Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci

Jonathan M. Mudge*[,1], Irwin Jungreis*[,2,3], Toby Hunt[1], Jose Manuel Gonzalez[1], James C. Wright[4], Mike Kay[1], Claire Davidson[1], Stephen Fitzgerald[5], Ruth Seal[1], Susan Tweedie[1], Liang He[2,3], Robert M. Waterhouse[6,7], Yue Li[2,3], Elspeth Bruford[1], Jyoti S. Choudhary[4], Adam Frankish[1], Manolis Kellis[2,3,†]

*These authors contributed equally to this work

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK;

[2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA;

[3]Broad Institute of MIT and Harvard, Cambridge, MA;

[4]Functional Proteomics, Division of Cancer Biology, Institute of Cancer Research, 123 Old Brompton Road, London SW7 3RP, UK;

[5]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK;

[6]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland;

[7]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland;

†Corresponding author: manoli@mit.edu

Running title: Discovery of protein-coding genes by PhyloCSF

Keywords: novel protein-coding genes, PhyloCSF, GENCODE

Abstract

The most widely appreciated role of DNA is to encode protein, yet the exact portion of the human genome that is translated remains to be ascertained. We previously developed PhyloCSF, a widely-used tool to identify evolutionary signatures of protein-coding regions using multi-species genome alignments. Here, we present the first whole-genome PhyloCSF prediction tracks for human, mouse, chicken, fly, worm, and mosquito. We develop a workflow that uses machine-learning to predict novel conserved protein-coding regions and efficiently guide their manual curation. We analyse over 1000 high-scoring human PhyloCSF regions, and confidently add 144 conserved protein-coding genes to the GENCODE gene set, as well as additional coding regions within 236 previously-annotated protein-coding genes, and 169 pseudogenes, most of them disabled after primates diverged. The majority of these represent new discoveries, including 70 previously-undetected protein-coding genes. The novel coding genes are additionally supported by single-nucleotide variant evidence indicative of continued purifying selection in the human lineage, coding-exon splicing evidence from new GENCODE transcripts using next-generation transcriptomic datasets, and mass spectrometry evidence of translation for several new genes. Our discoveries required simultaneous comparative annotation of other vertebrate genomes, which we show is essential to remove spurious ORFs and to distinguish coding from pseudogene regions. Our new coding regions help elucidate disease-associated regions, by revealing that 118 GWAS variants previously thought to be non-coding are in fact protein-altering. Altogether, our PhyloCSF datasets and algorithms will help researchers seeking to interpret these genomes, while our new annotations present exciting loci for further experimental characterisation.

Introduction

It has been almost two decades since the first high quality sequences from the human genome became available (Venter et al. 2001; International Human Genome Sequencing Consortium 2001). Nonetheless, efforts to decipher the information contained in our genome remain ongoing, and a key challenge is to identify regions that encode protein-coding sequences (CDS). At present, the two main human gene annotation projects, Ensembl/GENCODE (henceforth GENCODE) and RefSeq, as well as the UniProt protein resource (Zerbino et al. 2018; Frankish et al. 2018; Harrow et al. 2012; O'Leary et al. 2016; The UniProt Consortium

2019), disagree on the number of human protein-coding genes (Abascal et al. 2018), and even when a gene is agreed to be protein-coding it is often unclear which transcripts within the locus are translated (Mudge et al. 2013; Tress et al. 2017). It has been historically challenging to obtain protein sequences in the laboratory in a high-throughput manner, and it remains far easier to describe the structure of transcribed regions than to ascertain their coding potential. While the number of experimentally-derived peptide sequences found in online repositories such as PRIDE (Vizcaíno et al. 2016) has risen substantially in recent years, and such datasets have been used to discover novel proteins in genomes including human (Slavoff et al. 2012), difficulties remain in the creation and interpretation of high-quality 'proteogenomics' datasets (Nesvizhskii 2014). Meanwhile, ribosome profiling (RP) circumvents the experimental challenges in working with proteins, capturing sequence from the region of an RNA molecule that is attached to a ribosome (Ingolia et al. 2009). These data have been used to suggest the biological relevance of thousands of currently unannotated vertebrate open reading frames (ORFs) (Bazzini et al. 2014; Mackowiak et al. 2015; Raj et al. 2016; Fields et al. 2015; Ingolia et al. 2011). Nonetheless, it remains unclear to what extent ribosome attachment demonstrates production of a functional protein, i.e. one that makes a direct contribution to physiology (Bazzini et al. 2014), since ORFs can also undergo translation as part of gene regulation mechanisms, and a proportion of attachments could be stochastic 'noise' (Johnstone et al. 2016; Jackson and Standart 2015; Raj et al. 2016; Guttman et al. 2013).

CDS can also be identified through sequence conservation, and both the ratio of non-synonymous to synonymous substitutions (d_N/d_S) and codon substitution frequencies can be diagnostic of protein evolution (Lin et al. 2008). The power of such 'comparative annotation' has increased in recent years as the number of vertebrate genome sequences available has moved from single to triple figures. Previously, we developed PhyloCSF (Phylogenetic Codon Substitution Frequencies) to support CDS annotation based on multi-species genome alignments (Lin et al. 2011). PhyloCSF determines whether a given alignment is likely to represent a functional, conserved protein-coding sequence by determining its likelihood ratio under coding and non-coding models of evolution. Unlike the traditional d_N/d_S test, PhyloCSF uses precomputed substitution frequencies for every possible pair of codons, trained on whole-genome data. A particular advantage of PhyloCSF is that it can classify short portions of a CDS in isolation from the full sequence, which is necessary when considering

individual exons.

We previously demonstrated the ability of PhyloCSF and its predecessor, CSF, to add CDS annotation to genomes within the *Schizosaccharomyces* (Lin et al. 2011) and *Drosophila* lineages (Lin et al. 2007; The modENCODE Consortium et al. 2010; Jungreis et al. 2011, 2016), and also to identify novel human and mouse protein-coding genes based on the alignment of 29 mammalian genomes (Lindblad-Toh et al. 2011). Meanwhile, Mackowiak et al used PhyloCSF to find 2,000 candidates for conserved short open reading frames (sORFs) in the human, mouse, zebrafish, *Drosophila melanogaster*, and *Caenorhabditis elegans* genomes, (Mackowiak et al. 2015), while Bazzini et al used PhyloCSF to score ORFs within a set of RP translations observed in human and zebrafish (Bazzini et al. 2014). However, the efficacy of PhyloCSF has thus far been judged on its ability to recover *known* CDS, and few of the novel CDS predicted by these publications have undergone rigorous validation. GENCODE seeks to describe the true set of human protein-coding genes, not a larger set of plausible models. The inclusion of false CDS could have undesirable consequences for GENCODE's users, e.g. in the interpretation of clinical variants. Thus, externally-published novel CDS are always manually re-assessed according to GENCODE criteria. While we have found that such publications may report an excess of false-positive novel protein-coding genes (Uszczynska-Ratajczak et al. 2018), they are also likely to have underreported the set of true-positives awaiting discovery because they generally targeted existing transcript catalogs, reducing the discovery space to a few percent of the genome sequence. This is also generally true of mass spectrometry and RP-based projects.

Our goals in the current study were to develop algorithms that would allow PhyloCSF to be applied across whole genomes to find and prioritise candidate novel protein-coding regions, even in regions previously thought to be intergenic; to develop a workflow to enable manual annotators to investigate those candidates using modern transcriptomics and mass spectrometry datasets, as well as cross-species comparative annotation; and to use the resulting improved annotations to recharacterise 'non-coding' variants associated with traits or diseases as protein-altering.

We use the term ‘novel’ or ‘new discovery’ to describe coding genes, coding sequences, or pseudogenes that, at the time of this study, were not considered to be coding or, respectively, pseudogenic in the species under consideration in any of the major gene catalogs, or, as far as we could determine, in the peer-reviewed literature. By ‘novel’ we do *not* mean *de novo*, i.e. arising from non-coding sequence (Schlötterer 2015); in fact many of these sequences have known orthologs in other species or paralogs in the species under consideration.

Results

Whole-genome PhyloCSF finds candidate novel coding regions

In order to find novel coding genes, coding exons, and pseudogenes, we created a ranked list of candidate genomic regions that have the evolutionary signature of coding regions but were not previously annotated as coding or pseudogenes. Because transcriptional evidence for such regions might be incomplete or missing, we used a whole-genome method unbiased by known transcription.

We first calculated the PhyloCSF score of every codon of the hg38/GRCh38 human genome reference assembly in each of the six reading frames using alignments of 29 mammalian genomes. Each codon gets a positive score if the alignment of that codon is more likely to have arisen under a model of protein-coding evolution than under a model of non-coding evolution. Because individual codon alignments do not have enough information to distinguish coding from non-coding evolution with any confidence, we combined scores of nearby codons using a Hidden Markov Model (HMM) with states representing coding and non-coding regions. The intervals in which the most likely path through the HMM is in the coding state define a set of 596,426 genomic regions, “PhyloCSF Regions”, that likely include almost all conserved coding regions, both known and novel, that generate a PhyloCSF signal, as well as many false positives.

To restrict our list to *novel* regions, we excluded 205,043 PhyloCSF Regions overlapping protein-coding sequences in the same frame that were annotated in GENCODE v23. We also excluded 234,336 regions overlapping annotated coding sequences in the “antisense frame” (the frame on the opposite strand that shares the third codon position), and 23,443 overlapping pseudogenes, because PhyloCSF often reports a

protein-coding signal on their alignments even though the locus is no longer protein-coding. We excluded 52,548 regions shorter than nine codons since the signal on such short regions is unreliable. In order to eliminate regions that are antisense to *novel* coding regions, we trained a Support Vector Machine (SVM) to distinguish PhyloCSF Regions translated on their strand from those translated on the opposite strand, using the PhyloCSF scores on the two strands and the region length (see Methods and Supplemental Fig. S1). We excluded 11,469 regions that our SVM found to be considerably more likely to be coding on the other strand. Finally, for regions that were excluded because they partially overlap an annotation, we added back the portion that does not overlap, provided it is at least nine codons long and satisfies our antisense condition. There were 4,225 such fragments, which could be 5' extensions of annotated ORFs or extensions of known exons. This left us with 73,812 "PhyloCSF Candidate Coding Regions", henceforth 'PCCRs' (Figure 1A).

Seeking novel coding sequence in a whole genome scan is a needle-in-a-haystack problem. Known coding sequences comprise less than 0.25% of the 6-frame translation of the human genome, and *novel* coding sequences presumably comprise much less. Consequently, despite the high specificity of PhyloCSF, we expect most of our PCCRs to be false positives. To determine which PCCRs are most likely to be true novel protein-coding regions we ranked them using another SVM, this one trained to distinguish true coding PhyloCSF Regions from false positives using PhyloCSF scores on the two strands, the length of the region, and the branch length of the phylogenetic tree of species present in its local alignment (see Methods and Supplemental Fig. S1). Our algorithm considers PCCRs having lower ranks to be more likely to be real coding regions.

