Defining Functional Genic Regions in the Human Genome through Integration of Biochemical, Evolutionary, and Genetic Evidence

Zing Tsung-Yeh Tsai,[‡,1,2] John P. Lloyd,[1] and Shin-Han Shiu*[,1]

[1]Department of Plant Biology, Michigan State University, East Lansing, MI

[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[‡]Present address: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

***Corresponding author:** E-mail: shius@msu.edu.

Associate editor: Naruya Saitou

Abstract

The human genome is dominated by large tracts of DNA with extensive biochemical activity but no known function. In particular, it is well established that transcriptional activities are not restricted to known genes. However, whether this intergenic transcription represents activity with functional significance or noise is under debate, highlighting the need for an effective method of defining functional genomic regions. Moreover, these discoveries raise the question whether genomic regions can be defined as functional based solely on the presence of biochemical activities, without considering evolutionary (conservation) and genetic (effects of mutations) evidence. Here, computational models integrating genetic, evolutionary, and biochemical evidence are established that provide reliable predictions of human protein-coding and RNA genes. Importantly, in addition to sequence conservation, biochemical features allow accurate predictions of genic sequences with phenotypic evidence under strong purifying selection, suggesting that they can be used as an alternative measure of selection. Moreover, 18.5% of annotated noncoding RNAs exhibit higher degrees of similarity to phenotype genes and, thus, are likely functional. However, 64.5% of noncoding RNAs appear to belong to a sequence class of their own, and the remaining 17% are more similar to pseudogenes and random intergenic sequences that may represent noisy transcription.

Key words: **functional genomic region, conservation, chromatin state, random forest classification.**

Article

Introduction

Recent studies have revealed widespread biochemical activities associated with the human genome (ENCODE Project Consortium 2012; Hangauer et al. 2013). In particular, there is pervasive transcription beyond known genic regions (Djebali et al. 2012). This transcriptional activity may be indicative of the presence of novel genic regions (Mercer et al. 2009). However, some of this activity can also be transcriptional noise (van Bakel et al. 2010). In addition, there are more than 10,000 annotated noncoding RNA (ncRNA) regions within the human genome, many of which have not been experimentally characterized and have no known function. Thus, the functional significance of transcripts originating outside of annotated genes and of most annotated ncRNA regions is unclear.

The foremost challenge in identifying functional genomic regions is in defining what constitutes function, which has been a topic of considerable discussion (Doolittle et al. 2014; Kellis et al. 2014). In the ENCODE project, a genomic region is defined as having a biochemical function if reproducible biochemical activity, for example, transcriptional activity or particular chromatin states, can be detected (ENCODE Project Consortium 2012). This biochemical function definition, however, has drawn critique because the existence of an activity does not necessarily mean such activity is under selection (Eddy 2012; Doolittle 2013; Graur et al. 2013; Niu and Jiang 2013). It has also been suggested that evolutionary, biochemical, and genetic evidence provide complementary information for the functionality of a sequence (Kellis et al. 2014). However, this operational definition is criticized for not distinguishing between causal role functionality (what a component does) and selected effect functionality (how and why a component is subjected to natural selection) (Neander 1991; Doolittle et al. 2014). And because biochemical activity measures the causal role of a genomic region, it remains an open question as to what, if any, biochemical evidence is sufficient to identify functional genomic regions.

The feasibility of jointly considering biochemical activities and evolutionary evidence for detecting selection is illustrated by fitCons (fitness consequences of functional annotation) that provides an estimate of fitness consequence for a point mutation (Gulko et al. 2015). Nonetheless, it is unclear whether and how biochemical, evolutionary, and genetic evidence in combination may provide a more robust definition of functional genic sequences. Here, we examined the relative contributions of 21 conservation attributes, 14 sequence characteristics, and 35 biochemical signals in distinguishing between genetically defined functional regions (human

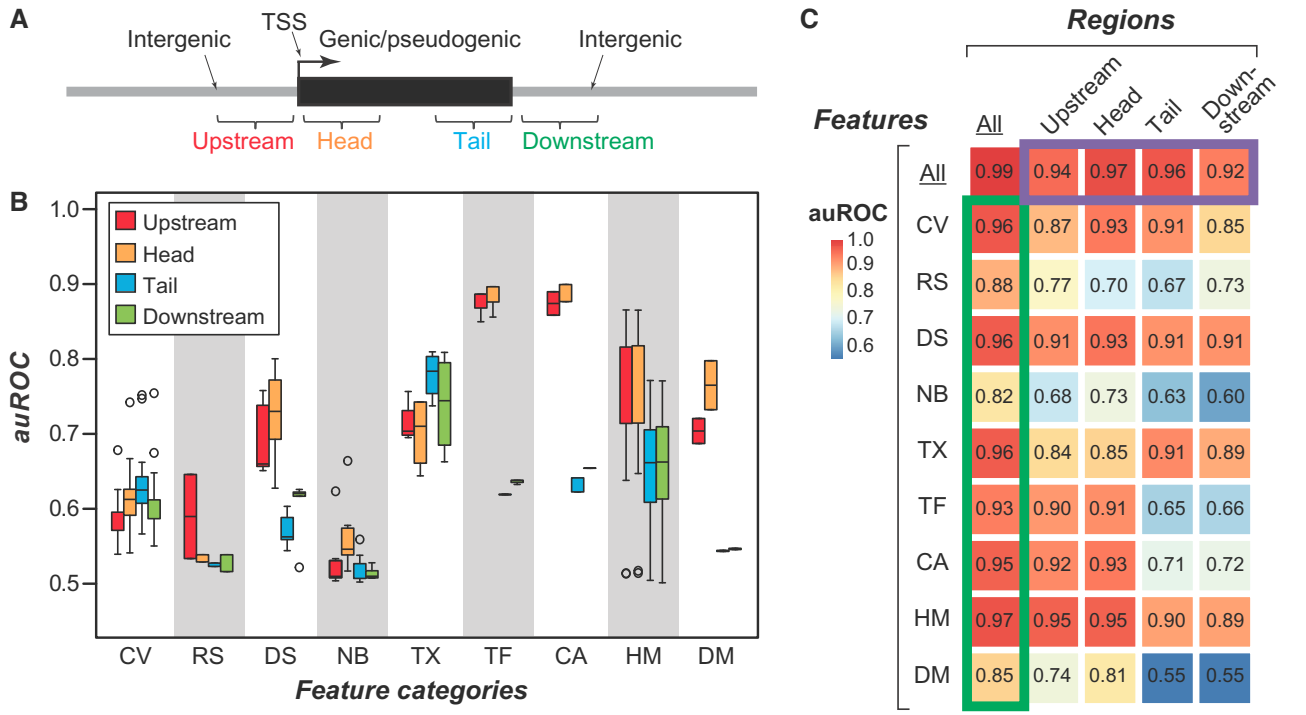

Fig. 1. Performance of classifying phenotype genes and pseudogenes (PS) by conservation, sequence property, and biochemical characteristics. (*A*) Schematic of the four 500-bp regions flanking the boundary of each entry in this study. (*B*) Boxplots of the auROC values for classifying phenotype genes and PS using each of the 70 targeted features in upstream, head, tail, and downstream regions including: 21 conservation (CV), two repeat and single nucleotide polymorphism (RS), five DNA structural property (DS), seven non-B DNA structure motif (NB), four transcription (TX), three transcription factor binding (TF), two chromatin accessibility (CA), 24 histone modification (HM), and two DNA methylation (DM) features. (*C*) The performance of random forest classification using combinations of features and regions. The first combination is using all features from all four regions (full model, top left box of panel *C*). The second combination consists of models built with all features from four regions separately (purple box; top of panel *C*). The third combination consists of models built with features from one category but with information from all four regions (green box; left of panel *C*). Finally, the fourth combination consists of models built with all features from a category in one region (below purple box and right of green box in panel *C*).

phenotype/disease genes) and likely nonfunctional sequences (pseudogenes [PS] and random intergenic sequences) in the human genome. In addition, machine learning models were established to investigate whether current annotated ncRNAs share evolutionary and biochemical features with functional sequences and may be considered genic or with nonfunctional sequences indicative of transcriptional noise.

