<table>
<thead>
<tr>
<th>Type</th>
<th>Ensembl Single</th>
<th>RefSeq Single</th>
<th>UniProtKB Single</th>
<th>Ensembl - RefSeq</th>
<th>Ensembl – UniProtKB</th>
<th>RefSeq - UniProtKB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt genome sequence</td>
<td>6</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antisense</td>
<td>7</td>
<td>260</td>
<td>84</td>
<td>96</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Duplicate (technical)</td>
<td>41</td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>IG/TR genes</td>
<td>6</td>
<td>120</td>
<td>33</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LncRNA</td>
<td>141</td>
<td>126</td>
<td></td>
<td>1</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Not in reference</td>
<td>9</td>
<td>160</td>
<td>104</td>
<td>78</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Other ncRNA</td>
<td>39</td>
<td>40</td>
<td></td>
<td>23</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Sense overlapping</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudogene</td>
<td>2</td>
<td>165</td>
<td>373</td>
<td>101</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Read-through</td>
<td>38</td>
<td>5</td>
<td>7</td>
<td>358</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Retroviral gene</td>
<td></td>
<td></td>
<td></td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sense intronic</td>
<td>2</td>
<td>31</td>
<td></td>
<td></td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

Table S1. The annotations of genes not classified as coding in all three sets
The table shows the alternative classification for those genes classified as coding by just one or two reference sets. Genes that are not present in other reference sets are labelled as “Not in reference”. Genes annotated, but not in the reference set are tagged as “Alt genome sequence”.

Figure S1. Novel human genes
There are sixteen coding genes that Compara highlights as novel human genes in GENCODE v24. All are single exon genes and are predicted by Ensembl automatic prediction programs. All but one (see AC009060.3 in the figure) would code for proteins with 124 amino acids. Although many are now “obsolete” in UniProt, some are annotated with the tag “FKSG”. UniProt currently annotates 22 FKSG proteins, all with more or less 124 amino acids, while GENCODE v24 annotates 11 of these UniProtKB FKSG genes. None of these novel human genes have their coding status supported by any of the databases or any by reliable peptide or antibody evidence. They do however have 90% identity to “proteins” in Corethrella appendiculata, Streptococcus pneumonia and Bacillus cereus.
Figure S2. UniProt evidence for genes classified as coding by different sets of manual annotators

The distribution of UniProt evidence codes across subsets of UniProtKB genes: those genes that are classified as coding by all three databases (Intersection), genes that are classified as coding by UniProtKB and by RefSeq, genes classified as coding by UniProtKB and by Ensembl/GENCODE, and genes are classified as coding solely by UniProtKB.
Figure S3. Potential non-coding genes in GENCODE 12 and GENCODE 24
A. The counts of GENCODE 24 potential non-coding genes that were potential non-coding genes in GENCODE 12 too (PNC in Gv12 and Gv24), of GENCODE 24 potential non-coding genes that were likely coding genes in GENCODE 12 (PNC in Gv24, but LCG in Gv12) and of genes that have been annotated in the coding reference set since GENCODE 12. B. The number of genes annotated with protein evidence codes Uncertain, Predicted and Homology by UniProtKB in the GENCODE 12 and GENCODE 24 reference sets. UniProtKB has reviewed the evidence codes for many UniProt entries.
Figure S4. The overlap between genes annotated as coding by three sets of manual annotators

In A the overlap between coding genes in the Ensembl/GENCODE, RefSeq and UniProtKB reference sets. In B the number of coding genes tagged as potential non-coding in the Ensembl/GENCODE reference and how they overlap with the other reference sets. In C the percentage of genes in each of the four relevant sets that are flagged as PNC genes.
Figure S5. Tissue expression of potential non-coding genes and likely coding genes

Genes binned by number of tissues in which transcripts were detected with a tissue count of more than 1TPM in the 36 tissues of the Human Protein Atlas RNAseq experiments. Tissue distribution shown for the likely coding genes (LCG Intersection), for potential non-coding genes annotated as coding by all three reference sets (PNC Intersection) and for potential non-coding genes annotated by just one or two sets of annotators (PNC Subsets).
Figure S6. Strength of tissue expression of different types of potential non-coding genes and likely coding genes

