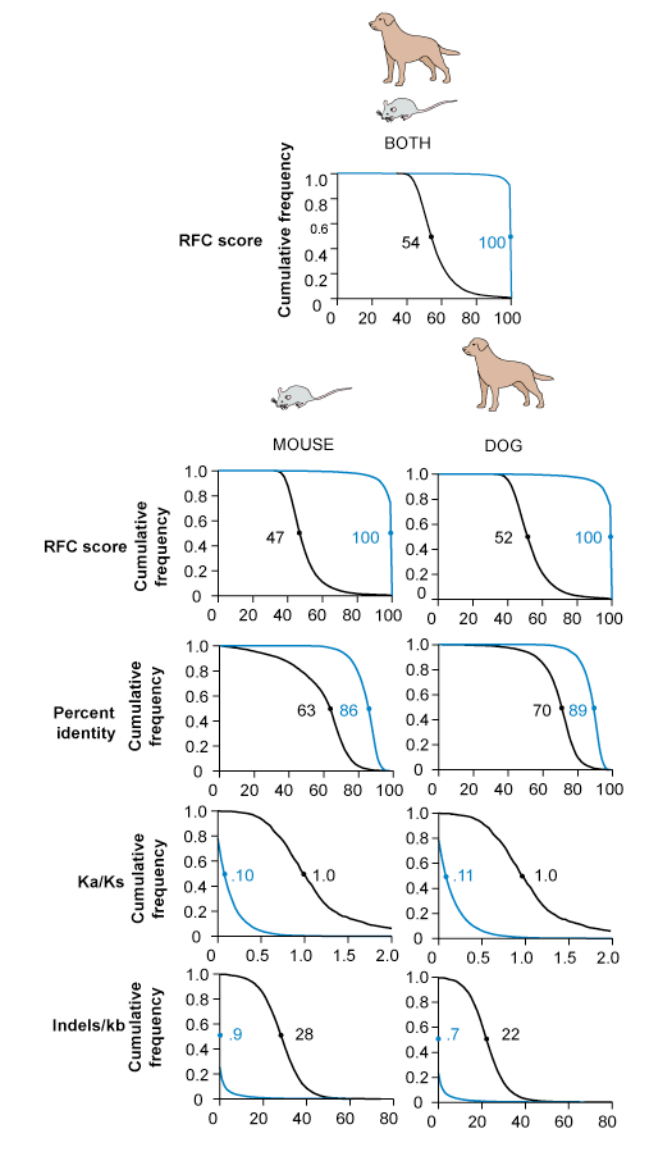
9. Pfam: clans, web tools and services Robert D. Finn, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, Mhairi Marshall, Ajay Khanna, Richard Durbin, Sean R. Eddy, Erik L. L. Sonnhammer and Alex Bateman. *Nucleic Acids Research* Database Issue **34**,D247-D25 (2006).

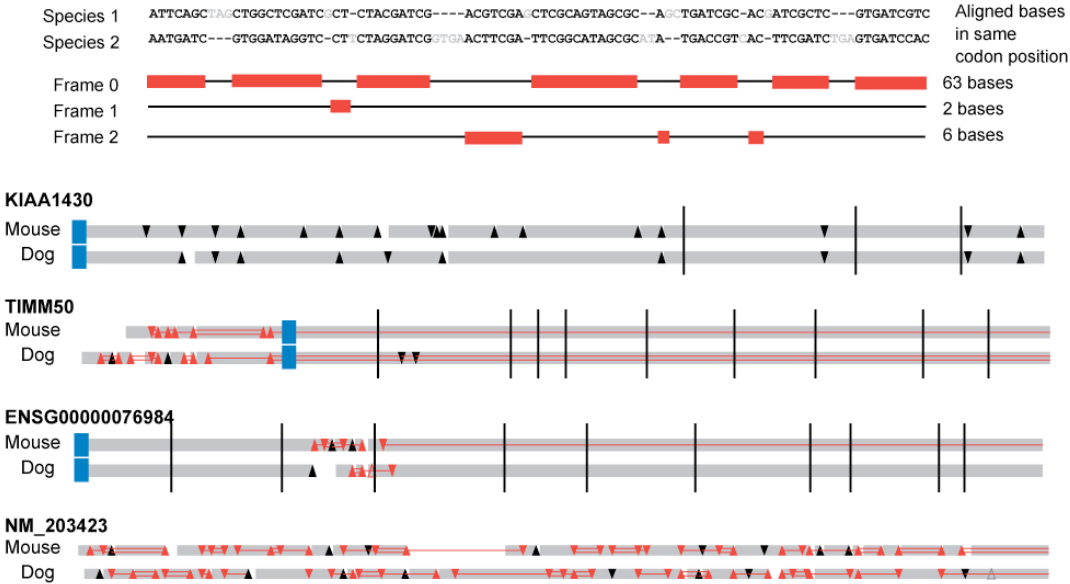
10. Eddy SR: Profile hidden Markov models. *Bioinformatics*, **14**,755-763 (1998).