Results and Discussion

No Single Feature Is Sufficient for Defining Functional Genic Regions

To define human genomic regions that function as genes (referred to as functional regions), we first assessed how well conservation attributes, sequence characteristics, and biochemical signatures (referred to as features, 70 total, see Materials and Methods and supplementary table S2, Supplementary Material online) could differentiate functional and nonfunctional regions. For functional regions, we used 3,046 phenotype and disease protein-coding genes with genetic evidence of functionality from the Human Phenotype Ontology database (Köhler et al. 2014). These phenotype/disease genes were referred to as HPO-p genes (supplementary table S1, Supplementary Material online). For nonfunctional regions, 4,399 human PS (Yates et al. 2016) were used (supplementary table S1, Supplementary Material online). We focused on defining 1,000-bp regions flanking the starting (upstream and head) and ending (tail and downstream) points of genes (fig. 1*A*), because features have distinctive patterns in the upstream and downstream regions (de Boer et al. 2014).

To assess how informative each feature was for predicting each of the four regions as HPO-p gene- or pseudogene-like, the area under Receiver Operating Characteristic curve (auROC) was used (fig. 1*B*). The auROC for a model that can make predictions perfectly is 1. At the other extreme, a model that does no better than random guesses has an auROC of 0.5 (see Materials and Methods). For each feature/region combination (e.g., nucleotide diversity calculated with sequences from the upstream region), an auROC value was calculated. In fig. 1*B*, instead of showing individual auROC values, we classified features into nine categories and showed the distribution of auROC values of all features in a category and from a particular region. Thus, looking across fig. 1*B*, we showed how informative features in a feature category/region combination were for distinguishing PS from genes.

In general, individual conservation-based features were among the least informative (average auROC = 0.61, fig. 1*B*), and the most informative feature was, the fitness Consequence

score (fitCons, auROC = 0.75) (Gulko et al. 2015), still has a 28.1% false positive rate (FPR, PS misclassified as genes) and 30% false negative rate (FNR, genes misclassified as PS). Meanwhile, transcription-related features were more informative than conservation (average auROC = 0.73; fig. 1B). However, the FPR and FNR of a model using the best performing feature (transcriptional coverage in tail regions) remained high at 34.1% and 17.5%, respectively. Like transcription, features related to transcription factor (TF) binding and chromatin accessibility are among the best in distinguishing HPO-p genes and PS (fig. 1B) but still have high FPR (≥13.9%) and FNR (≥12.1%). Taken together, no single feature was sufficient for defining functional regions, and the utility of features depended on which of the four regions (fig. 1A) was targeted.

Biochemical and Evolutionary Evidence Are Complementary in Defining Functional Genic Regions

Based on the observation that HPO-p genes and PS were clearly distinguishable in a principal component analysis using all features (supplementary fig. S1, Supplementary Material online), we next jointly considered all 70 features from all four regions with machine learning methods (the full model; see Materials and Methods). This is a binary classification model as it differentiates between human genomic sequences that are more similar to HPO-p genes (thus likely functional) or PS (likely nonfunctional). This full model significantly outperformed any single feature (auROC = 0.99, FPR = 4.5%, FNR = 8.4%, fig. 1C), distinguishing HPO-p genes and PS with high accuracy. To assess the relative contributions of different features and upstream, head, tail, and downstream regions, we established four additional types of models including: 1) Four region-specific models combining all features (purple rectangle, fig. 1C), 2) nine feature category-specific models combining all regions (green rectangle, fig. 1C) revealing histone modification-related features as the most informative category, 3) 36 feature category/region-specific models (fig. 1C) with highly variable performance, and 4) 280 "leave-one-out" models where each feature/region combination was removed to evaluate its importance (supplementary fig. S2, Supplementary Material online).

Interestingly, models considering all features, regardless of which regions we focused on, performed nearly as well as the full model. This was also true for some feature categories when all regions were considered. For example, consideration of all conservation features resulted in a well-performing classifier (fig. 1C). Similarly, biochemical signature-related categories, particularly histone modifications (fig. 1C), perform nearly as well as the full model. Although conservation or biochemical features alone are useful, 58 HPO-p genes were correctly predicted only by evolutionary conservation features and 168 HPO-p genes only by biochemical features. This finding echoes the suggestion that biochemical and evolutionary evidence are complementary in defining functional DNA sequences (Kellis et al. 2014).

Functional Likelihood (FL) Allows Prediction of Phenotype/Disease Protein-Coding Genes and PS

In the interest of predicting functionality of any genomic region aside from the HPO-p genes and PS, we devised an FL measure for a genomic sequence represented by a value between 0 (most likely nonfunctional) and 1 (most likely functional) (see Materials and Methods). The median FLs of HPO-p genes and PS were 0.97 (fig. 2A) and 0.01 (fig. 2B), respectively. To assess the error rates of using FLs to call a genomic region as functional or not, a threshold FL of 0.36 was determined as the FL value that leads to the maximal F-measure of a model (supplementary fig. S3, Supplementary Material online). F-measure is the harmonic mean of precision (proportion of sequences predicted as functional that are truly functional) and recall (proportion of functional sequences predicted as functional) values. Thus, a threshold FL based on maximum F-measure allows us to reduce both false positive and false negative predictions for a given model.

With this threshold, 94.5% of HPO-p genes are considered functional. For HPO-p genes classified as nonfunctional that are clearly false negatives, we speculated that low FL scores among these sequences might be a result of conditional or tissue-specific expression because PS tend to have highly restricted expression profiles. In this case, the specific conditions or tissues in which these sequences are functional may not be adequately captured by the data sets used when generating features. To assess this possibility, we investigated expression specificity, defined by how often a gene is expressed across multiple cell lines, and found that low FL HPO-p genes tend to be tissue-specific (one-sided Wilcoxon rank-sum test $P = 2.4 \times 10^{-29}$, supplementary fig. S4A, Supplementary Material online). This may suggest that the model is biased against narrowly expressed sequences. In addition, we also found that low FL HPO-p genes tend to have higher proportions of intron sequences in the gene body regions where features were calculated (one-sided Wilcoxon rank-sum test $P = 7.2 \times 10^{-5}$, supplementary fig. S4C, Supplementary Material online), which may result in low FL scores and false negative predictions.