To evaluate our ranking, we calculated the distribution of SVM scores of previously annotated coding genes (Figure 1B), and corresponding ranks, where the "rank of a novel coding gene" is defined to be the lowest rank of any PCCR that overlaps its CDS in the same frame, and the "rank of an annotated coding gene" is the rank it would have had if it had not been previously annotated, i.e., if we had not excluded PhyloCSF Regions overlapping that particular gene when constructing the PCCRs. We found that 93% of coding genes annotated in GENCODE v23 would have overlapped a PCCR, and 92% of *those* would have ranks among the best-ranked 1% of PCCRs, suggesting that most true *novel* coding genes could be discovered by examining the

best-ranked PCCRs, though PCCRs might not cover the entire CDS so further work would be needed to fully define the transcript models and CDS. Many higher-ranked regions could also indicate novel coding exons, extensions, and pseudogenes.

To facilitate the use of our whole-genome PhyloCSF scan to distinguish protein-coding regions, we created a track hub for the UCSC and Ensembl Genome Browsers (Casper et al. 2017; Zerbino et al. 2018) with tracks for the raw PhyloCSF score of every codon, the HMM-smoothed scores, the PhyloCSF Regions, the PCCRs, and splice site predictions using the maximum entropy method (Yeo and Burge 2004) (Figure 1C). We have also created browser tracks and PCCR lists for mouse, chicken, fly (*D. melanogaster*), worm (*C. elegans*), and mosquito (*A. gambiae*). The details page for each PCCR includes a link to view the color-coded alignment of the region in CodAlignView (I Jungreis, MF Lin, CS Chan, M Kellis 2016), and other relevant information. The PhyloCSF tracks differ from other conservation browser tracks such as phyloP (Pollard et al. 2010) and phastCons (Siepel et al. 2005) in that the PhyloCSF tracks represent a signal of constraint *specifically* for protein-coding function, whereas the signal represented by other tracks is independent of the cellular function imposing the constraint (Supplemental Fig. S2).

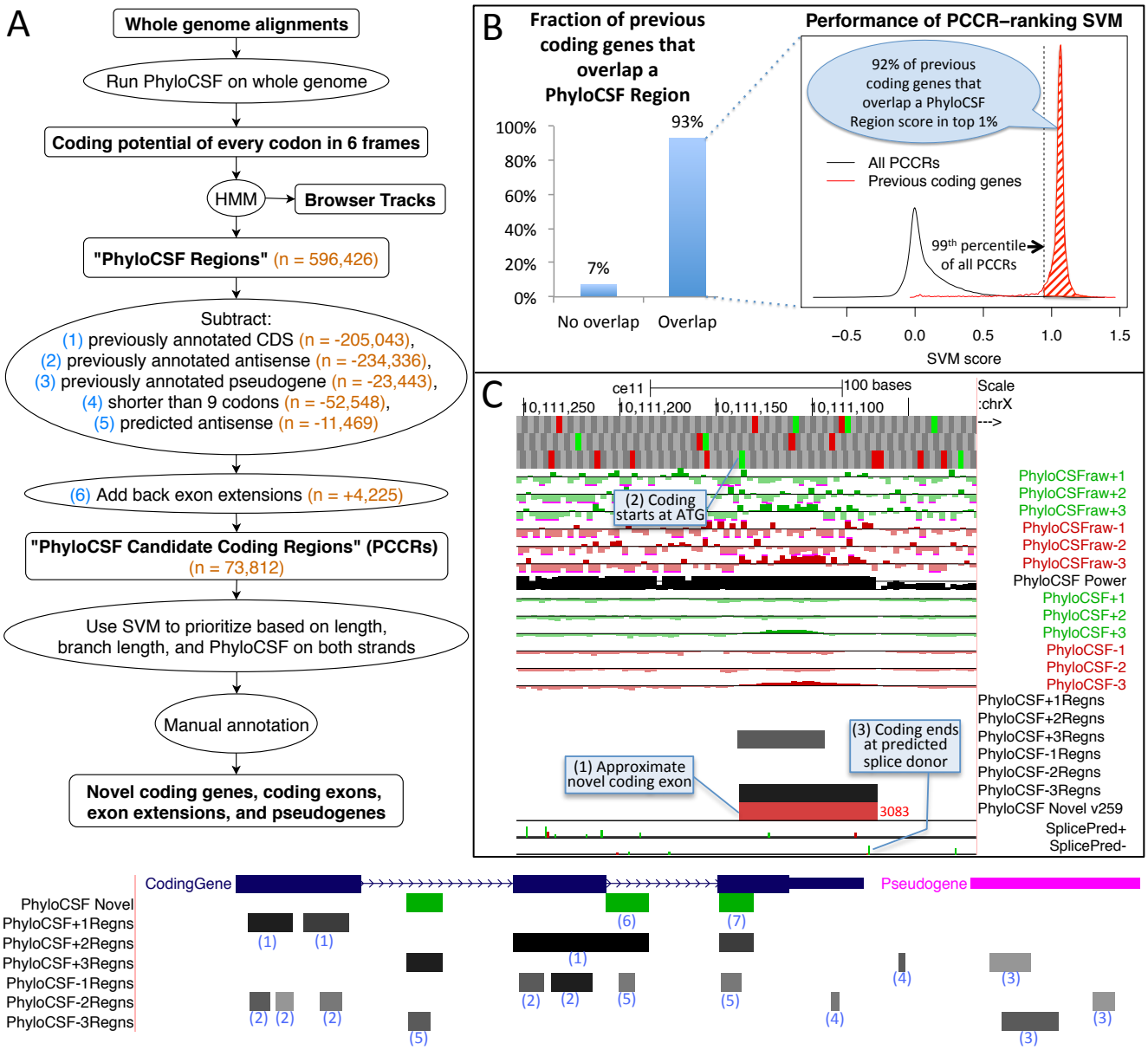


Figure 1. **Computing PhyloCSF Candidate Coding Regions. (A)** Flow chart of overall process. Numbers in orange are counts for the human hg38 assembly relative to the GENCODE v23 gene set. The hypothetical browser image at the bottom illustrates how the PhyloCSF Regions list is pruned to define PCCRs. In the vicinity of a coding gene (blue) and a pseudogene (pink), we initially have a set of intervals in each of the six possible reading frames ("PhyloCSF Regions") that are more likely to be in the coding state than non-coding state of the HMM (gray-scale intervals in the six PhyloCSF*Regns tracks). We then exclude any that overlap known coding genes in the same frame (1) or anti-sense frame (2), or that overlap known pseudogenes in any frame on either strand (3). Next, we exclude

regions less than nine codons long (4) and regions predicted by our antisense SVM to be likely antisense regions (5). Finally, we add back non-overlapping fragments of PhyloCSF Regions that partly overlap annotations, since these could be extensions of known exons (6). The resulting PhyloCSF Candidate Coding Regions are shown in green. These sometimes overlap known coding regions, and this is an indication that the PhyloCSF signal is in a different frame from the annotated one (7). The resulting PCCRs are then ranked by an SVM and investigated by expert manual annotators to find novel coding regions and pseudogenes. **(B)** Performance on previously annotated coding genes. Column chart on the left shows the fraction (93%) of GENCODE v23 coding genes that overlap at least one PhyloCSF Region; the remaining 7% could not have been identified by our workflow. Density plot on the right measures the efficiency of our PCCR-ranking SVM by showing SVM scores for all PCCRs (black) and scores of the highest-scoring PhyloCSF Region that overlaps each GENCODE v23 coding gene that overlaps at least one PhyloCSF Region (red). For 92% of such coding genes, the score is in the 99th percentile of scores of PCCRs (shaded area), indicating that manual examination of the top-ranked 1% of PCCRs would have uncovered each of these coding genes if it had not been known previously, and suggesting that most true *novel* coding genes could be identified by examining the best ranking PCCRs. **(C)** PhyloCSF Tracks in UCSC Genome Browser showing the '-' strand of *C. elegans* Chromosome X. Upper six green and red "PhyloCSFraw" tracks show the raw PhyloCSF score for each codon in each of six reading frames. The black "PhyloCSF Power" track indicates the relative branch length of the local alignment, a measure of the statistical power available to PhyloCSF; there is near full alignment for the first approximately ¾ of the track, but then there are fewer aligned species for the remaining ¼. Codons having relative branch length less than 0.1 show no scores. The next six green and red "PhyloCSF" tracks show the PhyloCSF scores smoothed by the HMM. The six "PhyloCSF*Regns" tracks show PhyloCSF Regions, with gray scale indicating the maximum probability of coding. The "PhyloCSF Novel" track shows the PCCRs in all six frames combined into a single track with green and red intervals indicating the plus and minus strands, respectively, and with the rank of the region within the list of PCCRs shown next to the region, lower ranks indicating stronger likelihood of coding. The two "Splice Pred" tracks show splice donor (green) and acceptor (red) predictions at GT and AG dinucleotides, respectively, on the plus and minus strands, with the height of each bar indicating the strength of the splice prediction. In the example shown, the tracks allow us to conjecture that there is a novel coding exon on the minus strand roughly coinciding with the 3083rd PCCR (1), extending from the ATG indicated by the small green rectangle in the third Base Position track at the top (2) up to the green splice donor prediction in the "SplicePred-" track (3).

Manual annotation of PhyloCSF regions

In order to find and annotate novel protein-coding regions, we manually examined many of the best-scoring PCCRs, clustered by chromosomal position. Firstly, we focused on the 658 clusters that contained all of the top 1000 ranked PCCRs. Secondly, we targeted the complete set of long intergenic non-coding RNA (lincRNA) models that overlapped PCCRs of any rank, in order to find mis-annotated non-coding genes. Thirdly, in order to investigate PCCRs in intergenic space, we analysed all remaining clusters up to rank 2,200 that did not

overlap any existing GENCODE annotation. Finally, we investigated several ad hoc clusters not tagged by PCCRs in the above three categories during preliminary efforts to compare the consequences of using different alignments (see Supplemental Methods section “PhyloCSF and browser tracks”).

Annotation was performed in accordance with the HAVANA guidelines for the GENCODE project (see Methods). However, an expanded approach was developed for this work that included a broad range of short-read and long-read datasets, plus detailed ‘comparative annotation’, including equivalent manual annotation of the mouse genome where possible and manual analysis of coding potential in additional vertebrate genomes (see Supplemental Methods section “Manual annotation overview”).

144 protein-coding genes and 228 kb of CDS added to GENCODE

Guided by these clusters of PCCRs, we added 144 new protein-coding genes to human GENCODE (Supplemental Data S2) and additional CDS within 236 previously annotated protein-coding genes, adding a total of 228,271 base pairs of CDS. We also added 169 new pseudogenes to GENCODE, and made extensions to 35 existing pseudogenes. The PCCR clusters analyzed and the resulting annotations are reported in Supplemental Data S1, and detailed information about each of the PCCRs in these clusters is reported in Supplemental Data S6. Supplemental Table S1 shows counts of PCCRs among the top-ranked 1000 that resulted in each kind of annotation, broken down by transcript region (overlapping CDS, extension of CDS, UTRs, etc.).

The 144 genes were classified as protein-coding because we believe that is the *most likely* interpretation of their functionality at the present time. In each case we were able to support the protein-coding status by producing either a multi-species or multi-paralog protein-sequence alignment, but we recognize that the true test of functionality for these loci will take place in the laboratory (Mudge et al. 2013). We note that PhyloCSF does not determine the transcript model containing the complete ORF, and may not even demarcate the entire translation; even a deeply conserved protein-coding gene may not have all codons or exons marked by PhyloCSF signals (see *EDDM13*; Supplemental Fig. S3A). Furthermore, a PhyloCSF signal indicates that a region has evolved *at some point in the past* as protein-coding sequence, and does not rule out that it has

been pseudogenized. This is important because vertebrate genomes are replete with pseudogenes (see below) (Pei et al. 2012; Sisu et al. 2014).

