Method

Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes

Michael F. Lin,[1,2] Pouya Kheradpour,[1,2] Stefan Washietl,[2] Brian J. Parker,[3] Jakob S. Pedersen,[3,5] and Manolis Kellis[1,2,4,6]

[1]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [2]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA;* [3]*Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark;* [4]*The Broad Institute, Cambridge, Massachusetts 02139, USA*

The degeneracy of the genetic code allows protein-coding DNA and RNA sequences to simultaneously encode additional, overlapping functional elements. A sequence in which both protein-coding and additional overlapping functions have evolved under purifying selection should show increased evolutionary conservation compared to typical protein-coding genes—especially at synonymous sites. In this study, we use genome alignments of 29 placental mammals to systematically locate short regions within human ORFs that show conspicuously low estimated rates of synonymous substitution across these species. The 29-species alignment provides statistical power to locate more than 10,000 such regions with resolution down to nine-codon windows, which are found within more than a quarter of all human protein-coding genes and contain ~2% of their synonymous sites. We collect numerous lines of evidence that the observed synonymous constraint in these regions reflects selection on overlapping functional elements including splicing regulatory elements, dual-coding genes, RNA secondary structures, microRNA target sites, and developmental enhancers. Our results show that overlapping functional elements are common in mammalian genes, despite the vast genomic landscape.

[Supplemental material is available for this article.]

It is often assumed that synonymous sites within protein-coding open reading frames (ORFs) evolve neutrally, since mutations in them do not change the amino acid translation. But, in fact, ORFs in many species simultaneously encode additional functional sequence elements within the codon sequence, often with strong evolutionary constraint on the synonymous sites (Chamary et al. 2006; Itzkovitz and Alon 2007). For example, mammalian ORFs are known to encode exonic splicing enhancers and silencers (Chen and Manley 2009), microRNA target sites (Lewis et al. 2005; Hurst 2006), A-to-I recoding sites (Rueter et al. 1999; Bass 2002), and transcriptional enhancers (Lang et al. 2005; Nguyen et al. 2007; Lampe et al. 2008; Tümpel et al. 2008; Dong et al. 2010). Several previous studies have observed strong genome-wide trends toward increased evolutionary constraint on such overlapping functional elements, by averaging across many loci pooled together (Baek and Green 2005; Xing and Lee 2005; Chen et al. 2006; Down et al. 2006; Goren et al. 2006; Parmley et al. 2006; Robins et al. 2008; Kural et al. 2009). However, these averaging approaches generally did not have the power to locate individual overlapping functional elements within specific genes.

In this study, we use the unprecedented discovery power provided by alignments of 29 mammalian genomes to provide a systematic annotation of individual functional elements embedded within protein-coding regions throughout the human genome. Since the average codon site in these multiple sequence alignments shows about four synonymous substitutions, we predict that overlapping functional elements will individually stand out as short, localized regions with exceptionally few synonymous substitutions—in much the same way that widely used methods such as GERP, phastCons, phyloP, and SiPhy locate conserved functional elements within a background of neutral nucleotide-level sequence evolution (Cooper et al. 2005; Siepel et al. 2005; Margulies et al. 2007; Garber et al. 2009; Pollard et al. 2010).

However, detecting overlapping evolutionary constraints within protein-coding ORFs is more difficult than detecting general nucleotide-level constraints, for two main reasons. First, since the majority of nucleotide sites in a typical human ORF are already highly conserved among mammals due to their protein-coding function, we must expect less statistical power to detect increased conservation for overlapping sequence elements of a given length. Second, it is important to account precisely for the protein-coding constraints on each nucleotide site by modeling the evolutionary process at the codon level, rather than analyzing the conservation of individual nucleotide sites independently of those surrounding them.

To address these challenges, we present a novel adaptation of statistical phylogenetic codon models widely used in evolutionary analysis of protein-coding genes (for recent reviews, see Anisimova and Kosiol 2009; Delport et al. 2009), which locates short windows within alignments of known human ORFs showing significantly reduced rates of synonymous substitution. Applying this new method to the 29-species alignments, we confidently locate more than 10,000 such regions, typically with 70%–90% reduced synonymous rates, down to a resolution of just nine codons. These putative "synonymous constraint elements" contain only ~2% of

[5]**Present address: Department of Molecular Medicine, Aarhus University Hospital, Skejby, Brendstrupgaardsvej 100, DK-8200 Aarhus N, Denmark.**
[6]**Corresponding author.**
E-mail manoli@mit.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.108753.110. Freely available online through the *Genome Research* Open Access option.