The prediction model classified 93.5% of PS as nonfunctional. Nonetheless, 6.5% PS with higher FL than the threshold were classified as functional. For these high FL PS, we considered both false positives (truly nonfunctional) and misannotation (annotated PS that were in fact functional) as explanations. Regarding false positives, these PS may be present in a similar chromatin context as nearby or overlapping genes and thus exhibit features similar to functional regions. Consistent with this, high FL PS tend to overlap with annotated genes twice as often compared to low FL PS (Fisher's Exact Test, $P < 1.6 \times 10^{-22}$; fig. 3A). After eliminating PS overlapping with annotated genes, only 2.1% of PS were classified as functional (fig. 3B). High FL PS also tend to be closer to annotated genes than low FL ones (fig. 3C and D). We also assessed the possibility that PS generated from more recent duplication events may still possess features like functional paralogs leading to high FLs. Contrary to this expectation, we found that in fact high FL PS were generated from more

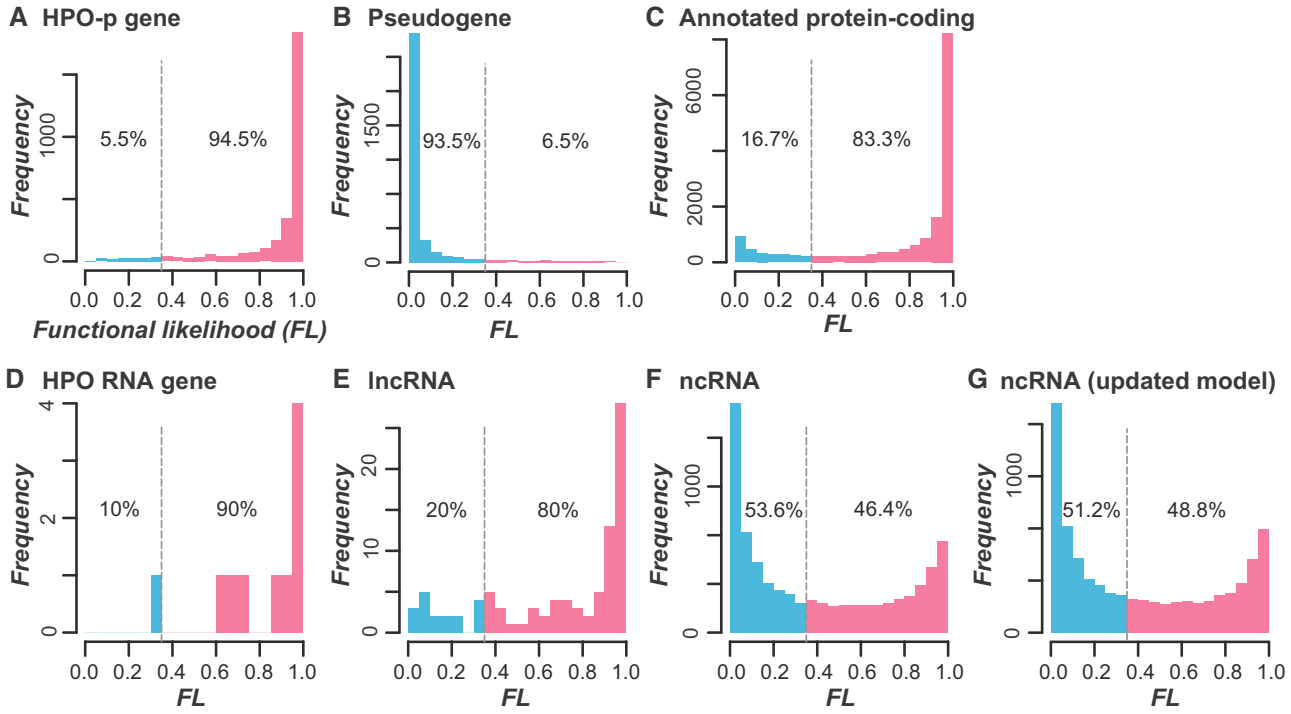

Fig. 2. Binary classification of sequences as likely functional or not. (*A*–*F*) Distributions of functional likelihood determined by a random forest model with all 70 features from all four regions and trained with HPO-p genes and PS: (*A*) HPO-p genes, (*B*) PS, (*C*) annotated protein-coding genes, (*D*) HPO RNA genes, (*E*) lncRNAs, and (*F*) ncRNAs. (*G*) The FL distribution of ncRNAs with an updated full model trained by a combined positive set (HPO-p, HPO RNA, and lncRNAs) and a combined negative set (PS and RIR). The vertical dashed lines indicate the FL threshold determined by maximizing *F*-measure to classify a sequence as functional or nonfunctional.

ancient duplication events compared to low FL PS (fig. 3*E*). One interpretation is that, because genes that tend not to have retained duplicates have higher essentiality (Lloyd et al. 2015), they might resemble essential genes more than an average gene in the genome and were misclassified. Finally, we cannot rule out the possibility that a small subset of predicted PS are misannotated and thus functional. Nonetheless, considering that the proportion of high FL PS decreases to 2.1% after controlling for overlap with annotated genes, the low FPR indicates that our model is capable of distinguishing functional and nonfunctional sequences in a highly accurate manner.

FL Also Allows Prediction of Annotated Protein-Coding and RNA Genes

After evaluating our model performance through cross-validation of hold-out HPO-p genes and PS, we next assessed the FLs of annotated protein-coding genes and RNA genes. We found that annotated protein-coding genes, after excluding HPO-p entries, are generally predicted as functional (fig. 2*C*). However, 17% of annotated protein-coding genes have lower FLs than the threshold and thus are predicted as nonfunctional (supplementary table S1, Supplementary Material online). Similar to results with low-scoring HPO-p genes, we find that protein-coding genes that are predicted as nonfunctional tend to be more tissue-specific in expression (one-sided Wilcoxon rank-sum test $P = 8.2 \times 10^{-177}$, supplementary fig. S4*B*, Supplementary Material online) and have higher intron proportions in the gene body regions used to calculate features (one-sided Wilcoxon rank-sum test $P = 5.4 \times 10^{-6}$, supplementary fig. S4*D*, Supplementary Material online) compared to predicted-functional protein-coding genes. Using a set of previously defined human retrogenes (Kabza et al. 2014), we found that low-FL protein-coding genes are enriched in retrogenes (Fisher's Exact Test, $P = 9.3 \times 10^{-4}$). Retrogene sequences are derived from reverse transcription and genomic reinsertion and, due to the lack of proper regulation in the new genomic context, most retrogenes are likely dead-on-arrival (Kaessmann et al. 2009). Furthermore, 557 of 2,784 low FL protein-coding genes are not annotated with specific functions or assigned to any known pathway. To further assess the functionality of these low FL annotated genes, we compared the growth effects of mutants in CRISPR global loss-of-function experiments (Gilbert et al. 2014) and mouse phenotype data (White et al. 2013). We found that annotated genes with low FLs tend to have higher growth rates when mutated compared to high FL genes ($r = -0.12, P < 2.2 \times 10^{-16}$, supplementary fig. S5, Supplementary Material online). Similarly, mouse orthologs of low FL genes tend to be nonessential (Fisher's Exact Test, $P = 9.6 \times 10^{-5}$, supplementary fig. S6, Supplementary Material online). Taken together, these findings indicate that FL values are an accurate estimate of the functional state for most annotated protein-coding genes, and a subset of annotated protein-coding genes may be false positive gene predictions.