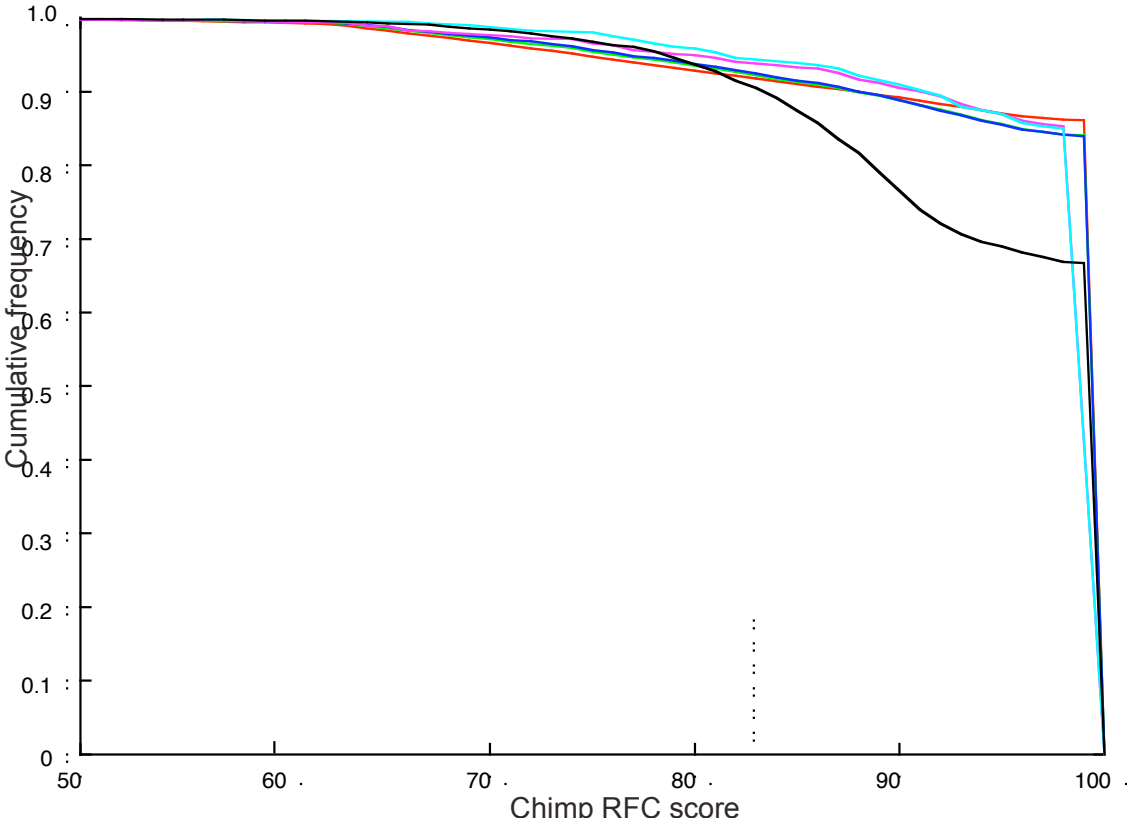
11. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. **13**,555-556 (1997).