21:1916–1928 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

all synonymous sites, but are found within more than a quarter of all human protein-coding genes. We also present numerous lines of evidence that they indeed play diverse functional roles in several biological processes such as splicing and translational regulation, dual-coding regions, RNA secondary structures, miRNA targeting, and developmental enhancers.

A few previous studies have also sought to locate individual overlapping functional elements in human genes based on their increased conservation, but their power was limited compared to what should now be possible by comparing the 29 mammals. Schattner and Diekhans (2006) analyzed pairwise alignments of human and mouse ORFs to identify about 200 regions of at least 60 codons tolerating at most one synonymous substitution. Similarly, many of the "ultraconserved elements" of Bejerano et al. (2004), stretches of at least 200 bp perfectly conserved between human and mouse, overlap known coding regions. Such pairwise species comparisons allow for much simpler statistical models, but their resolution of 180–200 nt seems so long (considerably longer than the typical exon length of ~120 nt) that many shorter overlapping functional elements were probably averaged out, even if they are highly conserved. Other related studies, including some multispecies approaches, have analyzed only a small fraction of mammalian genes (Hurst and Pál 2001; Chen and Blanchette 2007; Parmley and Hurst 2007b; Lin et al. 2008), while we undertake a comprehensive genome-wide analysis. Lastly, a few methods have been developed to identify examples of certain known classes of overlapping functional elements with predictable evolutionary signatures, including dual-coding ORFs (Chung et al. 2007; Ribrioux et al. 2008) and RNA secondary structures (Pedersen et al. 2004a,b). These are complementary to the rate-based approach we take here.

Estimating synonymous substitution rates in short windows within open reading frames

Our method uses phylogenetic codon models to find short windows within multi-species alignments of known ORFs that exhibit unusually low rates of synonymous substitution as measured by d_S, a composite rate commonly used to summarize the relevant parameters of such models (Yang and Bielawski 2000). Specifically, our method analyzes any window of adjacent codon sites within an alignment to compute the maximum likelihood estimate of a parameter λ_s, which is a scale factor on d_S, indicating how much slower or faster synonymous substitutions have occurred in that window relative to a null model representing typical protein-coding sequence evolution. For example, a particular window with $\lambda_s = 0.5$ is estimated to have evolved with a synonymous substitution rate only one-half that of the null model average (window $d_S = \lambda_s \times$ average d_S) (Fig. 1A). Our method is designed to estimate λ_s for short windows of adjacent sites—in this study, nine to 30 codons. (The method also estimates λ_n, the analogous parameter for non-synonymous substitutions, which we use to exclude potentially misaligned regions.)

Furthermore, we can associate a statistical significance with our estimate of λ_s for any window, using standard techniques for testing the goodness-of-fit of statistical phylogenetic models. Specifically, we can perform a likelihood ratio test (LRT) to assess precisely whether a small estimate of λ_s explains an observed alignment window better than the average rates assumed by the