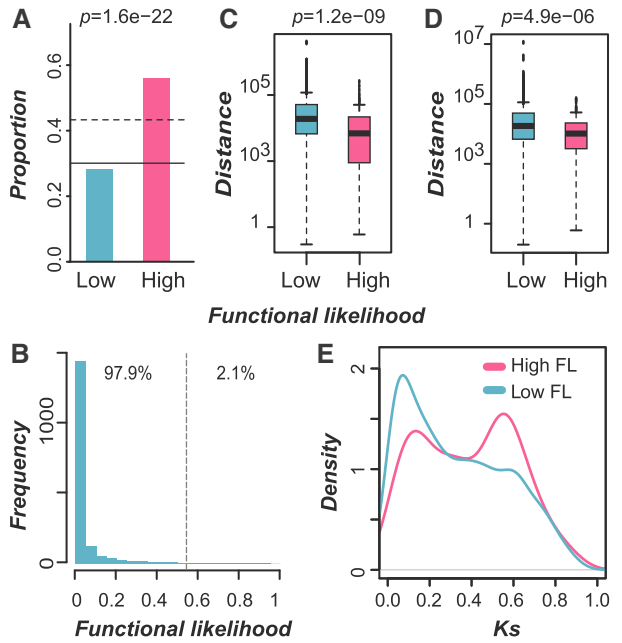

FIG. 3. Comparisons between high FL and low FL PS. (*A*) The proportion of PS with low or high FLs overlapping with annotated genes. A pseudogene was regarded as low FL if its FL value based on the full model was lower than the threshold FL that maximized the F-measure. Otherwise, it was regarded as high FL. The *P* value of a proportion test is shown. The horizontal dashed line and solid line indicate the proportion of PS and other gene regions, respectively, which overlap with annotated genes. (*B*) The FL distribution of PS that do not overlap with annotated genes. The distances between PS and their nearest (*C*) upstream or (*D*) downstream neighboring gene are shown in boxplots; *P* values are from one-sided Wilcoxon rank-sum tests. (*E*) The *Ks* distributions of high FL (pink) and low FL (blue) PS.

We next asked whether the full machine learning model can be applied to identify functional ncRNAs because of the importance of RNA genes (Fatica and Bozzoni 2013), the availability of >10,000 annotated human ncRNAs (Harrow et al. 2012), and the current debate on the functionality of pervasive transcription in the human genome (Graur et al. 2013; Niu and Jiang 2013; Doolittle et al. 2014). First, we applied the model to ten HPO RNA genes not included in the training data and nine, including *XIST* (Quinn and Chang 2015), were classified as functional (fig. 2*D*). To further verify the utility of the model in classifying RNA genes with a larger data set, we examined an additional 92 manually curated long ncRNAs that were annotated as functional (lncRNAs; Quek et al. 2015). We found that 80.5% of lncRNAs have FLs higher than the threshold and were predicted as functional (fig. 2*E*). These findings indicate that most known, functional RNA entries can be classified correctly, demonstrating that this integrated model can predict not only protein-coding but also RNA genes.

Nearly Half of the ncRNAs Are Predicted as Likely Functional Based on a Binary Classification Model

Given that the integrated model can distinguish HPO protein-coding and RNA genes and lncRNAs from PS, we next asked what proportion of the 10,924 annotated ncRNA entries were likely functional. Intriguingly, annotated ncRNA entries displayed a bimodal FL distribution (fig. 2*F*) where 46.4% were classified as HPO-p gene-like and are likely functional. But the rest (53.6%) more closely resembled PS, and thus were likely nonfunctional. To assess why the FL distribution of ncRNA regions is not as clear-cut as the other features, including HPO RNA and lncRNA genes, first we asked whether this was because the boundaries of ncRNAs were ill-defined and led to false predictions. If the boundaries were ill-defined and impacted the model significantly, we would expect that a model based on only the head/tail regions (fig. 1*A*) would outperform the full model because the head/tail model does not include up- and down-stream regions, which may contain genic regions. We found that the head/tail model led to results that were nearly identical to the original, full model (supplementary fig. S7, Supplementary Material online). Thus, the rather ambiguous FLs among ncRNAs are not simply due to ill-defined boundaries.

Another explanation is that the machine learning model based on protein-coding genes did not adequately capture the properties of RNA genes. This is unlikely as most known RNA genes were classified correctly (fig. 2*D* and *E*). Nonetheless, we further assessed this possibility by developing a new model trained with lncRNA genes as functional examples and PS as nonfunctional examples (supplementary fig. S8, Supplementary Material online). Although this lncRNA-based model led to higher error rates in predicting HPO-p entries (FNR = 11.6%, supplementary fig. S8*A*, Supplementary Material online), accuracy for lncRNA predictions was improved by 5.5%, as expected (FNR = 14.1%, supplementary fig. S8*E*, Supplementary Material online). Most importantly, the lncRNA-based model led to a 5% increase in the number of ncRNAs predicted as functional. However, the FL distribution for annotated ncRNAs remained bimodal and 48.6% of ncRNAs were still classified as nonfunctional (supplementary fig. S8*F*, Supplementary Material online). Taken together, up to 50% of the annotated ncRNAs are likely functional. Considering that the functionality of a great majority of these high FL ncRNAs is unknown, our findings indicate their resemblance to known protein-coding and RNA genes and provide additional evidence that they are likely bona fide genes. Consistent with this notion, high FL ncRNAs are enriched in ultraconserved noncoding sequences (Dimitrieva and Bucher 2013) compared to low FL ncRNAs (Fisher's Exact Test, $P = 3.9 \times 10^{-4}$). Meanwhile, the other 50% of the ncRNAs more closely resemble PS, raising the question of their functional significance.

Most ncRNAs Are More Similar to PS and Random Intergenic Sequences than They Are to Protein-Coding and RNA Genes

In the binary classification scheme above, an ncRNA was classified as either resembling the positive (HPO-p or lncRNA) or the negative (pseudogene) examples. Although the ncRNAs with low FLs were more similar to PS, it is also possible that they more closely resemble other genomic