For this figure we separated the olfactory receptors from all other genes. Olfactory receptors here are a control set; even if they do code for proteins almost all should be entirely expressed in the nasal tissue, which was not one of the tissues interrogated in the Human Protein Atlas experiment, so any coding from this tissue can be regarded as either biological or technical noise. Read-through genes and immunoglobulin and -t-cell receptors were left out of the figure for clarity. Remaining potential non-coding genes were split into four groups, those that were clearly “coding” (those for which we had evidence in PeptideAtlas and the Human Protein Atlas), those that were labelled in RefSeq as “antisense”, and the rest were labelled either “pseudogene” or general “non-coding” based on whether we could detect protein-like features such as protein structural or functional domains, or clear cross-species conservation signals. The pseudogene group was the largest subset of potential non-coding genes with 1,003 members. We plotted the maximum tissue expression of these four sets against the olfactory receptors and the set of likely coding genes. Genes in each set were ordered by their maximum TPM from the 36 tissues in the Human Protein Atlas experiments and we plotted the cumulative ranking of the genes in each set against this maximum TPM. The results can be seen in the figure above. The potential non-coding that were tagged as clearly coding (Coding PNC in dark red) have expression distribution that is highly similar to the known coding genes, while the possible pseudogenes, non-coding and antisense genes from the potential non-coding sets generally have much less expression, though most have much more expression than the olfactory receptors.
Figure S7. The overlap between three experimental protein detection methods
In A the number of likely coding genes detected in our analysis of the Human Protein Atlas (HPA) and PeptideAtlas (PA) databases. In B the potential non-coding genes detected in the Human Protein Atlas, PeptideAtlas databases. In C the percentage of the genes detected in each experiment that were potential non-coding genes.
Figure S8. Genomic variation in for subsets of genes in RefSeq and Ensembl/GENCODE
High impact variant percentage (yellow) and non-synonymous/synonymous ratios (blue) for genes classified as coding by all three databases (Intersection), by UniProtKB and by Ensembl/GENCODE only (Ensembl-UniProt), by UniProtKB and RefSeq only (RefSeq-UniProt), by Ensembl/GENCODE (Ensembl Unique) only and by RefSeq only (RefSeq Unique). The darker colours show the values for common variants and the lighter shades show the values for rare variants in each set of bars. 95% confidence intervals are shown.
Figure S9 – Human variation and possible non-coding features
The percentage of high impact variants for rare alleles (light yellow) and common alleles (orange) and the non-synonymous/synonymous ratios for rare alleles (light blue) and common alleles (blue) for genes tagged with a range of potential non-coding features. Read-through genes have lower non-synonymous to synonymous ratio in common alleles than in rare alleles and a somewhat lower percentage of high impact variants than the rest of the potential non-coding genes. Since read-through genes are generally composed of exons from two or more coding genes, a certain similarity with likely coding gene variation patterns is to be expected. Read-through genes were excluded from the other sets of potential non-coding features. Where there were fewer than 200 common variants in the genes that had a potential non-coding feature, this feature was excluded from the figure.
Figure S10. Genomic variation by RefSeq gene type
The percentage of high impact variants (yellow) and non-synonymous/synonymous ratios (blue) from the variants in the 1000 Genomes project for genes with RefSeq type “MODEL” (automatically predicted genes) that were unique to RefSeq (MODEL, RefSeq Unique), with RefSeq type “MODEL” and classified as coding in RefSeq and UniProtKB (MODEL, RefSeq-UniProt), with RefSeq type “MODEL” and coding in all 3 databases (MODEL, Intersection), along with GENCODE 24 automatically predicted genes that are present in all three databases (CURATE, Intersection). The darker colours show the values for common variants and the lighter shades show the values for rare variants. 95% confidence intervals are shown. Once again all sets apart from the CURATED RefSeq genes in the intersection appear to have a large proportion of genes that are under neutral selection.
Figure S11. Data supporting the non-coding status of CRIPAK
In A the alignment of the “zinc-binding” motifs in CRIPAK. In B the dot plot for CRIPAK showing the repetitive nature of the sequence. In C the expression patterns of neighbouring gene UVSSA from the Human Protein Atlas. In D the expression patterns of CRIPAK from the Human Protein Atlas. The CRIPAK and UVSSA expression patterns are highly similar suggesting a relation or confusion between the two genes.
Figure S12. The DLEU1 gene in Ensembl83
Transcripts annotated for the DLEU1 locus in the Ensembl browser taken from the Ensembl 83 archive (Ensembl 83 is the contemporary to GENCODE v24). Only one transcript (DLEU-033 in gold on the left of the picture) was protein coding, though DLEU-034 is a nonsense-mediated decay transcript.
Figure S13. PLK5 damaged functional domains and genetic variation

Polo-like kinase 5 (PLK5) is one of five human polo-like kinases, proteins that are characterized by an N-terminal kinase domain and two polo-box domains. In human PLK5 all three domains are damaged and two of the three domains have lost whole exons. Human-specific loss of coding exons is a very strong suggestion that human PLK5 is a classical unitary pseudogene. However, PLK5 is classified as coding because of a study on the role of mouse Plk5 that detected antibodies for human PLK5 \[1\]. Variants from genome-wide variation studies show that while other polo-like kinases have no high impact variants and very few non-synonymous variants, PLK5 appears to be under very different selection pressures. There are high impact common allele variants for PLK5 in both the 1,000 Genomes and ExAC [2] studies and non-synonymous/synonymous ratios are higher for common alleles than they are for rare alleles. These are small numbers of variants, just 12 common variants in 1,000 Genomes and 17 in ExAC, but they suggest that PLK5 is not subject to selective pressure and that PLK5 is probably not functional, even if it does code for a protein as some evidence suggests. In A: the structure of the human PLK2 kinase domain (PDB code: 4i6h) with the region missing in human PLK5 shown in white. B. The structure of the second human PLK2 polo-box domain (PDB code: 4xb0) with the human-specific loss in PLK5 in white. The important binding strands are shown in pink in the centre; although not lost, these residues will not fold correctly with half the structural domain missing. C. The percentage of high impact variants (yellow) and non-synonymous/synonymous ratios (blue) for the PLK gene family. To confirm the results from the 1000 Genomes for PLK5, we have also included the PLK5 variant proportions from the ExAC consortium. The darker colours show the values for common variants and the lighter shades show the values for rare variants.
Figure S14. Alignment of GVQW1 with nearest relatives from UniProtKB
The partial alignment between human GVQW1 gene product and primate genes annotated as GVQW1 in UniProtKB. There is little evidence of conservation considering how close the species are in evolutionary time. Only one of the six proteins (G3S0H5_GORGO) has the GVQW motif (shown in green).
Figure S15. Alignment of the 9 UniProtKB proteins used to identify novel coding genes in the CHESS database
We aligned just the most similar regions of each protein. Sequences were aligned with MUSCLE [3]. Seven proteins have a similar C-terminal regions, while there are two different types of N-terminal sequences, one with five sequences (Q8WTZ3, GVQW2, Q8N976, C9orf85, LINCO0269) and one with four sequences (LINCO0596, C16orf89, GVQW1 y UTY_PANTR).
References