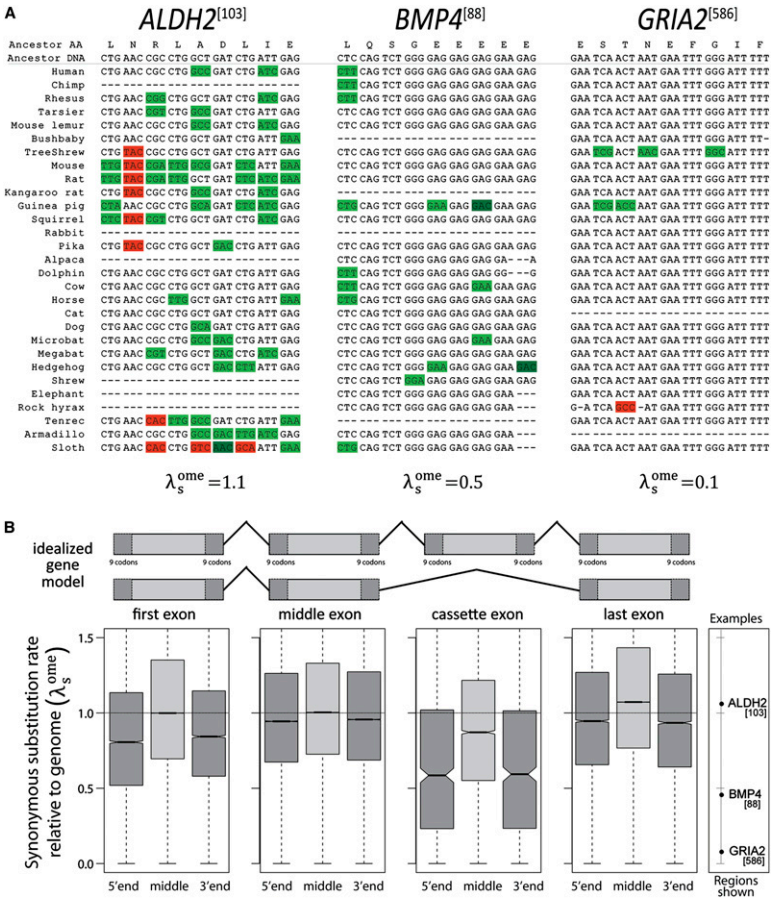

Figure 1. (*A*) Examples of local synonymous rate variation in alignments of 29 placental mammals for short nine-codon windows within the open reading frames (ORFs) of three known human protein-coding genes—*ALDH2*, *BMP4*, and *GRIA2*—with brackets denoting starting codon position within each ORF of shown alignment. (Bright green) Synonymous substitutions with respect to the inferred ancestral sequence; (dark green) conservative amino acid substitutions; (red) other nonsynonymous substitutions. The estimated parameter λ_s^{ome} denotes the rate of synonymous substitution within these selected windows relative to genome-wide averages. For example, the nine-codon window starting at codon 88 of the *BMP4* ORF shows $\lambda_s^{ome} = 0.5$, corresponding to an estimated synonymous substitution rate 50% below the genome average. (*B*) Variation in the estimated synonymous rate at different positions with respect to exon boundaries and translation start and stop, across all CCDS ORFs. For each class of regions, box-and-whisker plots show the observed distribution of λ_s^{ome}, including the median (*middle* horizontal bars), middle 50% range (boxes), extreme values (whiskers), and whether medians differ with high statistical confidence (nonoverlapping notches between two boxes). Estimated synonymous rates tend to be significantly reduced at the 5′ and 3′ ends of exons, and dramatically reduced in alternatively spliced exons, likely reflecting widespread splicing regulatory elements embedded within protein-coding regions.

null model. Then, if a window has a significantly reduced λ_s according to this test, we infer that its synonymous sites have probably been constrained by natural selection acting on an overlapping functional element. This is very similar to likelihood methods for nucleotide-level constraint detection (Garber et al. 2009; Pollard et al. 2010), extended to codon models so that we can disentangle the different evolutionary pressures on synonymous and non-synonymous sites.

The LRT provides an elegant way to avoid certain potential pitfalls in detecting individual regions with reduced synonymous rates. For example, it accounts for the uncertainty in rate estimates based on the exact set of informant species aligned for each window. To illustrate, consider two windows, one with estimated λ_s = 0.5 with all 29 species aligned, and the other with λ_s = 0.1 but only a human/chimpanzee alignment available. Even though the estimated λ_s is lower in the second window, it is almost surely less significant by the LRT, because it is based on far less informative data. The LRT also accounts for the expected constraint at each individual site based on the amino acid it encodes. For example, consider a hypothetical window coding exclusively for conserved methionine and tryptophan residues, which are encoded by nondegenerate codons (ATG and TGG, respectively). By definition, this window does not exhibit any synonymous substitutions and might therefore appear to have a very low synonymous substitution rate. But a reduced estimate of λ_s in this window would not be considered significant, because it does not provide a better explanation for the conservation of the nondegenerate sites.

We also designed our method to control for background variation in sequence composition and evolutionary rates across the genome (Lercher et al. 2001; Williams and Hurst 2002; Fox et al. 2008), as well as the possibility of selection on diffuse effects that can constrain all or most of an ORF, such as transcript structural stability or codon bias for translation efficiency (Chamary et al. 2006). To account for these, we evaluated each window against two null models, one representing the average sequence composition and evolutionary rates of the entire "ORFeome" (to obtain λ_s^{ome}), and the second estimated specifically from the individual ORF containing each window (to obtain λ_s^{ORF}). By identifying statistically significant rate reductions with respect to both null models, we required windows of interest to be exceptional with respect to genome-wide averages, on one hand, and also not explained by local biases in composition, rates, and codon usage, on the other hand.