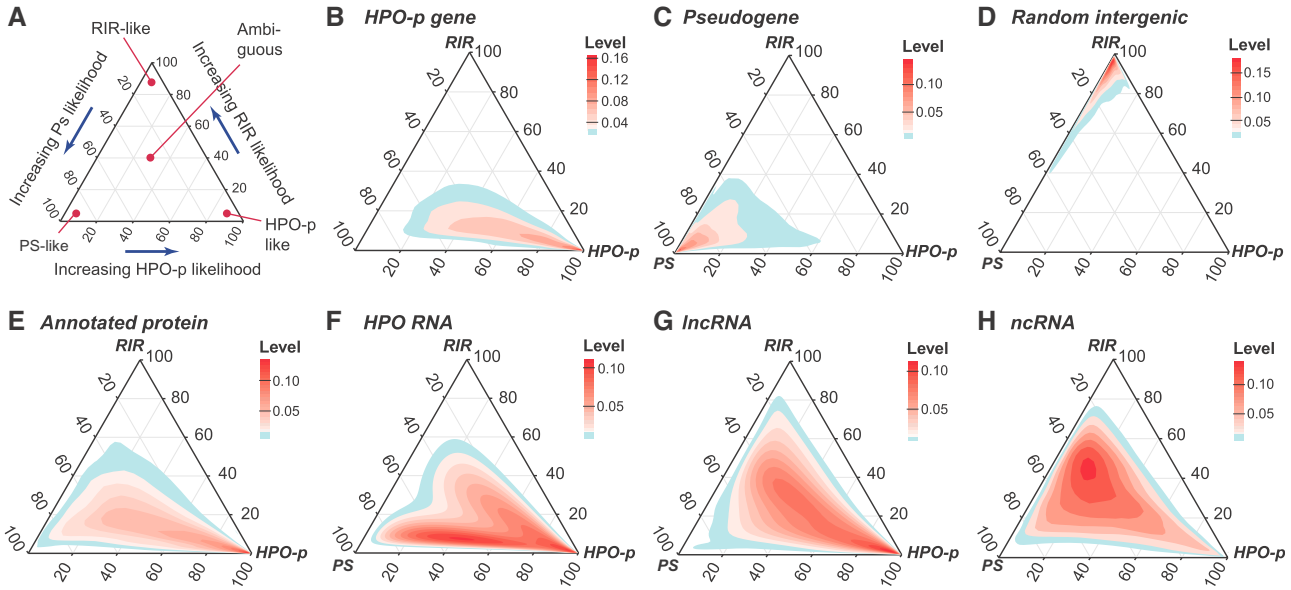

FIG. 4. Three-class classification of sequences. (*A*) An example output of the ternary likelihood distributions. The likelihood values were determined by a three-class random forest model trained with HPO-p genes, PS, and RIR. For each input sequence, the three-class model provided three likelihood scores that represent how similar a sequence entry is to HPO-p genes, PS, and intergenic sequences, respectively. The axes represent the likelihood score between 0 and 100 in which larger value indicates higher degree of similarity to RIR (top corner), PS (bottom-left corner), or HPO-p genes (HPO-p, bottom-right corner). (*B*–*H*) The ternary likelihood distributions for: (*A*) HPO-p genes, (*B*) PS, (*C*) RIR, (*D*) annotated protein-coding genes, (*E*) HPO RNA genes, (*F*) lncRNAs, and (*G*) ncRNAs. The darker red indicates increasing number of entries. The blue shade indicates the boundary of the distribution.

features, such as intergenic regions, or belong to a class of their own. To assess these possibilities, we established a three-class model trained by considering three different sequence types: HPO-p genes (phenotype/disease genes), PS, and random intergenic regions (RIR). For each input sequence to be evaluated, the three-class model provided three likelihood scores that represent how similar a sequence entry is to HPO-p genes, PS, or random intergenic sequences. This three-class model provided a comparison between two nonfunctional sequence types—PS and RIR. Moreover, it allowed us to assess whether functionally ambiguous genomic sequences identified from the binary model, such as some ncRNAs, more closely resemble RIR, PS, or HPO-p genes. To visualize our findings, ternary plots were generated that indicate the similarity of input sequences to these three sequence types (fig. 4). An input sequence would be closer to the top, bottom-left, and bottom-right corners if it more closely resembles RIR, PS, and HPO-p genes, respectively (fig. 4*A*). Like the two-class model, we found that the three-class model could accurately distinguish HPO-p genes (fig. 4*B*), PS (fig. 4*C*), and RIR (fig. 4*D*) because their likelihood values were distributed more densely in their respective corners in the ternary plots. Importantly, the three-class model provides additional resolution in resolving how these sequences differ and the small but apparent overlaps in the model space between HPO-p genes and PS.

Consistent with the binary classification results, most annotated protein-coding genes (fig. 4*E*), HPO RNA genes (fig. 4*F*), and lncRNA genes (fig. 4*G*) were more similar to HPO-p genes. The HPO RNA genes fell into two distinct clusters where sequences in one cluster were more similar to HPO-p genes and those in the other cluster sat halfway between HPO-p genes and PS (fig. 4*F*). In addition, some lncRNAs were in the classification space that was ambiguous, consistent with the fact that ~19.5% of lncRNAs were classified as nonfunctional in the binary model (fig. 2*E*). In contrast to HPO RNA and lncRNAs, most ncRNA entries were concentrated in a space half-way between intergenic regions and PS but far from the HPO-p corner (fig. 4*H*). To see if this was because the three-class model where HPO-p genes were used as part of the training data resulted in a biased model against RNA genes, we established another three-class model classifying lncRNA (instead of HPO-p genes), PS, and RIR. The same pattern was recovered from this new three-class model (supplementary fig. S9*A*–*G*, Supplementary Material online), indicating the placement of ncRNAs in the classification space did not simply result from undue influence of HPO-p training data. With this information, we updated our binary classification model to distinguish a combined positive set (HPO-p, HPO RNA, and lncRNAs) from a combined negative set (PS and RIR). In this new model (fig. 2*G*), 48.8% ncRNAs were considered functional, but 51.2% of them were still classified as more similar to the mostly nonfunctional PS and RIRs.

Four-Class Models Reveal That Some ncRNAs May Belong to a New Class of Genomic Feature

Before claiming these ncRNAs as transcriptional noise, we asked whether the low FL ncRNAs represent a new class of sequences that do not resemble known protein-coding and RNA genes. To assess this possibility, a four-class model was established for classifying HPO-p, PS, RIR, and ncRNAs (see

Materials and Methods). For each sequence, a likelihood value for each class was determined and the sequence was classified as the most likely class. Consistent with the binary (fig. 2) and the three-class (fig. 4) models, most HPO-p genes (88.4%, fig. 5A), PS (86.1%, fig. 5B), and RIR (98.7%, fig. 5C) were classified correctly. In the case of ncRNAs (fig. 5D), 18.5% were more similar to HPO-p genes, suggesting that they are most likely functional. This provides a conservative estimate for the functionality of ncRNAs relative to the binary classification scheme described above. In addition, 17% of ncRNAs more closely resembled PS and/or RIR and may be noisy transcription. The remaining ncRNAs (64.5%) could be separated from the other three classes, supporting the notion that a subset of ncRNAs have distinct characteristics and belong to a class of their own.