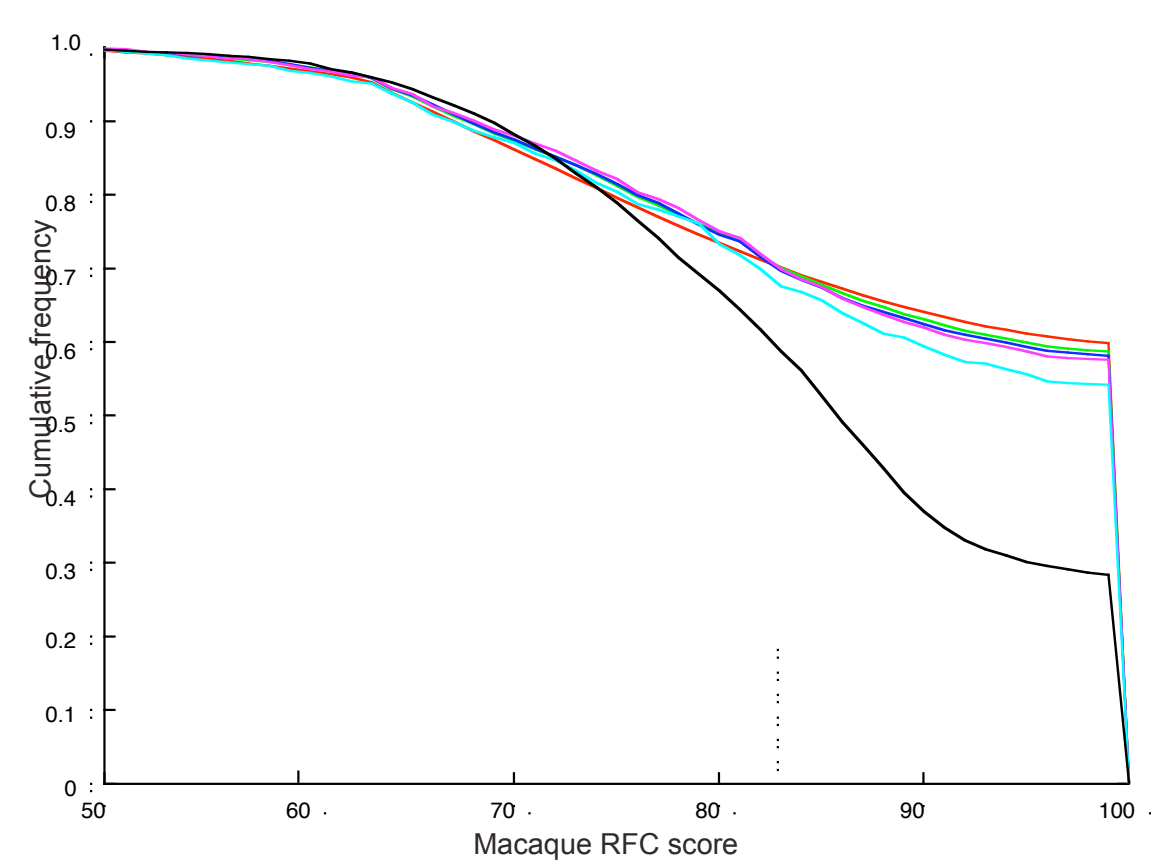
12. Curwen, V, *et al,* The Ensembl automatic gene annotation system. *Genome Re*s. **14**, 942-50 (2004).