Results

To annotate likely overlapping functional elements in human genes, we applied our local synonymous rate estimation procedure to sliding windows in all open reading frames in the human Consensus Coding Sequence (CCDS) catalog, a conservative set containing the ~85% of human gene annotations that are unanimously agreed on by the major gene catalogs (Pruitt et al. 2009). The vast majority of CCDS ORFs are aligned across at least 15 of the 29 placental mammals used in this study, with an average of four synonymous substitutions per codon site (Supplemental Fig. 1).

Genome-wide trends in synonymous rate variation

Before attempting to locate individual regions of synonymous constraint, we were immediately able to confirm a few notable genome-wide trends in synonymous rate variation that have been observed previously using different methods. Specifically, we observed marked reductions in average synonymous rates at the boundaries of coding exons, suggesting widespread evolutionary constraint on overlapping translation and splicing regulatory elements (Smith and Hurst 1999; Baek and Green 2005; Xing and Lee 2005; Chen et al. 2006; Parmley et al. 2006; Parmley and Hurst 2007a). For example, the median λ_s^{ome} estimate for the first nine codons following the start codon in each ORF is only 0.81, indicating that the first several codons in a typical mammalian gene appear to have tolerated synonymous substitutions at a rate 19% below average ($P < 10^{-15}$, Mann-Whitney U test) (Fig. 1B). The median estimated synonymous rate is also reduced by 13% in nine-codon windows spanning exon–exon junctions ($P < 10^{-15}$). Strikingly, the first nine to 12 codons of alternatively spliced cassette exons annotated in CCDS typically have an estimated synonymous rate 41% below average ($P < 10^{-15}$). We will revisit the possible translation and splicing regulatory roles suggested by these overall trends after establishing the statistical significance of the synonymous constraint in each individual window.

More than one-third of CCDS ORFs contain short windows with statistically significant synonymous constraint

We next applied LRTs to identify individual windows that show statistically significant evidence of reduced synonymous substitution rates (Fig. 2). Using three window sizes of nine, 15, and 30 codons, and sliding across each ORF by one-third of the window length, we selected windows passing three likelihood ratio tests, for the hypotheses that λ_s^{ome} is significantly below one, that λ_s^{ORF} is also significantly below one, and that λ_n^{ome}, the relative rate of nonsynonymous substitutions, is not significantly above one. We applied appropriate corrections for multiple testing (see Methods), and simulation and permutation benchmarks confirmed the robustness of our approach for detecting significantly reduced synonymous rates (Supplemental Material S3).

At the intermediate window length of 15 codons, 1.7% of the windows in CCDS ORFs meet these criteria. Overlapping significant windows collapse into 10,757 separate regions throughout human ORFs, covering 2.8% of all approximately 28 million CCDS coding nucleotide positions. More than one-third of CCDS genes (6033/16,939) contain at least one such region. Notably, although the test threshold is $\lambda_s^{ome} < 1$, the median λ_s^{ome} among the windows passing the test is only 0.23, corresponding to a 77% reduced rate of synonymous substitution compared to the genome-wide average. Furthermore, the estimates of λ_s^{ORF} in significant windows are also very low (best-fit line $\lambda_s^{ORF} = 0.78\lambda_s^{ome} + 0.06$ with R^2 = 0.81), confirming that these regions are generally not explained by ORF- or region-specific variation in sequence composition or evolutionary rates. Finally, the locally estimated synonymous and non-synonymous rates are not strongly correlated (Pearson coefficient between λ_s^{ORF} and λ_n^{ORF} of 0.04 in all windows, and 0.06 in significant windows), suggesting that our method largely succeeds in disentangling evolutionary pressures on synonymous and nonsynonymous sites, when controlling for regional biases in composition and rates. (Due to such biases, the estimates of λ_s^{ome} and λ_n^{ome} do correlate somewhat, with Pearson coefficient 0.24 in all windows and 0.22 in significant windows.)