However, compared to the other three classes (fig. 5A–C), the ncRNAs in this "ncRNA class" (median ncRNA likelihood = 0.48, fig. 5D) were classified with substantial ambiguity because they mostly have high pseudogene likelihood (median = 0.18), random intergenic region likelihood (median = 0.15), and HPO-p gene likelihood (median = 0.14). This pattern may be because, despite the use of multiple biochemical and conservation features, some crucial distinguishing characteristics remain to be discovered. This is consistent with the observation that some HPO-p genes (fig. 5A) and annotated protein-coding genes (fig. 5E) have minor (not dominant) but appreciable ncRNA likelihood (yellow). Interestingly, although most HPO RNA genes (80%, fig. 5F) and lncRNAs (56.6%, fig. 5G) were classified as HPO-p like, both sets of functional RNA genes have higher median ncRNA likelihoods (0.23 and 0.27, respectively) compared to HPO-p (0.11, fig. 5A), suggesting that there are some common properties between these RNA entries. This is corroborated by the pattern from the three-class model where a portion of the lncRNA distribution in the classification space (fig. 4G) overlaps with the peak region of ncRNAs (fig. 4H). Thus, there is a clear continuum between some ncRNAs and lncRNAs given the features we have examined, raising the question whether some of these ncRNAs are precursors from which novel genes may evolve.

Conclusion

In summary, computational models considering conservation, sequence-structural, and biochemical features allow accurate predictions of known protein-coding and RNA genes from nonfunctional sequences. Features relevant to evolutionary conservation and those based on biochemical activities can be used independently for building models with comparable performance and are complementary. However, this does not mean that the presence of a biochemical activity suggests that a genomic region is under selection and thus has functional significance. Rather, consideration of multiple biochemical features in combination allows identification of genic sequences likely under strong purifying selection and may serve as an alternative measure of selection.

By applying these models, we answer the question of what proportion of expressed genomic regions, particularly those

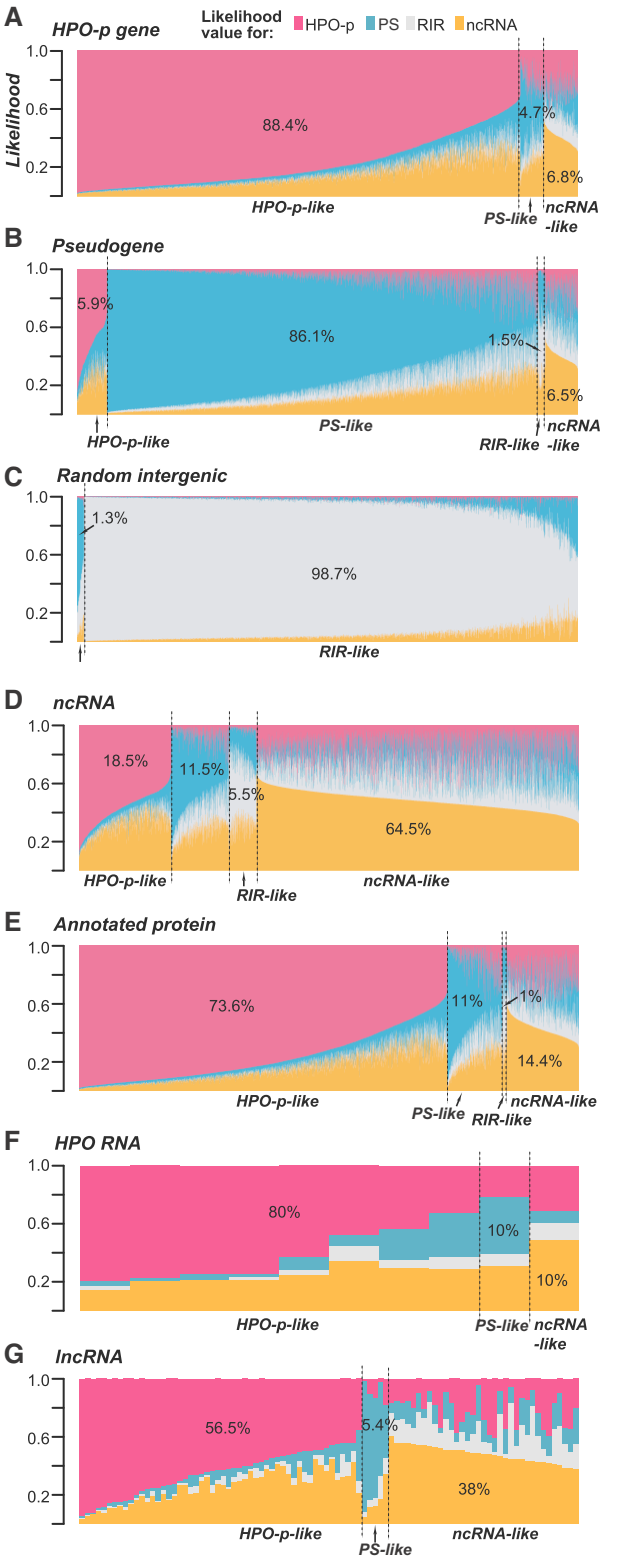

FIG. 5. Four-class classification of sequences. Likelihood values (*y* axis) that a sequence (*x* axis) belongs to each of the four classes: HPO-p genes (red), PS (blue), RIR (gray), and ncRNAs (yellow) for (*A*) HPO-p genes, (*B*) PS, (*C*) RIR, (*D*) ncRNAs, (*E*) annotated protein-coding genes, (*F*) HPO RNA genes, and (*G*) lncRNAs. Each vertical bar represents the likelihoods of the four classes for a single sequence. Each sequence was classified into one of the four classes based on the highest likelihood value. Percentages indicate the proportion of sequences that belong to each class.

annotated as ncRNAs, is likely functional. We find support for the functionality of 18.5% annotated ncRNAs based on their resemblance to known genes. Nonetheless, the functional significance of the remaining ncRNAs remains unclear. While most dissimilar from functional disease-gene regions, these ncRNAs are apparently also distinct from PS and RIR and could represent a novel class of sequences with unknown functional significance. Regardless, as these sequences do not have sufficient similarity to known functional sequences, our null hypothesis is that they represent transcriptional noise, which can be falsified once experimental evidence of their functionality is established.

Materials and Methods

Sequence, Annotation, and Training Data

The human genomic sequence and annotation data used in this study (GRCh37/hg19) were obtained from Ensembl (Yates et al. 2016). Additional annotation data used included the phenotype/disease gene annotations from the Human Phenotype Ontology (HPO) database (Köhler et al. 2014) and the functional RNA annotation from the lncRNAdb (Quek et al. 2015). All annotated entries used in the study were $\geq$1 kilobase (kb). For machine learning, the positive, functional examples include 3,046 HPO protein-coding genes (referred to as HPO-p) and 92 annotated functional RNA genes from the lncRNAdb (referred to as lncRNAs) (Quek et al. 2015). These are genes with known disease or phenotypic consequences when mutated. The negative, nonfunctional examples include 4,399 entries of the Ensembl pseudogene biotype. In cases where PS were also annotated as functional RNA in the lncRNAdb, they were treated as lncRNAs only.

We also used 2,500 intergenic regions randomly chosen from human genomic regions that did not overlap with any Ensembl, HPO, or lncRNAdb annotations as putative negatives. In addition to the above sequences, for machine learning, 16,618 Ensembl genes in the protein-coding biotype (annotated protein coding), 10,924 Ensembl entries in the ncRNA biotype (annotated ncRNA), and 10 HPO RNA genes were also examined. Any Ensembl-annotated protein-coding gene that was also annotated in HPO was taken out of the Ensembl category and treated as an HPO-p. Similarly, any ncRNA annotated in the lncRNAdb was taken out of the annotated ncRNA category and treated as an lncRNA. Information on the identifiers, source databases, and locations of the sequences used are in supplementary table S1, Supplementary Material online.