Properties of the newly added protein-coding genes

Several properties of the 144 newly added protein-coding genes help explain why they were not found previously. Firstly, the genes are enriched for small CDS: 50 translations are under 100 aa, and the median size is only 140.5 aa; less than half of the 387-aa median of all GENCODE CDS. Two examples are *SMIM31* (Figure 2A) and *SMIM41* (Supplemental Fig. S3B); both CDS were discovered within existing 'non-coding' transcript models. Small CDS are harder to identify in both manual and computational annotation pipelines (Mudge and Harrow 2016), and this problem is confounded by the fact that 28 of these 50 loci are single exon genes. It is also probable that protein size thwarted our proteogenomics pipeline (see below), as small proteins may be harder to identify in 'shotgun' mass spectrometry experiments (Nesvizhskii 2014).

Secondly, 78 out of 144 protein-coding loci were missed due to a previous lack of transcript evidence. While most GENCODE annotation is based on cDNA and EST libraries, our new workflow integrated multiple modern transcriptomics datasets. In 20 instances, the CDS was discovered after an existing 'non-coding' GENCODE model was extended to incorporate the entire reading frame. In 15 other cases, CDS annotation required the discovery of an alternatively spliced transcript within a 'non-coding' locus, as illustrated by *C10orf143* (Figure 2B) and *EDDM13* (Supplemental Fig. S3A). In 44 cases the protein-coding gene was entirely new to GENCODE -- i.e. it was not previously found as a lncRNA or pseudogene -- with the prior absence of most of these genes being due to their restricted expression (Supplemental Data S2), as illustrated by *C1orf232* (Supplemental Fig. S3C) and *CCDC201* (Figure 2C). Transcription of *C1orf232* appears to be limited to brain and eye tissues in human and mouse, while *CCDC201* is apparently transcribed only in female reproductive tissues and early developmental cells.

Finally, 13 protein-coding genes were identified within the UTRs of extant protein-coding loci. *H2BE1* is a particularly exciting discovery, being a novel histone protein with expression apparently limited to early development (Figure 2D). In nine 5' UTR cases, transcriptomics data indicates that the new CDS and the

previously known downstream CDS consistently share the same RNA molecule, as illustrated by the CDS identified within the *ALDOA* 5' UTR (Supplemental Fig. S3D). UTR-associated ORFs are extensively detected in vertebrate RP studies (Ingolia et al. 2011, 2009). However, it remains unclear what proportion of these are regulatory ORFs that do not produce functional peptides and instead compete with the canonical CDS for ribosome binding, or else are simply stochastic interactions (Calvo et al. 2009; Johnstone et al. 2016; Bazzini et al. 2014). In contrast, PhyloCSF detects the evolutionary signature of function at the amino acid level, so UTR ORFs identified by PhyloCSF are highly likely to be CDS that produce functional peptides. In fact, our 5' UTR-associated examples include 3 cases where protein existence has been confirmed by others through laboratory work: within the 5' UTRs of *MIEF1 (Rathore et al. 2018)*, *MKKS (Akimoto et al. 2013)* and *RAB34 (Zougman et al. 2011)*. The CDS within the *MKKS* 5' UTR produces a mitochondrial protein, while *MKKS* itself is involved in cytokinesis. This observation is a reason why GENCODE chose to represent the UTR-associated CDS as distinct protein-coding genes.

Novel protein-coding genes do not always get high SVM scores

Among the clusters containing the 1000 highest-scoring PCCRs, 81.6% led to some annotation update, whereas this was true of only 38.1% of the less-well ranked clusters we investigated. Broadly, this confirms that ranking according to SVM score is an effective way to direct manual annotators to the regions most likely to be productive.

However, not all of the protein-coding genes we identified ranked this well. In fact, during our survey of all lincRNAs, 8 protein-coding genes were identified based on clusters with a best rank greater than 3000. Analysis of these cases identified two scenarios whereby protein-coding genes may have low PhyloCSF scores. Firstly, the score can be lowered due to the loss of the gene in a sizable subclade, as this causes a gap in the underlying genome alignments. For example, *FAM240C* was apparently lost at the base of the rodent / lagomorph clade, and was identified based on a cluster with a top rank of 22,742 (Supplemental Data S2). Secondly, while multispecies alignments aim to capture '1:1' orthology between genome sequences, they can be compromised by paralogy. Thus, *ETDA* and *ETDB* were identified as primate-specific duplications of a single-copy ancestral protein-coding gene, and it was apparent that the genome alignments producing their

PhyloCSF signals were incorrect. We subsequently found evidence that certain high-ranking PCCRs were also based on alignments corrupted by paralogy, especially among the small cysteine and glycine repeat containing family members found in a cluster on Chromosome 2. In fact, local homology-based searching found three additional novel protein-coding genes within this cluster supported by PCCRs beyond the rankings studied here (*SCYGR1*, *SCYGR5* and *SCYGR7*), and also identified *ETDC* as an additional paralog to *ETDA* and *ETDB*. These genes are included in Supplemental Data S2.

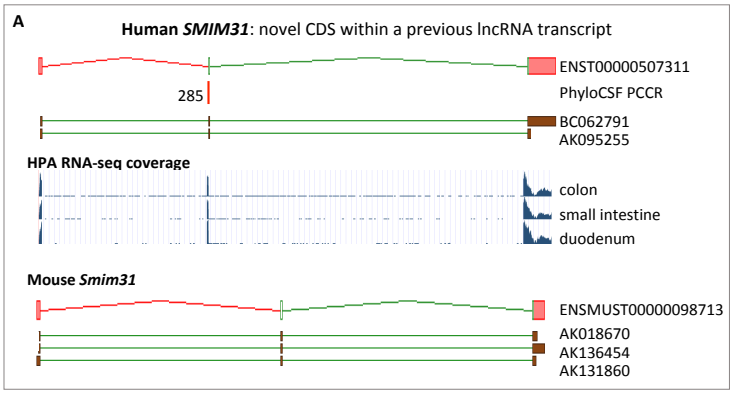
70 protein-coding genes are new discoveries

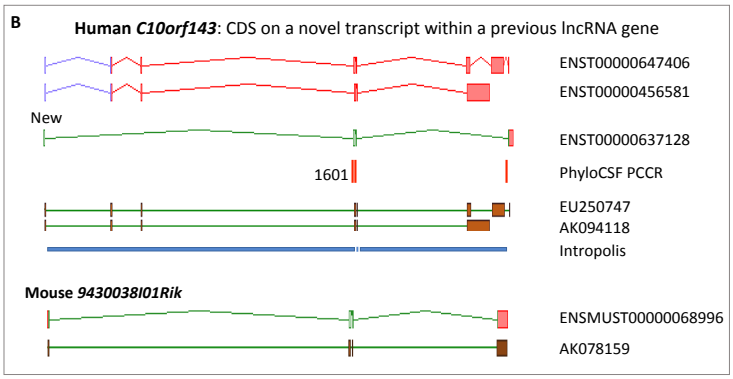
We believe that 70 of the 144 protein-coding genes added to GENCODE in this study are new discoveries, in that they were not considered to be coding loci in human before they were annotated and made publicly available by GENCODE (the sources we searched in order to come to this conclusion are listed in Supplemental Methods section "Assessing the novelty of annotations"). We found that 61 of the 144 genes existed prior to this study in either the RefSeq or UniProt catalogs, or were previously characterised as open reading frames by Mackowiak et al based on their usage of PhyloCSF (Mackowiak et al. 2015). However, it appears that 19 of these 61 genes have had their 'correct' CDS resolved for the first time as part of this study. Next, as previously noted, we found that the CDS identified within the 5' UTRs of *MKKS* and *MIEF1* had already been reported in published studies (Akimoto et al. 2013; Andreev et al. 2015; Delcourt et al. 2018), although these findings had not propagated into any annotation catalogs. Finally, we rediscovered 5 out of the 16 protein-coding loci that we recently reported (Wright et al. 2016) based on a concurrent reanalysis of large 'draft proteome' peptide datasets (Kim et al. 2014; Wilhelm et al. 2014), and all 6 loci from our analysis of testis data from the Chromosome-Centric Human Proteome Project (Weisser et al. 2016) (Supplemental Data S1 and 2).

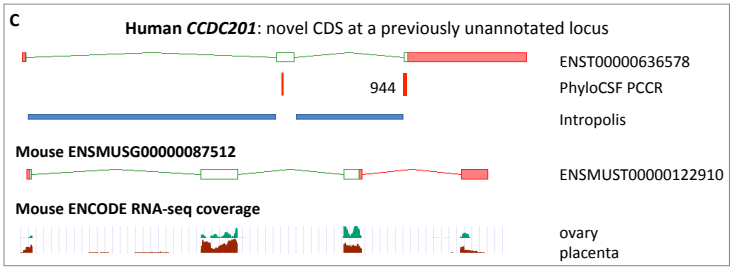
Four of the 70 novel protein-coding genes were independently reported subsequent to our identification. *SMIM38* was reported as translated based on proteomics data (Ma et al. 2016), while *SPAAR*, *STRIT1*, and *MYMX* were experimentally characterised (Matsumoto et al. 2017; Nelson et al. 2016; Bi et al. 2017). We recognise that such experimental analyses will be important to confirm the functionality of all 144 protein-coding genes. Finally, we note that *FAM240C, SMIM28,* and *AC138647.1* were annotated as protein-coding in

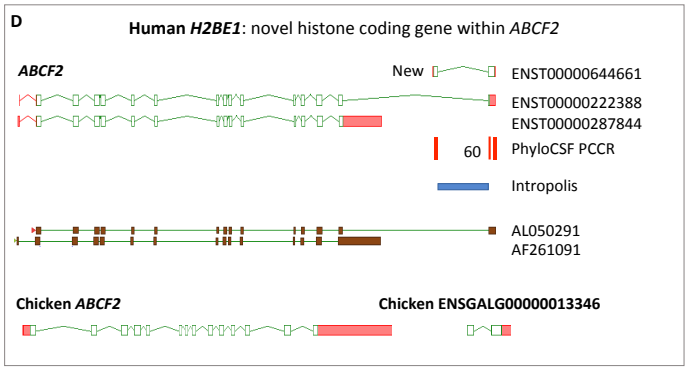
earlier versions of GENCODE and RefSeq. We included them as ‘new discoveries’ because at the time of this study annotators in both groups had subsequently reconsidered all three loci to be non-coding.

Approximately half of these 70 genes have an ortholog with some form of prior gene annotation in another species including 19 in mouse GENCODE, although with an incorrect translation in many cases (Supplemental Data S2). Following our comparative annotation, with the exception of certain paralogs discussed above, all except 15 of the total 144 human protein-coding genes now have annotated orthologs in mouse GENCODE; the missing cases apparently represent gene loss events. Furthermore, we found that at least 71 of the 144 arose prior to the mammalian radiation, and we were able to describe 15 zebrafish orthologs as part of the HAVANA / ZFIN annotation efforts (Howe et al. 2013).