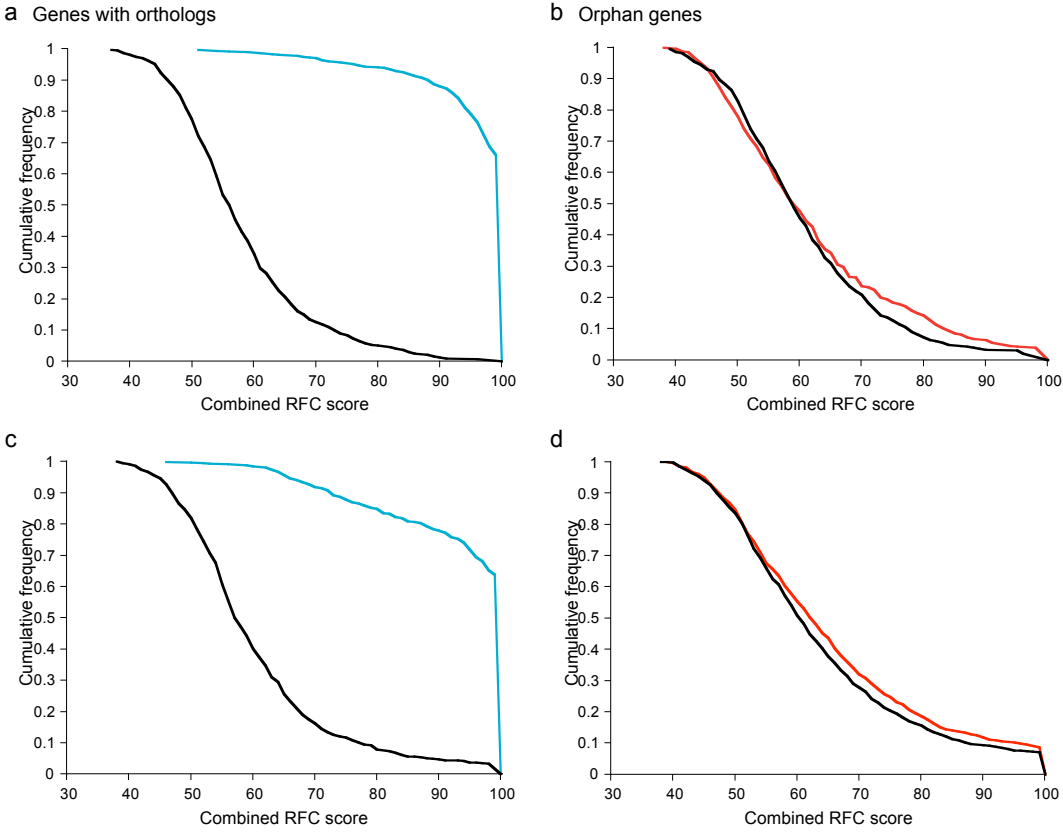

Fraction ORFs over the x value
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400
ORF length

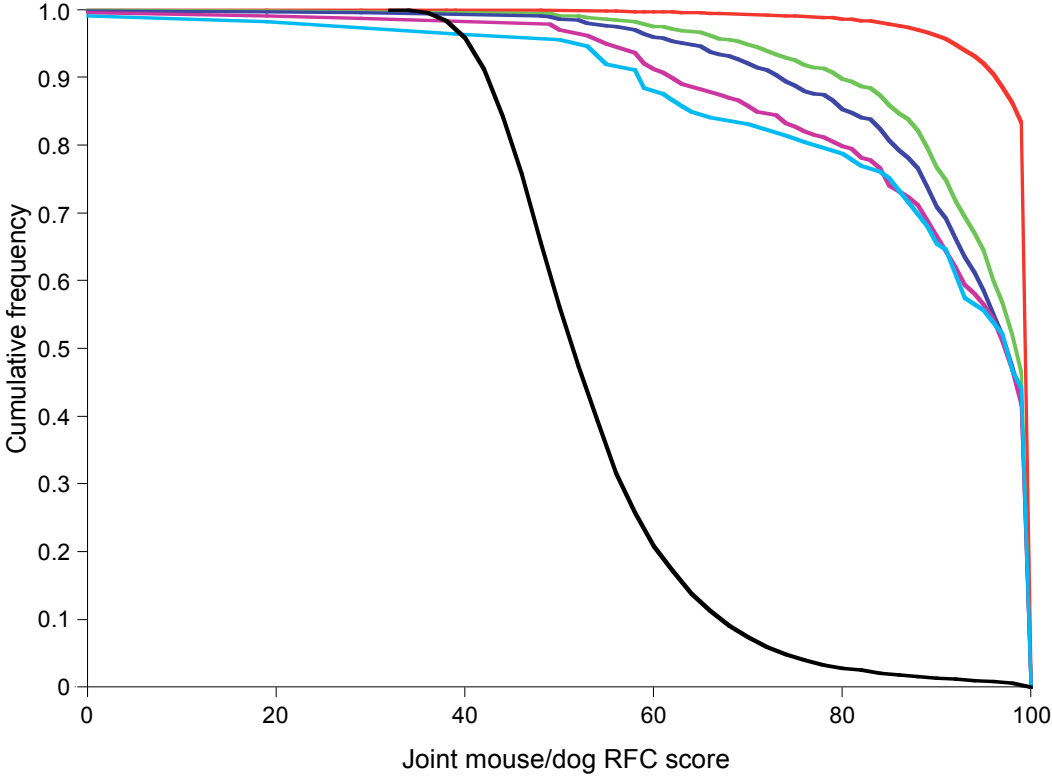

BOTH
RFC score
Cumulative frequency
54
100
MOUSE
DOG
RFC score
47
100
52
100
Percent identity
63
86
70
89
Ka/Ks
.10
1.0
.11
1.0
Indels/kb
.9
28
.7
22


Species 1
Species 2
Aligned bases in same codon position
Frame 0
Frame 1
Frame 2
63 bases
2 bases
6 bases
KIAA1430
Mouse
Dog
TIMM50
Mouse
Dog
ENSG00000076984
Mouse
Dog
NM_203423
Mouse
Dog


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Cumulative frequency
50
60
70
80
90
100
Chimp RFC score


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Cumulative frequency
50
60
70
80
90
100
Macaque RFC score


a
Genes with orthologs
b
Orphan genes
c
d
Cumulative frequency
Combined RFC score
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
30
40
50
60
70
80
90
100


Cumulative frequency
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
Joint mouse/dog RFC score