Conservation Features

There were three types of conservation features used. The first type was the nucleotide sequence identity of each human sequence compared with their putative orthologs in five primates (*Pan troglodytes* CHIMP2.1.4, *Gorilla gorilla* gorGor3.1, *Pongo abelii* PPYG2, *Macaca mulatta* MMUL_1, *Callithrix jacchus* C_jacchus3.2.1) and nine eutherians (*Mus musculus* GRCm38, *Rattus norvegicus* Rnor_5.0, *Oryctolagus cuniculus* OryCun2.0, *Canis familiaris* CanFam3.1, *Felis catus* Felis_catus_6.2, *Equus caballus* EquCab2, *Bos taurus* UMD3.1, *Ovis aries* Oar_v3.1, *Sus scrofa* Sscrofa10.2). Thus, for each sequence, 14 identity values were used as features. The second type was phastCons scores of alignments between human and 99 other vertebrate genomes (the phastCons100way data set; Siepel et al. 2005). For each sequence, phastCons scores were used to generate three feature values. The first was the third quartile of phastCons scores instead of the median to capture sequences with relatively small parts that were under selection. The second was the percentage of positions having a phastCons score $>$0.5. The third was the percentage of positions having a phastCons score $>$0.75.

A nucleotide position outside of a 100-way aligned region, and therefore without phastCons data, was assigned a value of zero to capture that lack of conservation evidence. The third type of conservation features was fitness Consequence (fitCons) score, a summary statistic used to represent the effect of a mutation (Gulko et al. 2015). Although the fitCons score is an ensemble measure based on multiple features not directly related to conservation, it measures the fitness consequence of mutations of a nucleotide position and is thus considered in this category. For each sequence, the average and maximum fitCons scores of all nucleotide positions were calculated. Specifically, the i6 scores from fitCons were used that integrated across HUVEC, H1-hESC, and GM12878 cell types.

Sequence Property Features

There are four types of sequence property features. The first is the coverage of simple sequence repeats including low-complexity sequences and interspersed repeats identified by RepeatMasker (Smit et al. 2013) version 4.0.3. The second type is the single nucleotide polymorphism (SNP) density (SNP number/kb) as annotated in the dbSNP database (Sherry et al. 2001) build 146. The third and fourth types of sequence property features are related to DNA structure. The third type is the DNA dinucleotide structural information consisting of 125 conformational and thermodynamic dinucleotide properties collected from the DiProDB database (Friedel et al. 2009). The dimensionality of the DiProDB data set was first reduced with principal component analysis. Over the length of each sequence, the values of the top five principal components (PCs) (explained 83.3% of variation) were calculated every two nucleotides (window size of two) with a step size of one base.

For each of the five PCs, the values over all windows for each sequence were calculated. The five PCs mainly correspond to DNA major groove geometry, free energy, twist and roll, DNA minor groove geometry, and tilt and rise, respectively (Tsai et al. 2015). The fourth type is non-B DNA secondary structure that may cause DNA rearrangements and increased mutational rates (Zhao et al. 2010). For each sequence, the density of each of seven sequence motifs forming non-B DNA secondary structures (number of motif occurrences/kb) was calculated using the precomputed data from the non-B DB database (Cer et al. 2013). The seven non-B secondary structure forming motifs included: A-phased, direct, G-quadruplex forming, inverted, mirror, short tandem,

and Z-DNA motifs. A counted motif occurrence had ≥1 bp overlap with the targeted sequences.

Transcription-Related Features

The ENCODE RNA-sequencing (RNA-seq) tracks (CSHL Long RNA-seq) from 19 human cell lines (A549, AG04450, BJ, B cells CD20+, GM12878, H1-hESC, HeLa-S3, HepG2, HMEC, HSMM, HUVEC, IMR90, K562, MCF-7, Monocytes CD14+, NHEK, NHLF, SK-N-SH, and SK-N-SH RA) were obtained from the UCSC genome browser (Speir et al. 2016). These reads were paired-ends and have been mapped against the hg19 assembly (Parkhomchuk et al. 2009). As the RNA-seq data report strand of origin for transcripts, sense and antisense transcripts were analyzed separately. The features including expression levels over 19 cell types, RNA-seq read coverage, and read counts were calculated separately for sense and antisense reads. Thus, there were four RNA-seq-based features for each sequence. A genomic position without RNA-seq reads mapped to it was assigned a value 0. For expression level, the maximum sense or antisense read coverage over all bases in a sequence was first determined. Maximum read depth was then averaged over all cell lines to represent expression level.

For read coverage, we calculated the percentage of positions having sense or antisense read depth >0 in ≥1 cell line for each sequence. In addition to RNA-seq, the chromatin-immunoprecipitation sequencing (ChIP-seq) data for 161 TFs in 91 cell types were the ENCODE TxnFactor ChIP Track from the UCSC genome browser (Wang et al. 2012). For each ChIP-seq peak, the expScore based on the input signal values was used to represent binding intensity (0–1,000, provided in the UCSC genome browser). For each sequence, three ChIP-seq-based feature values were calculated. The first was the average binding intensity defined as the average expScore of the sites bound by each TF. The second is the average number of binding sites per kilobase among all TFs in the targeted region. The third is the number of TFs with expScore >0 in the targeted region.

Chromatin Accessibility, Histone Modification, and DNA Methylation

The ENCODE DNase I Hypersensitivity Clusters in 125 cell types (v.3; Thurman et al. 2012) from the UCSC genome browser were obtained to calculate two features for each sequence. The first was the coverage of accessible region defined as the proportion of base pairs with a DNaseClusters track value >0. The second was the maximum DNaseClusters track value over all positions of the sequences in question. For histone modifications, the ChIP-seq data generated by ENCODE/Broad Institute (Ram et al. 2011) were obtained from the UCSC genome browser that contained 156 processed data sets for 12 marks (CTCF, H2A.Z, H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me2, H3k4me3, H3k79me2, H3k9ac, H3k9me3, and H4k20me1) across 13 cell lines (GM12878, H1-hESC, HeLa-S3, HepG2, HMEC, HSMM, HUVEC, K562, Monocytes CD14+, NHA, Nhdfad, NHE, and NHLF).

For each sequence, two features were calculated for each histone mark. The first is an average score showing the intensity of histone modification (ChIP-seq read coverage, normalized between 0 and 1,000, position with missing values excluded). The second is the proportion of positions having an average histone modification intensity score >0. For DNA methylation, the reduced representation bisulfite sequencing data of 15 cell lines (AG04450, BJ, GM12878, H1-hESC, HeLa-S3, Hepatocytes, HepG2, HMEC, HSMM, K562, IMR90, MCF-7, Osteobl, SK-N-SH, and SK-N-SH RA) were obtained from the UCSC genome browser as the ENCODE DNA methylation tracks (Meissner et al. 2008). For each sequence, density of DNA methylation site (number of sites/kb) and the average DNA methylation score (bisulfite sequencing read depth, normalized between 0 and 1,000) across cell lines were calculated.