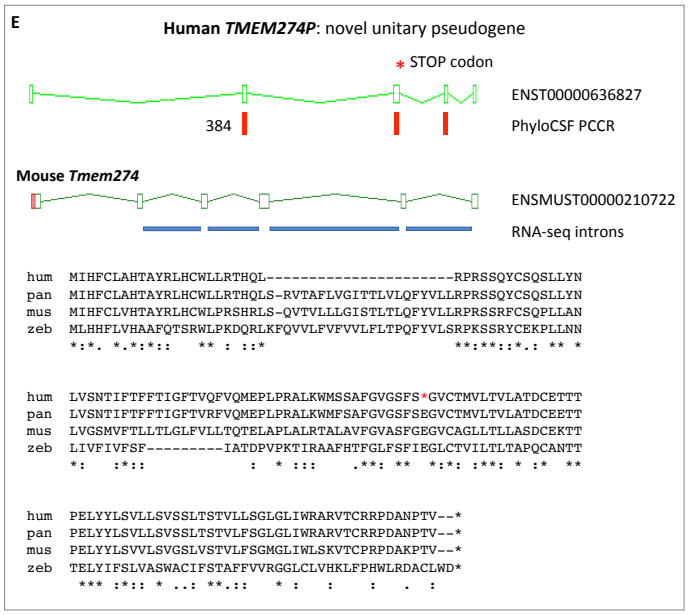


Figure 2. Novel protein-coding loci. Browser images show CDS (open green rectangles), UTRs (pink), supporting PCCRs (red), top rank (black), cDNA evidence (brown), and RNA-seq-supported introns (blue rectangles). Additional transcript models omitted for clarity. Multi-species protein alignments showing conservation of complete ORFs are in Supplemental Fig. S4. **(A)** novel coding gene *SMIM31*, previously a cDNA-supported GENCODE lincRNA, was changed to protein-coding without a change of transcript structure due to a 71-aa CDS (ENST00000507311) conserved to coelacanth. The protein-coding cDNA-supported ortholog was added to mouse GENCODE (*Smim31*). PhyloCSF does not detect coding potential in the second coding exon, but multi-species protein alignment and preponderance of 3mer indels provide evidence this exon is coding. Human Protein Atlas (HPA) RNA-seq and human and mouse FANTOM5 CAGE data demonstrate high transcription in gastrointestinal tissues. **(B)** novel coding gene *C10orf143* was previously a GENCODE lncRNA (*LINC00959*), with two cDNA-derived models (ENST00000647406 and ENST00000456581). Discovery of the 108-aa CDS required adding a transcript model (ENST00000637128), supported by Intropolis short-read data. The original lncRNA transcripts have been reannotated as nonsense-mediated decay targets (purple ORFs), based on a premature stop codon in a cassette exon. The orthologous cDNA-supported mouse locus had previously been recognised as protein-coding (*9430038I01Rik*). The gene has a broad expression profile in both species. **(C)** *CCDC201* is a novel human gene with a 187-aa CDS conserved to birds, previously missed due to lack of spliced cDNA or EST evidence. The ancestral stop codon has been lost in rodents, adding a 30-aa extension in novel mouse protein-coding gene ENSMUSG00000087512. Introns are supported by Intropolis short-read RNA-seq, limited to female reproductive tissues and certain developmental cells. Mouse ENCODE RNA-seq supports placenta and ovary expression only, and the mouse locus (in the guise of a ncRNA) had previously been identified as a target for the germ cell specific transcription factor *Figla* (Joshi et al. 2007). **(D)** *H2BE1* is a novel histone HB2 family member protein-coding gene with a 122-aa CDS (model ENST00000644661), whose first exon was identified in this study. Intropolis supports the transcript structure, with expression limited to oocytes and embryonic cells (e.g. SRR499827). Human FANTOM5 CAGE data lacks experiments from developmental stages, which may explain the absence of TSS evidence. Overlapping model ENST00000222388 had previously been annotated as an alternative transcript of *ABCF2* (ancestral CDS represented by model ENST00000287844) based on cDNA AL050291, with putative translation in the shared exon following the coding frame of *ABCF2*. PhyloCSF indicates that the 122-aa CDS is translated in a different frame, so the translation of ENST00000222388 is potentially spurious. While the 122-aa CDS is conserved to birds, the locus has apparently been lost in rodents. There is no evidence for transcriptional connectivity between the orthologous Ensembl chicken models *ABCF2* and ENSGALG00000013346 (bottom). ENST00000222388 has been reclassified as a 'readthrough' transcript, and Intropolis data indicate that such readthrough between human *ABCF2* and *H2BE1* is rare. **(E)** *TMEM274P* is a novel human unitary pseudogene, orthologous to novel mouse protein-coding gene *Tmem274*. CDS alignments to RefSeq models such as scallop LOC110448246 and trichoplax XP_002113670.1 suggest this gene may predate vertebrate evolution, although orthology is presumptive due to lack of synteny beyond coelacanth. The gene has at best weak expression data in all species examined, but all but one of the mouse splice junctions is supported by minimal ENCODE RNA-seq

data from pooled sources, and all splice sites display mammalian conservation. An alignment of human (hum) to chimp (pan), with outgroups mouse (mus) and zebrafish (zeb), shows that human has a premature stop codon that is not a known SNP in the fourth exon of the ancestral CDS (red asterisk in diagram and alignment), and has also lost the second coding exon (large gap in human sequence); both events are unique to human. The zebrafish sequence in the alignment is from XP_017212190, while the chimp translation is from the genome sequence.

PhyloCSF finds additional CDS within known protein-coding genes

While our main focus in this manuscript is on the set of protein-coding genes added to GENCODE, the majority (59%) of CDS base-pairs added to GENCODE were in fact added to 236 previously-annotated protein-coding genes. For 118 of these genes, the added CDS was a new discovery, in that it was not already present in the RefSeq or UniProt databases either. An extreme example is the *RP1* locus, linked to retinitis pigmentosa, where an additional transcript model containing 22 conserved novel coding exons was added to both the human and mouse gene sets. The bulk of these coding exons had been regarded by GENCODE and RefSeq as a separate protein-coding gene in human (LOC107984125), but our transcriptomics analysis indicates that these are not separate loci. Similarly, we were able to resolve the previously separated *BTBD8* / *KIAA1107* and *LCOR* / *C10orf12* gene pairs into single loci.

PhyloCSF identifies pseudogenic regions

We added 169 pseudogenes to human GENCODE, according to the observation of non-polymorphic truncating deletions, premature termination codons, or frame-disrupting changes in the human CDS in comparison to an inferred ancestral model (see Supplemental Methods section “Manual annotation overview”). Of these 169 pseudogenes, 149 appear to be new discoveries in that they were not included in the RefSeq catalog either. We also extended the structure of 24 previously annotated human pseudogenes, and found evidence for ‘pseudo-exons’ within 32 protein-coding genes, i.e. cases where a portion of the ancestral CDS was lost within a gene that has apparently continued to encode a functional protein. While 44 of the 169 pseudogenes are orthologs of ancestral protein-coding genes disabled in the human lineage (‘unitary pseudogenes’), the other 125 are duplicative (‘unprocessed’) pseudogenes, for which the PhyloCSF signal resulted from non-syntenic alignment to protein-coding paralogs. The inclusion of these 44 increased the number of unitary pseudogenes in human GENCODE by almost a quarter (Supplemental Data S3). To our

knowledge, 39 of these unitary pseudogenes are not found in other *human* databases, but 29 have protein-coding mouse orthologs recognised in either the GENCODE or RefSeq catalogs. We also added 6 mouse orthologs for these human loci, 2 of which are also unitary pseudogenes. One of these is the remnants of *crescent*, previously characterised in chicken (Pfeffer et al. 2002) and a recognised mammalian pseudogenisation event (Kuraku and Kuratani 2011). The other 4 cases are mouse protein-coding genes that apparently represent new discoveries. For example, *Tmem274* has an ancient CDS; conservation may even extend beyond vertebrates, yet the pseudogenisation appears unique to human (Figure 2E). Meanwhile, *Pfn5* is a novel profilin-like protein-coding gene in mouse with a novel unitary pseudogene counterpart in human, *PFN5P* (Supplemental Fig. S3E). Studying the function of these genes in those species that have retained them could help us understand how their loss has affected the evolution of our species.

In certain cases the protein-coding versus pseudogene decision was difficult, and Supplemental Data S2 highlights 9 'edge cases' for which further experimental analysis will be especially important. These include pseudo-exon cases, and also genes where the disruption to the ancestral CDS in human or mouse was relatively minor. It can be difficult to infer how the loss of CDS affects a protein-coding gene, as exemplified by *KIF25*, in which we found 8 pseudo-exons upstream of the previously annotated human CDS that are apparently not transcribed in higher primates despite showing vertebrate conservation, and yet there is published evidence that the human locus produces a functional protein; we infer this must be a truncated molecule (Decarreau et al. 2017). Finally, we also recognise that certain reclassifications of lncRNAs as protein-coding genes would seem to contradict the findings of previous studies; this includes *TUNAR (Lin et al. 2014)* and *TINCR (Kretz et al. 2013)*, both of which have ascribed non-coding functions. Their CDS are small – 48aa and 87aa respectively – and yet both are conserved beyond the mammalian order. In fact, we do not rule out the possibility that these loci function at both the protein and RNA levels.

Proteomics data validates CDS annotations

The GENCODE proteomics pipeline provided additional support for six of the protein-coding genes that we did

not already report in our parallel mass spectrometry-based protein-discovery efforts (Wright et al / Weisser et al) (Supplemental Data S2 and S4), including two of the 70 new discoveries. We also found support for CDS annotations added to 29 existing protein-coding genes (Supplemental Data S4). The GENCODE proteomics pipeline reprocesses the raw peptide spectral peptide data from Kim et al. This covers thirty tissues, allowing us to find (for example) peptide support for SMIM36 in retina to match the eye specific transcription profile, and peptide support for SMIM39 in frontal cortex to match the brain / central nervous system specific expression profile. Nonetheless, our transcriptomics analysis indicates that many of the protein-coding genes are expressed in tissues from which peptide data are not yet available. Furthermore, as a result of our work, many of our 70 new discoveries now have corresponding entries in the neXtProt protein database (Gaudet et al. 2017), which aims to provide functional support for all human proteins. neXtProt protein sequences are taken from UniProt, which targets new GENCODE CDS (such as our 70 new discoveries) for curation, and their mass spectrometry data is incorporated from PeptideAtlas (Desiere et al. 2006). We found that an additional 7 of these genes currently have peptide support according to neXtProt / PeptideAtlas criteria (Supplemental Data S2), although these are less stringent and include samples from cancer cell lines. Finally, we used the SORFS.org database of ORFs under 100aa predicted from a comprehensive set of ribosome profiling studies (Olexiouk et al. 2017) to find evidence of translation for 6 of our 50 CDS matching this size criterion (Supplemental Data S2).

PhyloCSF regions that did not support annotation

Many of the high-scoring PCCRs that did not correspond to open reading frames and are presumed to be non-coding false-positives overlapped predicted promoter and enhancer regions. In the former case, we believe this is because the high GC content and density of triples containing CpG at promoters (i.e. CpG islands) can result in codon frequency distributions similar to those of coding regions, and also because we used PhyloCSF's "fixed" option for branch lengths in the underlying phylogenetic tree, which is faster and more accurate than the "mle" option on single codons but is more sensitive to the level of sequence conservation. Thus, the elevated conservation typical of promoter and enhancer regions improves their fit to PhyloCSF's

coding model of evolution, increasing their scores. Subsequently, we have found that the ranks of CpG island-associated false-positive PCCRs can be downgraded by running PhyloCSF with the "mle" option, which scales all branch lengths by a maximum-likelihood estimated factor and is far less sensitive to sequence conservation. However, the overall consequences of using the fixed vs mle options remain to be ascertained.