Windows with significant synonymous constraints are not unusually enriched on any of the individual human autosomes, although they are about twofold depleted on chromosome 19 (Supplemental Table 1), which frequently stands out in genome-wide analyses owing to several unusual properties. In particular, the apparent depletion of synonymous constraints on this chromosome may be due to a calibration bias in our method arising

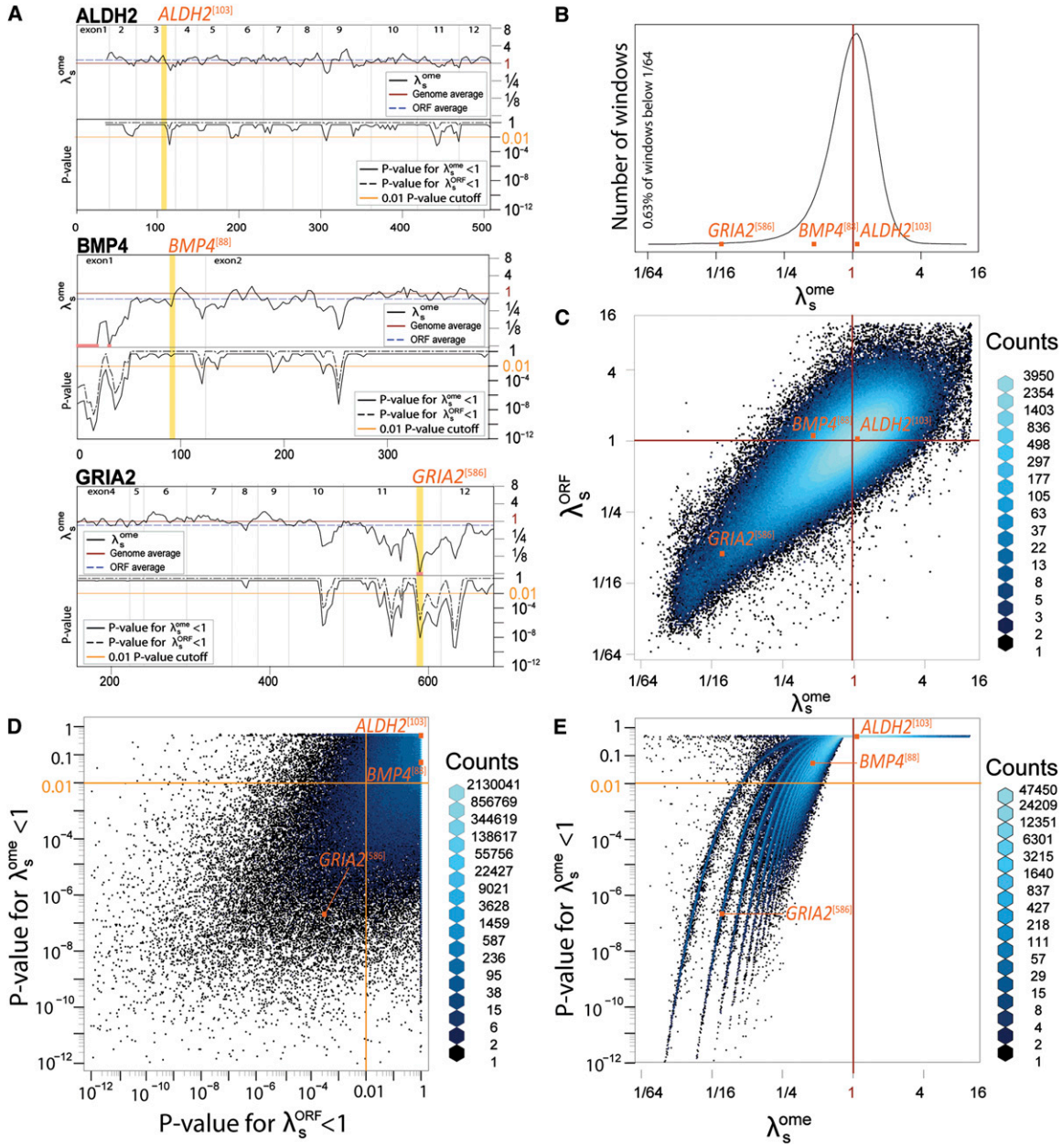

Figure 2. Identifying individual windows with statistically significant synonymous constraint. (*A*) Estimated synonymous rate relative to genome average (λ_s^{ome}) and corresponding *P*-value for the hypothesis $\lambda_s^{ome} < 1$ evaluated in nine-codon windows along the entire protein-coding regions of *ALDH2*, *BMP4*, and *GRIA2*, highlighting the windows corresponding to the three examples in Figure 1. For each plot, the *top* portion shows the λ_s^{ome} estimate for each window (black curve), the genome average (red line at $\lambda_s = 1$), and the ORF average (blue dashed line). The *bottom* portion shows the statistical significance of the reduction in the synonymous rate estimate in each window, accounting for evidence in the cross-species alignments, using a likelihood ratio test for the hypothesis $\lambda_s^{ome} < 1$ (continuous black curve, using the genome average as the null model), and for the hypothesis $\lambda_s^{ORF} < 1$ (dashed black curve, using the ORF average as the null model). (Vertical gray lines) Exon boundaries; (orange) regions where λ_s^{ome} drops below 1/16th toward the 5′ end of *BMP4* and the 3′ end of *GRIA2*. (*B*) Overall distribution of λ_s^{ome} estimates for all nine-codon windows across all CCDS genes. Heavy *left* tail indicates an excess of windows with very low estimated synonymous rates, shifting the mean ($\lambda_s^{ome} = 1$) to the *left* of the distribution mode, which likely represents neutral rates. (*C*) Comparison of synonymous rates estimated relative to genome-wide (λ_s^{ome}) and ORF-specific (λ_s^{ORF}) null models, each point denoting one nine-codon window, and density of overlapping points denoted by color. Joint distribution shows that low λ_s^{ome} estimates also usually correspond to low λ_s^{ORF} estimates, and therefore that the heavy tail observed in *B* does not reflect regional or ORF-wide deceleration, but instead localized constraints in small windows within each ORF, also visible in the three examples of *A*. (*D*) Comparison of *P*-values for synonymous rate reduction with respect to genome-wide (*y*-axis) and ORF-specific (*x*-axis) null models. Candidate synonymous constraint windows are selected when synonymous rate reductions are significant at $P < 0.01$ with respect to both null models (orange lines). Note that many windows are significant with respect to one null model but not the other. (*E*) Correspondence between λ_s^{ome} and the associated significance estimate for the each nine-codon window. The visible stripes in this plot arise from windows that are perfectly conserved except for one, two, three, or