Machine Learning Approach, Functional Likelihood, and Model Performance Metrics

A machine learning framework based on random forest was developed to predict whether a genomic region would be functional or not. Random forest was chosen because of its efficiency on large data sets, its ability to report the importance of each feature, and accuracy in predictions (Breiman 2001). To avoid potential bias due to class imbalance, equal numbers of positive and negative examples were used for each training round. We utilized 10-fold cross-validation, where functional prediction models were built using 90% of positive and negative class sequences. To assess performance, the model was then applied to the withheld 10% of sequences. The trained model was also applied to the rest of the sequence entries not used for training or testing to predict whether they belonged to the positive or the negative class. For each sequence entry not in the training set, the proportion of positive decision tree predictions in the random forest model was calculated. By repeating the procedure 1,000 times, we then calculated the average proportion and defined it as FL. For multiclass models, the overall procedure was the same except that multiple classes were defined. Two three-class models were defined including one classifying HPO-p genes, PS, and RIR and the other replacing HPO-p genes with lncRNAs. In each run, the model would classify a test case into one of the three classes. In the four-class model, ncRNA was added as a new class. Following the same procedure as the binary classification, a confidence score for each of the classes for each sequence was determined. The confidence scores of classes (three and four for the three- and four-class models, respectively) for each sequence would add up to one. The random forest analyses were conducted in R using the "party" package (Strobl et al. 2009) and the "PRROC" package (Grau et al. 2015).

For evaluating the performance of features, multiple metrics were used including true positive rate (TPR), FNR, FPR, auROC, and F-measure. To determine the auROC of a model based on a particular feature, we first used multiple threshold values of the feature in question to determine corresponding TPRs and FPRs. A Receiver Operating Characteristic curve (ROC) was then drawn by plotting the TPRs against their

corresponding FPRs. The auROC was calculated based on the ROC. In this framework, a feature that can perfectly distinguish genes from PS has auROC of 1. A completely uninformative feature will have an auROC of 0.5. F-measure is the harmonic mean between the proportion of sequences predicted as genes that are truly genes (precision) and the proportion of true positive genes predicted as genes (recall). For a random forest model, a threshold FL was determined based the maximum F-measure of a model using multiple FL values ranging from 0 to 1. This approach allowed us to consider both false positive and false negative values at the same time when determining a threshold FL to classify functional sequences.

Evaluation of the Impact of Dependence between Features in Machine Learning

As several features discussed in this study were not independent, we adopted conditional random forest model and conditional permutation variable-importance measure in the "party" R package which have been demonstrated to be particularly suitable for correlated predictor variables (Strobl et al. 2009). To investigate the potential impact of feature dependency to prediction performance, we also developed prediction models with independent features using two methods. In the first method, we applied PC analysis and developed a PC model using all PCs, which are orthogonal with each other and therefore independent. In the second method, an independent component (IC) model was generated by utilizing 250 ICs that were calculated with the fastICA package in R. The auROC for the PC and IC models were 0.998 and 0.892, respectively, compared to 0.988 for the full model using the original, dependent features. Thus, the ability to distinguish between functional and nonfunctional sequences was not negatively impacted by the use of dependent predictor variables. Given that the full model using original features that were dependent could in some cases reveal the relative contributions of evolutionary and biochemical features in defining functional region, we utilized results from the model built with the original, untransformed features in all following analyses.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank David Arnosti, Melissa Lehti-Shiu, Monique Floer, and Amy Ralston for critical reading of the manuscript and suggestions. This work was supported in part by the National Science Foundation (MCB–1119778, IOS-1126998, and IOS-1546617), a Michigan State University Discretionary Funding Initiative grant (to S.-H.S.), and the Taiwan Ministry of Science and Technology Postdoctoral Research Abroad Program MOST-104-2917-I-564-070 (to Z.T.-Y.T.).

References

Breiman LEO. 2001. Random forests. *Mach Learn*. 45:5–32.
Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT, et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res*. 41:D94–D100.
de Boer CG, vanBakel H, Tsui K, Li J, Morris QD, Nislow C, Greenblatt JF, Hughes TR. 2014. A unified model for yeast transcript definition. *Genome Res*. 24:154–166.
Dimitrieva S, Bucher P. 2013. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Res*. 41:D101–D109.
Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* 489:101–108.
Doolittle F. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA*. 110:5294–5300.
Doolittle F, Brunet TDP, Linquist S, Gregory TR. 2014. Distinguishing between "Function" and "Effect" in genome biology. *Genome Biol Evol*. 6:1234–1237.
Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. *Curr Biol*. 22:R898–R899.
ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.
Fatica A, Bozzoni I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet*. 15:7–21.
Friedel M, Nikolajewa S, Sühnel J, Wilhelm T. 2009. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res*. 37:D37–D40.
Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. 2014. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* 159:647–661.
Grau J, Grosse I, Keilwagen J. 2015. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* 31:2595–2597.
Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of encode. *Genome Biol Evol*. 5:578–590.
Gulko B, Hubisz MJ, Gronau I, Siepel A. 2015. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet*. 47:276–283.
Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet*. 9:e1003569.
Harrow J, Frankish A, Gonzalez JM, Frazer KA. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 22:1760–1774.
Kabza M, Ciomborowska J, Makałowska I. 2014. RetrogeneDB—a database of animal retrogenes. *Mol Biol Evol*. 31:1646–1648.
Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet*. 10:19–31.
Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA*. 111:6131–6138.
Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GCM, Brown DL, Brudno M, Campbell J, et al. 2014. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*. 42:D966–D974.
Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H. 2015. Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. *Plant Cell* 27:2133–2147.
Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008.

Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 454:766.
Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet*. 10:155–159.
Neander K. 1991. Functions as selected effects: the conceptual analyst's defense. *Philos Sci*. 58:168–184.
Niu D-K, Jiang L. 2013. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem Biophys Res Commun*. 430:1340–1343.
Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res*. 37:e123.
Quek XC, Thomson DW, Maag JLV, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME. 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*. 43:D168–D173.
Quinn JJ, Chang HY. 2015. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet*. 17:47–62.
Ram O, Goren A, Amit I, Shoresh N, Yosef N, Ernst J, Kellis M, Gymrek M, Issner R, Coyne M, et al. 2011. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* 147:1628–1639.
Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 29:308–311.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, HillierL W, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15:1034–1050.
Smit AFA, Hubley R, Green P. 2013. RepeatMasker Open-4.0. Available from: http://www.repeatmasker.org.
Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, Lee BT, Learned K, Karolchik D, Hinrichs AS, et al. 2016. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res*. 44:D717–D725.
Strobl C, Hothorn T, Zeileis A. 2009. Party on! A new, conditional variable-importance measure for random forests available in the party package. *R J*. 1:14–17.
Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489:75–82.
Tsai ZT-Y, Shiu S-H, Tsai H-K. 2015. Contribution of sequence motif, chromatin state, and DNA structure features to predictive models of transcription factor binding in yeast. *PLoS Comput Biol*. 11:1–22.
van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol*. 8:e1000371.
Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*. 22:1798–1812.
White JK, Gerdin A-K, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ, Podrini C, et al. 2013. Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* 154:452–464.
Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucleic Acids Res*. 44:D710–D716.
Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci*. 67:43–62.