There are also 26 PCCR clusters that we consider highly likely to be genuine that did not yet lead to productive annotation (category 'under investigation' in Supplemental Data S1). These include 12 protein-coding genes where PCCRs suggest that translation initiates upstream of the annotated ATG initiation codon but no alternative upstream ATG was apparent. These PCCRs could represent upstream alternative splicing events that have not yet been captured in transcript libraries, or perhaps demonstrate the usage of non-ATG initiation codons (Kearse and Wilusz 2017). We also found compelling evidence for translation events either within or overlapping with previously annotated coding exons of *POLG, PCNT, PLEKHM2, ASXL1* and *ASXL2* in alternative reading frames. These cases remain difficult to interpret.

Variation evidence supports recent protein-coding selection

Evidence from human nucleotide variation indicates that purifying selection at the amino acid level has continued to act on the newly added CDS, in aggregate, in the human population, as well as on the subset consisting of just the 70 novel coding genes. In particular, we found that variants in new CDS show a strong bias to be synonymous if translated in the predicted reading frame (Supplemental Fig. S5A) and derived allele frequencies for nonsense variants are significantly lower than those of missense variants, which are in turn significantly lower than those of synonymous variants (Supplemental Fig. S5B and S5C).

New annotations reveal 118 protein-altering GWAS variants

An important application of gene annotation is to connect variants associated with disease via family studies or genome-wide association studies (GWAS) to changes in proteins. We searched the UK Biobank GWAS summary statistics and EBI GWAS catalogs for single-nucleotide variants (SNVs) within our new coding annotations that had previously been found to have genome-wide significant association with diseases or other

traits. We identified 118 variants that affect the protein sequence, including one splice-disrupting variant, two nonsense variants, and 115 missense variants (Figure 3A, and Supplemental Data S5). Note that some variants might already have been classified as protein-coding at the time of the GWAS because we have been releasing the updated annotations described here in GENCODE versions 24 through 28, and because some of the variants lie in regions previously classified as coding by RefSeq.

Recognition of these variants as protein-disrupting may prove crucial in understanding the mechanism by which they affect disease. For example, a 2013 GWAS study found rs11145465 to be associated with refractive error and myopia, and had classified it as non-coding (Verhoeven et al. 2013; Tedja et al. 2018). However, we now recognize that it is a missense mutation in a previously unidentified protein-coding transcript of *TJP2* (Figure 3B and 3C). This gene has been implicated in a wide range of diseases, including cancer, hearing loss, liver disease, and immune disorders (González-Mariscal et al. 2017). The novel coding transcript is expressed only in eye tissues (Figure 3B), while the GENCODE transcripts described prior to this work show negligible expression in eye.

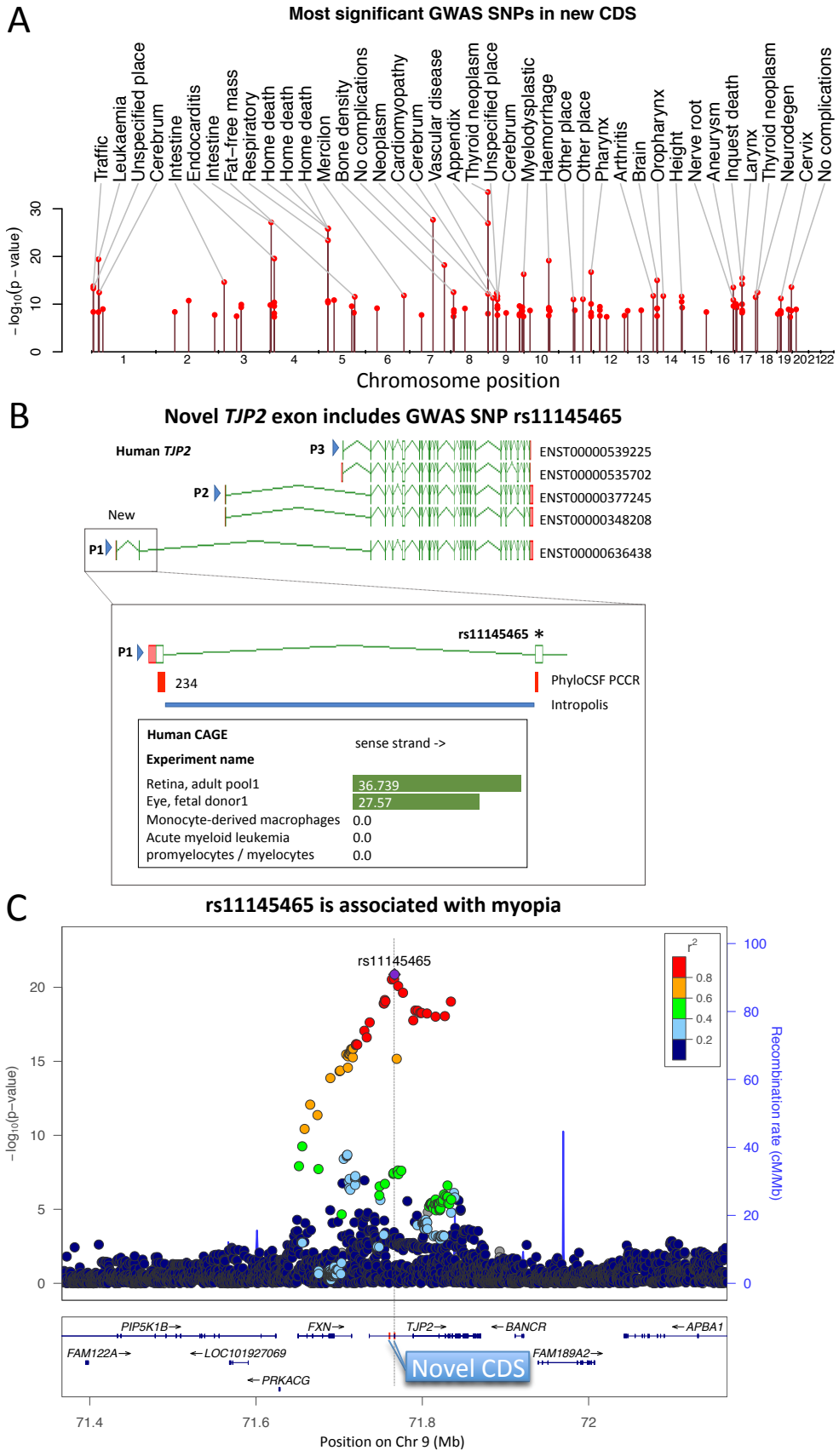


Figure 3. Protein-altering disease variants. (A) Chromosomal positions and strength of association for the 118 SNVs in newly

annotated CDS that were previously found to be significantly associated with diseases or other traits, with the trait abbreviation from Supplemental Data S5 listed for the 40 most significant associations. **(B)** Novel coding sequence added to human *TJP2* locus includes eye-disease associated variant. Previous GENCODE annotation represented by models ENST00000539225, ENST00000535702, ENST00000377245, and ENST00000348208. Additional transcriptional complexity omitted for clarity. PhyloCSF PCCRs indicated the presence of two additional coding exons (dotted box and inset) that led to annotation of novel coding transcript model ENST00000636438, which lacks cDNA or EST support but whose intron is confidently supported by short read data in Intropolis (blue rectangle) mostly from a retinal study (Farkas et al. 2013), and whose TSS (P1) is supported by FANTOM5 CAGE data, limited to retina and eye (data from ZENBU browser, precisely redrawn for clarity; scores represent sequence read counts, with zeros for next three experiments included for comparison). In contrast, TSSs P2 and P3 have negligible CAGE support for eye expression, with profiles dominated by monocyte and central nervous system expression. FANTOM5 CAGE also demonstrates eye-specific expression for an equivalent mouse model added as part of this study, also supported by eye-experiment ESTs (e.g. BU505208.1). The second coding exon added to human GENCODE contains GWAS variant rs11145465, identified in a study of refractive error and myopia with a p-value of 7×10^{-9} (Verhoeven et al. 2013). In that study the variant had been interpreted as non-coding based on RefSeq annotation, but it can now be reclassified as a missense mutation of an amino acid that is perfectly conserved in the mammal and avian clades. **(C)** Regional association plot for eye disease. All SNPs in an 800 kb window with their strength of association with refractive error and myopia in a more recent study (Tedja et al. 2018) show that rs11145465 has the strongest association. The positions of the novel coding exons of ENST00000636438 have been added in red.

Novel CDS in other species

We have created PhyloCSF browser tracks and PCCR lists for chicken, fly (*D. melanogaster*), worm (*C. elegans*), and mosquito (*A. gambiae*). A cursory examination of top-ranked PCCRs in these lists suggests that implementing our complete workflow could prove useful for discovering hundreds of novel CDS and pseudogenes in those genomes. We describe a few examples from these species to indicate the potential value of such an effort (Figure 4, Supplemental Fig. S6). These examples were identified from the alignments using the PhyloCSF signal, splice site predictions, and conservation of start codons, stop codons, splice sites, and reading frame, without reference to transcriptional data, so we cannot rule out that some of these are pseudogenes or that the true transcript models deviate from our predicted models.

Many of the best-ranked PCCRs in each of these species suggest novel pseudogenes (Supplemental Fig. S7), which is particularly notable since *D. melanogaster* and *A. gambiae* have a paucity of known pseudogenes.

We have also created PhyloCSF browser tracks and a PCCR list for the mouse genome. Our analysis of the human PCCR list has already resulted in many novel annotations in the mouse genome, and the mouse PCCR list could prove to be valuable for identifying novel annotations in regions of the mouse genome that have been lost in human. GENCODE plans to implement a full survey of mouse PCCRs.