more synonymous substitutions observed in the extant species, while the position along each stripe reflects variation in the λ_s^{ome} estimate and its significance, determined by the species coverage, codon composition, and observed codon substitutions in each window. (*B–E*) The three example regions highlighted in *A* are shown in each distribution and density plot, with horizontal and vertical axes aligned. The orange line in plots *A*, *D*, and *E* denotes the statistical significance cutoff of $P < 0.01$, and the red line in plots *A*, *B*, *C*, and *E* denotes the genome-wide average $\lambda_s^{ome} = 1$ and $\lambda_s^{ORF} = 1$ for *B*. The *ALDH2*[103] synonymous rate is not significantly reduced either relative to the genome or to the *ALDH2* ORF; *BMP4*[88] is reduced relative to the genome but not relative to its ORF, which shows an overall reduced rate; *GRIA2*[586] is >80% reduced relative to both the genome and its ORF, resulting in significant *P*-values for both.

from the chromosome's above-average G+C content and mutation rates (Lercher et al. 2001; Castresana 2002), or it could reflect a genuine biological tendency related to the chromosome's unusually large complement of genes from a few tandem families (Grimwood et al. 2004). We analyzed ORFs on the X chromosome using a null model estimated from coding sites on that chromosome only and found that its resulting proportion of significant windows is lower than most autosomes, but still much greater than chromosome 19. We did not analyze the fewer than 100 protein-coding genes on the Y chromosome due to their eccentric, fast-evolving properties (Hughes et al. 2005, 2010; Kuroki et al. 2006).

In the longer windows of 30 codons, our method has increased statistical power to detect synonymous constraints since it combines evidence from more sites, and a larger proportion of windows reach significance with somewhat higher typical synonymous rate estimates, although they collapse into fewer separate regions. Conversely, a smaller proportion of the shorter nine-codon windows reach significance, with even lower estimated synonymous rates (Table 1). We also attempted our analysis with even smaller windows of six and three codons, but vanishingly few reached significance. Evidently, in these very short windows, even perfect conservation across the available species is usually not adequate to infer synonymous constraints using our current methodology and alignments.

Since the three window sizes lead to different trade-offs between resolution and discovery power, it is reasonable to expect them to identify somewhat different sets of regions as significant. In fact, of the regions obtained by collapsing overlapping significant windows at the 15-codon resolution, 24% are not detected at either the longer or shorter resolution. Similarly, 33% of the 30-codon regions and 28% of the nine-codon regions are detected only at those resolutions. As expected, the intermediate 15-codon resolution has the most overlap with the others, including 67% of the 30-codon and 72% of the nine-codon regions.