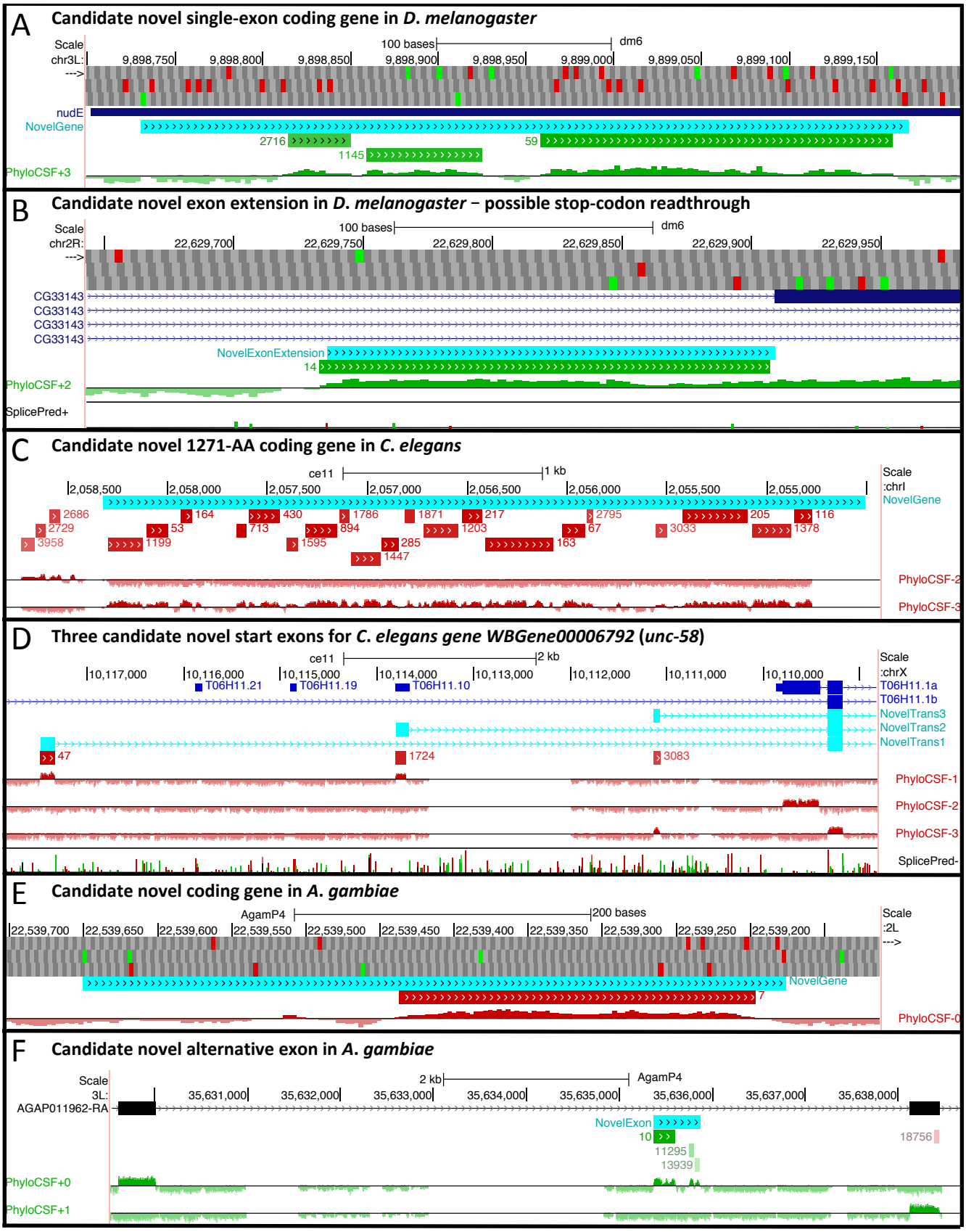


Figure 4. Potential novel CDS in other species. Browser images show proposed novel CDS (cyan) suggested by PCCRs (green/red for

+/- strand; rank next to region); smoothed PhyloCSF browser tracks; splice site predictions where useful (green donor, red acceptor, height indicating prediction strength); and ATG (green) and stop (red) codons. Supplemental Fig. S6 has color-coded alignments for each example. **(A)** A cluster of three PCCRs in the 5'-UTR of *D. melanogaster nudE* suggest there is a single-exon novel protein-coding gene or an additional *nudE* cistron with ORF at positions 9898731-9899168. Although there is no PhyloCSF signal in the first 28 codons, the high frame conservation despite several indels provides ample evidence of purifying selection for protein-coding function. **(B)** A PCCR just 5' of an exon of *D. melanogaster* transcript F of *CG33143* suggests that there is a novel coding transcript including an exon 173 nt longer than the annotated exon. This exon includes an in-frame TAG stop codon, suggesting translational stop codon readthrough. We have previously estimated that roughly 6% of *D. melanogaster* genes undergo stop codon readthrough (Jungreis et al. 2016). The stop codon is perfectly conserved, and is followed immediately by a cytosine residue, both of which are known correlates of readthrough. **(C)** A large cluster of PCCRs on the '-' strand of *C. elegans* Chromosome I suggest there is a 1271 amino acid single-exon gene with ORF at positions 2054512-2058327. There is no alignment for a few codons on each end of the PhyloCSF signal, so to construct the putative ORF we have extended the region 5' to the nearest ATG and 3' to the nearest stop codon. **(D)** Three PCCRs within an intron of *C. elegans* gene *WBGene00006792* (*unc-58*) shown on the '-' strand of Chromosome X suggest alternative start exons for that gene. The coding region of each of these putative exons begins with a perfectly conserved ATG and ends at a perfectly conserved GT having high splice-prediction score. All three end with a 1-nt partial codon, which allows them to splice to the next exon of transcript *T06H11.1b* while preserving the reading frame. **(E)** A PCCR in *A. gambiae* suggests that 22539177-22539650 on the '-' strand of Chromosome 2L is protein-coding, forming either a novel gene or the first coding exon of the previously incompletely annotated gene *AGAP005849*. Subsequent curation confirmed the latter. Frame conservation provides strong evidence of coding function in the early portion of the putative transcript where the PhyloCSF signal is weak. **(F)** A cluster of three PCCRs in an intron of *A. gambiae* gene *AGAP011962* suggests an additional coding exon at positions 35635374-35635874 of Chromosome 3L, confirmed through subsequent curation to be part of a previously missed alternative transcript.

Discussion

We have presented the first whole-genome PhyloCSF resources for the human, mouse, chicken, *D. melanogaster*, *C. elegans*, and *A. gambiae* genomes, and demonstrated the utility of the human resource and our workflow in finding hundreds of high-confidence novel CDS and pseudogenes within a genome that had already been intensely scrutinised. This analysis has several advantages over previous studies having similar goals. We have achieved high sensitivity by using PhyloCSF on the whole genome to find novel CDS that either fully or partially lie outside existing transcript catalogs. We have achieved high specificity by computationally filtering out identified sources of PhyloCSF false positives (including antisense signals, known pseudogenes, and low alignment branch length) and by manual examination of every candidate, retaining only

those that were supported by both transcriptional and comparative data. Our integrated annotation workflow has allowed us to achieve more reliable and comprehensive results than could be achieved by either fully automatic or manual methods acting separately. In particular, while it is apparent that most PhyloCSF signals remaining after computational filtering are false positives, we have demonstrated that our ranking algorithm is a highly efficient approach to isolate true positives. Meanwhile, our preview of the top-ranked PCCRs for *D. melanogaster*, *C. elegans*, and *A. gambiae*, suggests that the deployment of a similar manual annotation-centred workflow guided by PCCRs could be a key step in completing the catalogs of conserved protein-coding genes for these species. A similar effort for the chicken genome is already underway (Vignal and Eory 2019).

Our whole-genome resources are already helping researchers investigating novel transcript sets to distinguish those with protein coding potential without having to install and run PhyloCSF (Perry et al. 2018; Makarewich et al. 2018; Huang et al. 2019a; McCorkindale et al. 2019; Kang et al. 2019; Huang et al. 2019b; van Heesch et al. 2019; Lin et al. 2019; Wang et al. 2019; Vignal and Eory 2019). Transcripts that do not overlap a PCCR or any annotated coding gene are unlikely to have conserved protein-coding function, while transcripts that overlap top-ranked PCCRs are the best candidates for translational potential. We recommend that no gene be considered protein-coding based on PCCR overlap alone; rather, an overlap is the starting point for constructing a potential CDS. In this regard, CodAlignView is a valuable tool for exploring multi-species alignments for signals of coding potential (I Jungreis, MF Lin, CS Chan, M Kellis 2016), while the PhyloCSF browser tracks may be especially useful for examining PCCRs in the context of transcriptomics data. Indeed, we stress the value of an integrated transcriptomics analysis: many of our novel protein-coding genes previously existed as non-coding models that were inaccurately or incompletely described. Conversely, short-read transcriptomics data is not in itself sufficient to identify protein-coding genes with high confidence, and even when the locus-level identification of coding potential is correct, we have found that the actual CDS predicted is commonly inaccurate. A confounding factor here is the existence of extensive alternative transcription within protein-coding genes. The proportion of this complexity that represents stochastic ‘noise’ remains to be ascertained (Wan and Larson 2018), and while it could be that only a minority of transcript isoforms are translated into mature proteins, this remains highly debated (Mudge et al. 2013; Blencowe 2017;

Tress et al. 2017). In fact, we believe that PhyloCSF and our PCCR list have enormous potential both to discover additional *novel* protein-coding alternatively spliced transcripts in known genes and to distinguish those *known* transcripts that generate conserved protein products from those that do not (see Supplemental Fig. S2 for one example); our present work has only scratched the surface in this regard.

We recognise that not all novel protein-coding genes can be found by our workflow, and a brief survey of the 7% of previously annotated protein-coding genes that do not overlap a PhyloCSF Region found that many are recent paralogs lacking sufficient evolutionary history to produce a signal. We also reiterate that the fidelity of PhyloCSF is linked to the accuracy of the underlying genome alignments, and while 'serendipitous' PhyloCSF signals resulting from paralogous alignments were of value to this study, we caution that this behaviour cannot be relied upon. Furthermore, PhyloCSF confirms the *provenance* of a genomic region to be a protein-coding sequence, not whether it remains protein-coding in a particular species. An examination of variation burden indicates that our novel CDS, in aggregate, have continued to be subject to purifying selection at the amino acid level in the human population, but does not have adequate statistical power to show that each individual gene is still producing a functional protein. Demonstrating that candidate CDS are not pseudogenic regions remains a judgement call until true confidence in the coding potential of a given gene can be obtained in the laboratory, ideally via single gene studies. In the meantime, confidence in CDS annotation can be gained through the incorporation of orthologonal datasets. While others have sought to discover or validate prospective CDS using RP datasets (Mackowiak et al. 2015; Bazzini et al. 2014), our own experience is that these remain difficult to interpret in a biological context, certainly when the goal is to create 'high confidence' reference annotation (Mudge and Harrow 2016). However, we do not doubt the potential usefulness of RP data; indeed, we have shown that at least some ORFs initially suggested by RP are likely to be true proteins. Meanwhile, we and others have previously found novel CDS using mass spectrometry (Wright et al. 2016; Weisser et al. 2016; Slavoff et al. 2012; Kim et al. 2014; Wilhelm et al. 2014). Our work here provides further demonstration of the value of this approach, and it has the potential to be extended in the future via 'targeted' proteogenomics, e.g. using synthetic peptides. Furthermore, GENCODE annotation is utilised by several projects seeking to provide catalogs of protein function, including neXtProt (via UniProt) and the Human

Protein Atlas (Gaudet et al. 2017; Thul et al. 2017); we anticipate that such resources will in due course provide valuable insights into these genes based on experimental data.

Ultimately, experimental characterization of novel CDS is vital, as gene annotation supports virtually all attempts to understand the mechanisms of human disease. Our preliminary work has shown that CDS discovery can shed light on disease associated loci, and we hope that our reclassification of many disease-associated variants as protein-altering will lead to further investigation of their mechanism of action, and eventually to clinically beneficial consequences.