Hereafter, we refer to the collapsed significant regions as "synonymous constraint elements" (SCEs) and undertake numerous downstream analyses to show that they correspond to overlapping functional elements with diverse biological functions. We will perform most of these analyses based on SCEs identified at the 15-codon resolution, since they include most of the other sets, but we will also use the nine-codon and 30-codon resolutions based on the expected length of different types of overlapping functional elements. Similarly, we expect that each can be useful in different contexts for future follow-up studies.

Table 1. Sliding windows in CCDS ORFs were tested for significantly reduced synonymous substitution rate estimates at different resolutions and stringencies

Resolution (window size)	Short (9 codons)	Intermediate (15 codons)	Long (30 codons)
No. of windows tested	2,915,773	1,727,202	842,475
Proportion of windows significant	1.04%	1.72%	2.75%
Median significant λ_s^{ome}	0.1361	0.2299	0.3584
Median significant λ_s^{ORF}	0.1443	0.2572	0.4120
Maximum significant λ_s^{ome}	0.5497	0.6323	0.7134
Maximum significant λ_s^{ORF}	0.4682	0.5805	0.6520
No. of nonoverlapping synonymous constraint elements	11,882	10,757	8933
Proportion of 27,812,282 CCDS nucleotide positions within a synonymous constraint element	1.79%	2.82%	4.48%
Proportion of 16,939 CCDS ORFs containing a synonymous constraint element	35.8%	35.6%	33.3%

Sequence composition and codon usage in SCEs

Since a major design goal for this study was to control for the specific codon sequence in each window, we thoroughly examined the composition of the regions reported as significant (Supplemental Material S6). Briefly, the SCEs exhibit certain biases in nucleotide, dinucleotide, and amino acid composition, for example, slightly below-average G+C content, but these biases are within the range of variation seen between protein-coding regions from different parts of the genome. Other compositional properties allowed us to rule out certain possible artifactual explanations for the low divergence in SCEs, including tandem and microsatellite repeats, biased conversion in recombination hotspots, and codon usage bias. In particular, the effective number of codons (ENC) (Wright 1990; Fuglsang 2006) shows that there is slightly less codon usage bias in SCEs compared to other coding regions, indicating that our method excluded regions explained by this effect, as intended. Overall, our analysis of sequence composition and codon usage did not suggest any debilitating shortcomings of our overall approach, and it is likely that the compositional differences that are seen largely reflect the sequence-dependent biological nature of the overlapping functional elements encoded by the SCEs.

Characteristics of genes containing SCEs

We next studied overall properties of the 6033 genes containing SCEs (at 15-codon resolution). Compared to the remaining CCDS genes, the typical gene containing an SCE has a much longer ORF (median 558 vs. 356 codons). This is not actually longer than expected based on drawing genes randomly weighted by their ORF length, but the genes containing SCEs also have more introns (median nine vs. five), lengthier individual introns (1727 nt vs. 1261 nt), and they span much larger genomic regions (36,000 nt vs. 11,000 nt), suggesting that the overall length distribution is entangled with the well-established correlations between gene length and other relevant characteristics including conservation, functional categories, and expression levels (Supplemental Material S7; Castillo-Davis et al. 2002; Urrutia and Hurst 2003; Stanley et al. 2006; Pozzoli et al. 2007). The genes containing SCEs also appear to be under stronger purifying selection on their amino acid sequences, as the median estimate of $\omega = d_N/d_S$ measured across each complete ORF is 0.068, much lower than the 0.138 for other genes, despite containing regions with greatly reduced d_S.

Next, we analyzed Gene Ontology (GO) annotations for the genes containing SCEs. While 36% of CCDS genes contain a 15-codon SCE, they include 70% genes annotated with the term "chromatin modification," a twofold enrichment (Bonferroni-corrected hypergeometric $P < 7.9 \times 10^{-13}$). Additionally, they include most of the genes in these and related categories: "ubiquitin-protein ligase activity" (1.8-fold, $P < 2.9 \times 10^{-6}$), "ion channel complex" (1.7-fold, $P < 5.6 \times 10^{-5}$), "nervous system development" (1.6-fold, $P < 6.8 \times 10^{-9}$), "transcription factor activity" (1.6-fold, $P < 3.5 \times 10^{-6}$), and "RNA splicing" (1.5-fold, $P < 8.4 \times 10^{-4}$). These enrichments remain strongly significant when controlling for the varying ORF lengths (Supplemental Table 4) and suggest a few interesting hypotheses

about genes that encode overlapping functional elements. For example, the enrichment for genes encoding chromatin modification and RNA splicing functions could suggest the existence of autoregulatory circuits for many such genes, similar to known examples such as ADARB1, which edits its own pre-mRNA and causes a change in its splicing (Rueter et al. 1999), and DGCR8, which binds its own mRNA and causes it to be cleaved by Drosha (also know as RNASEN) (Han et al. 2009). Also, the enrichment for ion channel genes suggests a connection with A-to-I editing, since several such genes are known targets of this recoding mechanism (Bass 2002); we explore this further below.