Methods

PhyloCSF

PhyloCSF software and parameters were obtained from GitHub (Lin 2012). PhyloCSF was run using the "fixed" option on every codon in each frame on both strands of each chromosome and scaffold in the primary genome assembly. We used the “fixed” option because it is faster and, on single codons, more accurate than the “mle” option (though the “mle” option is more accurate on longer regions). Alignments used are specified in Supplemental Methods section “PhyloCSF and browser tracks”.The scores were smoothed using a Hidden Markov Model (HMM) having 4 states, one representing coding regions and three representing non-coding regions. The emission of each codon is its PhyloCSF score. The ratio of the emissions probabilities for the coding and non-coding models are computed from the PhyloCSF score, since it represents the log-likelihood ratio of the alignment under the coding and non-coding models. The three non-coding states have identical emissions probabilities but different transition probabilities (they can only transition to coding) to better capture the multimodal distribution of gaps between same-frame coding exons. Intuitively, the emissions probabilities of the three states can be thought of as roughly capturing the gaps between a coding exon and the next coding exon on the same strand in the same genomic frame if they are consecutive exons in the same gene, non-consecutive exons in the same gene, or in different genes. However, the algorithm does not actually use this information and instead uses Expectation Maximization to find the best approximation of this gap distribution as a mixture model of three exponential distributions.

The HMM defines a probability that each codon is protein coding, based on the PhyloCSF scores of that codon and nearby codons on the same strand in the same frame, without taking into account start codons, stop codons, or potential splice sites. The smoothed PhyloCSF browser tracks show the log-odds that each codon is in the coding state according to the HMM. PhyloCSF Regions are defined as the intervals in which the most likely path through the HMM is in the coding state.

PhyloCSF Candidate Coding Regions relative to a particular set of gene annotations were created as follows. All PhyloCSF Regions were compared to CDS and pseudogene annotations from the specified gene set, and those contained in annotated CDS regions in the same or antisense frame, or in annotated pseudogene regions in any frame or strand, were excluded. If only part of a region was contained in the annotated CDS or pseudogene, the region was trimmed to the unannotated portion. Regions shorter than nine codons were excluded.

We trained a Support Vector Machine (SVM) to distinguish PhyloCSF Regions that are more likely to be a novel coding region than antisense to a novel coding region, using as features the average PhyloCSF score per codon, the per-codon difference between the PhyloCSF score and the score in the antisense frame, and the length of the region. The length is relevant because antisense “ghost” regions tend to be shorter than true protein-coding regions. We trained the SVM using 10,000 randomly selected PhyloCSF Regions overlapping annotated CDS in the same frame as positive examples, and an equal number overlapping annotated CDS in the antisense frame as negative examples. We then excluded from the PhyloCSF Candidate Coding Regions set any regions that our antisense SVM scored below 0.3, a threshold chosen so as to keep almost all of our positive training examples (99%), while excluding most of our negative training examples (94%) (Supplemental Fig. S1A).

We trained an SVM to distinguish the PhyloCSF Candidate Coding Regions most likely to be protein-coding (Supplemental Fig. S1B) using four features, namely the average PhyloCSF score per codon, the per-codon difference between the PhyloCSF score and the score in the antisense frame, the length of the region, and the branch length of the species in the local alignment of the region. These features were chosen because true

protein-coding regions tend to have higher PhyloCSF score, greater difference between PhyloCSF scores on the two strands, greater length, and greater alignment branch length than false positives (Supplemental Fig. S1C). We trained the SVM using 10,000 randomly selected PhyloCSF Regions overlapping annotated CDS in the same frame as positive examples, and an equal number of regions that do not overlap any CDS annotations in the same frame or antisense frame or any pseudogene annotations in any frame on either strand as negative examples. We then ranked the PhyloCSF Candidate Coding Regions using the scores from this SVM. Both SVMs were trained using the "R" language svm function from the cran "e1071" package with default parameters (R Core Team 2017). To test whether the SVM performance statistics reported in Figure 1B were influenced by overfitting, we redid those calculations excluding the 10,000 training regions. There were only 60 known coding genes (0.3%) that overlapped at least one of the training regions but did not overlap any other PhyloCSF regions, and excluding training regions when scoring known coding genes had a negligible effect on the results.

For each assembly, the annotation version used to compute PhyloCSF Candidate Coding Regions and whether the PhyloCSF scores used by the SVMs were the original "fixed" scores or scores recomputed using the "mle" option are reported in the Supplemental Methods section "PhyloCSF and browser tracks". The counts reported in the Results section are from the hg38 human assembly using GENCODE v23.

Annotation

To aid manual annotation, PCCRs were clustered based on 10 kb sliding windows; this was because novel coding regions are often found as multiple exons of the same gene. All annotation was produced manually according to the guidelines developed by the HAVANA group for the GENCODE / ENCODE projects (Harrow et al. 2012). A detailed annotation workflow is provided in Supplemental Methods section "Manual annotation overview". Briefly, in addition to sequences from the GenBank repository, annotation was also supported by SLR-seq (Tilgner et al. 2015), capture-seq Pacific Biosciences (PacBio) data (Lagarde et al. 2017) and a vast collection of publicly available short-read RNA-seq datasets as processed by the Intropolis project (Nellore et al. 2016). Transcription start sites were annotated based on Cap Analysis of Gene Expression (CAGE) libraries generated by FANTOM (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. 2014), while

polyadenylation sites were identified using PolyA-seq data (Derti et al. 2012). Insights into tissue specificity were chiefly gained from the CAGE and Intropolis datasets. Comparative analysis was also performed on non-GENCODE genomes and transcriptomes. Potential orthologs were initially sought using BLASTP (Altschul et al. 1990) on the vertebrate Protein database at NCBI (Coordinators 2017), and examined in their genomic context using the UCSC (Casper et al. 2017) and Ensembl (Zerbino et al. 2018) Genome Browsers. Orthologs were also identified based on manual cross-species genome alignments. The accuracy of these provisional models were examined using whatever experimental data were available for that species. Multi-species protein alignments were created using Clustal Omega (Sievers et al. 2011). Additional scrutiny was applied to annotations that overlap transposons (Supplemental Methods section "Overlap of novel annotations with transposon sequences"). Transposon overlaps were found by comparing novel CDS to RepeatMasker regions (Smit et al. 2013) obtained from the UCSC Genome Browser (Casper et al. 2017), excluding regions of repeat class `Low_complexity` and `Simple_repeat`.

Proteomics analysis

The raw data published by Kim et al. (Kim et al. 2014) covering 30 tissues in 85 higher-energy collision dissociation (HCD) mass spectrometry experiments were downloaded from PRIDE (PXD000561, PXD002967) and converted to mzML format. These mzML spectra were searched using multiple search engines in a high confidence OpenMS workflow as described by Wright et al. (Wright et al. 2016) and Weisser et al. (Weisser et al. 2016). The spectra were searched against a sequence database composed of all GENCODE v27 CDS transcripts combined with PhyloCSF sequences; an equally sized decoy database generated using DecoyPYrat (Wright et al. 2016) was concatenated and used to control FDR. Peptides were filtered to a posterior error probability of less than 0.01 and required to be significant in multiple search engines; a minimum and maximum length of 6 and 30 amino acids respectively was set; a maximum of 2 missed cleavages were allowed, and certain modifications such as deamidation were filtered out. The final list of peptides were then manually inspected and curated against the PhyloCSF sequences and CDS.

Human variation

Germline single-nucleotide variants in the CDS portion of a newly annotated coding gene or of a previously

annotated coding gene containing new CDS were obtained from Ensembl release 91. For analysis of purifying selection, only variants having the "MAF" and "MA" tags in the Ensembl VCF file were used. Variants associated with disease were found by searching for SNVs in new CDS or adjacent splice sites having p-value less than 5×10^{-8} in the EBI GWAS catalog and autosomes in the UK Biobank GWAS summary statistics for 2419 traits provided by the Neale lab. Additional details are in Supplemental Methods section "Human variation".

Data Access

The PhyloCSF tracks for the hg38 human assembly generated using the 58-mammals alignments, and the tracks for the hg19 (human), mm10 (mouse), galGal4 (chicken), dm6 (fly), and ce11 (worm) assemblies may be viewed in the UCSC (http://genome.ucsc.edu) or Ensembl (http://www.ensembl.org) Genome Browsers by loading the "PhyloCSF" public track hub. The URL for this hub is https://data.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub/hub.txt.

The tracks for hg38 using the 100-vertebrates alignment are available at https://data.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub_hg38_100/hub.txt.

The tracks for hg38 generated by lifting over scores generated in hg19 using the 29-mammals alignment are available at https://data.broadinstitute.org/compbio1/PhyloCSFtracks/trackHub_hg38_29/hub.txt.

An assembly hub for viewing PhyloCSF tracks for the AgamP4 mosquito assembly in the UCSC or VectorBase Genome Browsers is available at https://data.broadinstitute.org/compbio1/AssemblyHubs/AgamP4/hub.txt.

A repository has been created for spreadsheets containing the list of PhyloCSF Candidate Coding Regions for each species and annotation set, with pertinent information for each PCCR such as the PhyloCSF and SVM scores. It is our intention to add PCCR lists for additional species or newer annotations sets as they become available. The repository includes a README file that describes the spreadsheet fields. The PCCRs that were the primary focus of this study are those in PCCRs.H_sapiens.hg38.GENCODE23.txt.gz. The repository is available at https://data.broadinstitute.org/compbio1/PhyloCSF_Candidate_Coding_Regions.

All human annotations described in this study are included in GENCODE (www.gencodegenes.org) release

v29, though most also appeared in earlier releases beginning with v24. All mouse protein-coding genes are in release M19 or earlier. All annotations were first publicly available via the GENCODE Annotation Updates trackhub, which is updated every 24 hours (http://ftp.ebi.ac.uk/pub/databases/gencode/update_trackhub/hub.txt).

Scripts implementing the HMM, SVM, and splice-prediction algorithms are in Supplemental Code S1 and also in the public GitHub repository, https://github.com/iljungr/PhyloCSFCandidateCodingRegions.git. Also included are a README file containing step by step instructions for using these scripts, and a script that works through examples.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U41HG007234. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional support was provided by the Wellcome Trust grant number WT108749/Z/15/Z, NIH grant R01 HG004037 (IJ), the Swiss National Science Foundation grant PP00P3_170664 (RMW), National Human Genome Research Institute grant U24HG003345 (EB, RS and ST), Wellcome Trust grant number 208349/Z/17/Z (EB, RS and ST), and the European Molecular Biology Laboratory. We would like to thank Jade Vinson for help with splice prediction software, Lel Eory for help with chicken data, and Mike Lin, Clara Chan, Mark Diekhans, John Rinn, and Jennifer Harrow for helpful input.

References

Abascal F, Juan D, Jungreis I, Martinez L, Rigau M, Rodriguez JM, Vazquez J, Tress ML. 2018. Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res*. http://dx.doi.org/10.1093/nar/gky587.

Akimoto C, Sakashita E, Kasashima K, Kuroiwa K, Tominaga K, Hamamoto T, Endo H. 2013. Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochim Biophys Acta* **1830**: 2728–2738.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN, Baranov PV. 2015. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife* **4**: e03971.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981–993.

Bi P, Ramirez-Martinez A, Li H, Cannavino J, McAnally JR, Shelton JM, Sánchez-Ortiz E, Bassel-Duby R, Olson EN. 2017. Control of muscle formation by the fusogenic micropeptide myomixer. *Science* **356**: 323–327.

Blencowe BJ. 2017. The relationship between alternative splicing and proteomic complexity. *cell.com*. http://www.cell.com/trends/biochemical-sciences/fulltext/S0968-0004(17)30070-1.

Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106**: 7507–7512.

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. 2017. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res*. http://dx.doi.org/10.1093/nar/gkx1020.

Coordinators NR. 2017. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **45**: D12–D17.

Decarreau J, Wagenbach M, Lynch E, Halpern AR, Vaughan JC, Kollman J, Wordeman L. 2017. The tetrameric kinesin Kif25 suppresses pre-mitotic centrosome separation to establish proper spindle orientation. *Nat Cell Biol* **19**: 384–390.

Delcourt V, Brunelle M, Roy AV, Jacques J-F, Salzet M, Fournier I, Roucou X. 2018. The protein coded by a short open reading frame, not by the annotated coding sequence is the main gene product of the dual-coding gene MIEF1. *Mol Cell Proteomics*. http://dx.doi.org/10.1074/mcp.RA118.000593.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173–1183.

Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. 2006. The PeptideAtlas project. *Nucleic Acids Res* **34**: D655–8.

Farkas MH, Grant GR, White JA, Sousa ME, Consugar MB, Pierce EA. 2013. Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC Genomics* **14**: 486. http://dx.doi.org/10.1186/1471-2164-14-486.

Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, Raychowdhury R, Hacohen N, Carr SA, Ingolia NT, et al. 2015. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell* **60**: 816–827.

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2018. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky955/5144133.

Gaudet P, Michel P-A, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, Duek PD, Gateau A, Gleizes A, Hinard V, et al. 2017. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res* **45**: D177–D182.

González-Mariscal L, Miranda J, Raya-Sandino A, Domínguez-Calderón A, Cuellar-Perez F. 2017. ZO-2, a tight junction protein involved in gene expression, proliferation, apoptosis, and cell size regulation. *Ann N Y Acad Sci* **1397**: 35–53.

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**: 240–251.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**: 498–503.

Huang H-H, Chen F-Y, Chou W-C, Hou H-A, Ko B-S, Lin C-T, Tang J-L, Li C-C, Yao M, Tsay W, et al. 2019a. Long non-coding RNA HOXB-AS3 promotes myeloid cell proliferation and its higher expression is an adverse prognostic marker in patients with acute myeloid leukemia and myelodysplastic syndrome. *BMC Cancer* **19**. http://dx.doi.org/10.1186/s12885-019-5822-y.

Huang J, Li J, Li Y, Lu Z, Che Y, Mao S, Lei Y, Zang R, Zheng S, Liu C, et al. 2019b. Interferon-inducible lncRNA IRF1-AS represses esophageal squamous cell carcinoma by promoting interferon response. *Cancer Lett* **459**: 86–99.

I Jungreis, MF Lin, CS Chan, M Kellis. 2016. CodAlignView. *CodAlignView: The Codon Alignment Viewer*. http://data.broadinstitute.org/compbio1/cav.php (Accessed April 30, 2016).

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218–223.

Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789–802.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Jackson R, Standart N. 2015. The awesome power of ribosome profiling. *RNA* **21**: 652–654.

Johnstone TG, Bazzini AA, Giraldez AJ. 2016. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J* **35**: 706–723.

Joshi S, Davies H, Sims LP, Levy SE, Dean J. 2007. Ovarian gene expression in the absence of FIGLA, an oocyte-specific transcription factor. *BMC Dev Biol* **7**: 67.

Jungreis I, Chan CS, Waterhouse RM, Fields G, Lin MF, Kellis M. 2016. Evolutionary Dynamics of Abundant Stop Codon Readthrough. *Mol Biol Evol* **33**: 3108–3132.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res* **21**: 2096–2113.

Kang C-M, Bai H-L, Li X-H, Huang R-Y, Zhao J-J, Dai X-Y, Zheng L, Qiu Y-R, Hu Y-W, Wang Q. 2019. The binding of lncRNA RP11-732M18.3 with 14-3-3 β/α accelerates p21 degradation and promotes glioma growth. *EBioMedicine* **45**: 58–69.

Kearse MG, Wilusz JE. 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev*

31: 1717–1731.

Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. 2014. A draft map of the human proteome. *Nature* **509**: 575–581.

Kretz M, Siprashvili Z, Chu C, Webster DE, Zehnder A, Qu K, Lee CS, Flockhart RJ, Groff AF, Chow J, et al. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* **493**: 231–235.

Kuraku S, Kuratani S. 2011. Genome-wide detection of gene extinction in early mammalian evolution. *Genome Biol Evol* **3**: 1449–1462.

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet*. http://dx.doi.org/10.1038/ng.3988.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**: 476–482.

Lin H, Jiang M, Liu L, Yang Z, Ma Z, Liu S, Ma Y, Zhang L, Cao X. 2019. The long noncoding RNA Lnczc3h7a promotes a TRIM25-mediated RIG-I antiviral innate immune response. *Nat Immunol* **20**: 812–823.

Lin MF. 2012. PhyloCSF GitHub repository. *PhyloCSF GitHub repository*. https://github.com/mlin/PhyloCSF (Accessed August 28, 2014).

Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE, et al. 2007. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res* **17**: 1823–1836.

Lin MF, Deoras AN, Rasmussen MD, Kellis M. 2008. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput Biol* **4**: e1000067.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: i275–82.

Lin N, Chang K-Y, Li Z, Gates K, Rana ZA, Dang J, Zhang D, Han T, Yang C-S, Cunningham TJ, et al. 2014. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol Cell* **53**: 1005–1019.

Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, et al. 2015. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol* **16**: 179.

Ma J, Diedrich JK, Jungreis I, Donaldson C, Vaughan J, Kellis M, Yates JR 3rd, Saghatelian A. 2016. Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal Chem* **88**: 3967–3975.

Makarewich CA, Baskin KK, Munir AZ, Bezprozvannaya S, Sharma G, Khemtong C, Shah AM, McAnally JR, Malloy CR, Szweda LI, et al. 2018. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. *Cell Rep* **23**: 3701–3709.

Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, Saghatelian A, Nakayama KI, Clohessy JG, Pandolfi PP. 2017. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**: 228–232.

McCorkindale AL, Wahle P, Werner S, Jungreis I, Menzel P, Shukla CJ, Abreu RLP, Irizarry RA, Meyer IM, Kellis M, et al. 2019. A gene expression atlas of embryonic neurogenesis in Drosophila reveals complex spatiotemporal regulation of lncRNAs. *Development* **146**. http://dx.doi.org/10.1242/dev.175265.

Mudge JM, Frankish A, Harrow J. 2013. Functional transcriptomics in the post-ENCODE era. *Genome Res* **23**: 1961–1973.

Mudge JM, Harrow J. 2016. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **17**: 758–772.

Nellore A, Jaffe AE, Fortin J-P, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips RA III, Karbhari N, Hansen KD, Langmead B, et al. 2016. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol* **17**: 266.

Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, Reese AL, McAnally JR, Chen X, Kavalali ET, et al. 2016. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**: 271–275.

Nesvizhskii AI. 2014. Proteogenomics: concepts, applications and computational strategies. *nature.com*. https://www.nature.com/articles/nmeth.3144.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–45.

Olexiouk V, Van Criekinge W, Menschaert G. 2017. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res*. http://dx.doi.org/10.1093/nar/gkx1130.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. 2012. The GENCODE pseudogene resource. *Genome Biol* **13**: R51.

Perry RB-T, Hezroni H, Goldrich MJ, Ulitsky I. 2018. Regulation of Neuroregeneration by Long Noncoding RNAs. *Mol Cell* **72**: 553–567.e5.

Pfeffer PL, De Robertis - International Journal of … EM, 2002. 2002. Crescent, a novel chick gene encoding a Frizzled-like cysteine-rich domain, is expressed in anterior regions during early embryogenesis. *ijdb.ehu.es*. http://www.ijdb.ehu.es/web/paper/9240561.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, Stephens M, Gilad Y, Pritchard JK. 2016. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5**. http://dx.doi.org/10.7554/eLife.13328.

Rathore A, Chu Q, Tan D, Martinez TF, Donaldson CJ, Diedrich JK, Yates JR 3rd, Saghatelian A. 2018. MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry* **57**: 5564–5575.

R Core Team. 2017. R: A language and environment for statistical computing. https://www.R-project.org.

Schlötterer C. 2015. Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet* **31**: 215–219.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high□quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

Sisu C, Pei B, Leng J, Frankish A, Zhang Y, Balasubramanian S, Harte R, Wang D, Rutenberg-Schoenberg M, Clark W, et al. 2014. Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A* **111**: 13361–13366.

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. 2012. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat Chem Biol* **9**: nchembio.1120.

Smit AFA, Hubley R, Green P. 2013. 2013--2015. RepeatMasker Open-4.0.

Tedja MS, Wojciechowski R, Hysi PG, Eriksson N, Furlotte NA, Verhoeven VJM, Iglesias AI, Meester-Smoor MA, Tompson SW, Fan Q, et al. 2018. Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nat Genet* **50**: 834–848.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**: 1787–1797.

The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**: D506–D515. http://dx.doi.org/10.1093/nar/gky1049.

Thul PJ, Åkesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, et al. 2017. A subcellular map of the human proteome. *Science* **356**. http://dx.doi.org/10.1126/science.aal3321.

Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, Bustamante CD, Rasmussen M, Snyder MP. 2015. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol* **33**: 736–742.

Tress ML, Abascal F, Valencia A. 2017. Most Alternative Isoforms Are Not Functionally Important. *Trends Biochem Sci* **42**: 408–410.

Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. 2018. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet*. http://dx.doi.org/10.1038/s41576-018-0017-y.

van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, Kirchner M, Maatz H, Blachut S, Sandmann C-L, et al. 2019. The Translational Landscape of the Human Heart. *Cell* **178**: 242–260.e29.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291**: 1304–1351.

Verhoeven VJM, Hysi PG, Wojciechowski R, Fan Q, Guggenheim JA, Höhn R, MacGregor S, Hewitt AW, Nag A, Cheng C-Y, et al. 2013. Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* **45**: 314–318.

Vignal A, Eory L. 2019. Avian Genomics in Animal Breeding and the End of the Model Organism. In *Avian Genomics in Ecology and Evolution: From the Lab into the Wild* (ed. R.H.S. Kraus), pp. 21–67, Springer

International Publishing, Cham.

Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, Mayer G, Perez-Riverol Y, Reisinger F, Ternent T, et al. 2016. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**: 11033.

Wang A, Bao Y, Wu Z, Zhao T, Wang D, Shi J, Liu B, Sun S, Yang F, Wang L, et al. 2019. Long noncoding RNA EGFR-AS1 promotes cell growth and metastasis via affecting HuR mediated mRNA stability of EGFR in renal cancer. *Cell Death Dis* **10**: 154.

Wan Y, Larson DR. 2018. Splicing heterogeneity: separating signal from noise. *Genome Biol* **19**: 86.

Weisser H, Wright JC, Mudge JM, Gutenbrunner P, Choudhary JS. 2016. Flexible Data Analysis Pipeline for High-Confidence Proteogenomics. *J Proteome Res* **15**: 4686–4695.

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**: 582–587.

Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, Choudhary JS, Harrow J. 2016. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat Commun* **7**: 11778.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377–394.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754–D761.

Zougman A, Mann M, Wisniewski JR. 2011. Identification and characterization of a novel ubiquitous nucleolar protein “NARR” encoded by a gene overlapping the rab34 oncogene. *Nucleic Acids Res* **39**: 7103–7113.

Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci

Jonathan M. Mudge, Irwin Jungreis, Toby Hunt, et al.

Genome Res. published online September 19, 2019
Access the most recent version at doi:10.1101/gr.246462.118