More than one-third of short SCEs can be provisionally assigned roles in transcript splicing or translation initiation

As expected, based on the aforementioned general trends in synonymous rate variation (above; Fig. 1B), many SCEs can be provisionally classified as possible splicing regulatory elements based on their location within gene models. In particular, 34.7% of the nine-codon SCEs span an exon–exon junction, compared to only 20.3% of a set of random control regions placed uniformly throughout CCDS ORFs, with matching length distribution and total number.

Interestingly, the introns flanked by these SCEs tend to have weaker 3′ splice acceptor sites (nine-codon resolution; $P = 3.9 \times 10^{-7}$, Mann-Whitney U test), based on analysis of the sequence information content of the surrounding nucleotides (Yeo and Burge 2004). This is consistent with the hypothesis that while "strong" splice sites are constitutively recognized by the splicing machinery, the activity of weaker splice sites is more reliant on additional nearby *cis*-regulatory sequences, increasing the possibilities for their combinatorial and condition-specific regulation (Fairbrother et al. 2002; Chen and Manley 2009). The SCEs spanning exon–exon junctions also show an increased frequency of exonic splicing enhancer motifs compared to other coding regions, although this is difficult to distinguish from correlated compositional biases (Supplemental Material S8).

We also examined individual exons with multiple alternative acceptor or donor sites, for which the coding sequence of the longest exon isoform additionally encodes splice sites for shorter isoforms and perhaps additional splicing regulatory elements (e.g., Fig. 3A). Of 551 such alternative donor sites in RefSeq transcripts encoded within the CCDS exons we analyzed, 84 (15.2%) fall within SCEs, compared to only 1% in the random regions ($P < 10^{-21}$). Similarly, 57 of 576 (9.9%) alternative acceptor sites lie within SCEs (1.7% random; $P < 10^{-21}$).

SCEs are also enriched for elements potentially involved in translation initiation. The first nine-codon window (beginning at the first site following the start codon) in 3.9% of CCDS ORFs is found to be under synonymous constraint, a strong enrichment compared to 1.04% of all windows. Additionally, there are 744 RefSeq-annotated internal translation initiation sites encoded within longer coding exons of CCDS ORFs, of which 5.4% fall within SCEs (e.g., Fig. 3B) compared to 1.6% in the random regions ($P < 10^{-9}$).

Taken together, these provisional classifications of possible splicing and translation regulatory elements account for slightly more than one-third of the SCEs (at nine-codon resolution). Conversely, nearly two-thirds are not found in locations that directly suggest such roles, although it should be noted that some splicing regulatory sequences may act at considerable distances (Parmley and Hurst 2007a; Parmley et al. 2007).

Synonymous constraint at an alternate translation start site in BRCA1

One noteworthy example of an SCE with a possible role in translation initiation is found within the tumor suppressor gene *BRCA1*. Hurst and Pál (2001) first observed that an extended region within this ORF shows unusually low synonymous substitution rates, based on sliding windows of 100 codons in pairwise comparisons among the human, mouse, and rat orthologs. Recently, however, Schmid and Yang (2008) raised certain issues with their statistical methods and argued that their result was artifactual.

Our analysis is based on much more data than both previous studies and strongly supports the original conclusion of Hurst and Pál (2001). Within the most significant 15-codon window in *BRCA1*, the estimated rate of synonymous substitution in placental mammals is reduced by 80% ($\lambda_s^{ome} = 0.20$, $P < 9.9 \times 10^{-8}$), ranking among the slowest 1% of windows in human ORFs, and slower than average for SCEs at this resolution. Like all SCEs, the window also has a very low synonymous rate estimate compared to the *BRCA1* ORF specifically ($\lambda_s^{ORF} = 0.23$, $P < 2.3 \times 10^{-5}$). Due to the controversy over this region in particular, we also performed auxiliary permutation tests that further confirmed its statistical significance (Supplemental Material S4).

Furthermore, the much higher resolution of our analysis precisely localizes the significant region of synonymous constraint to the annotated translation initiation site of a CCDS-supported alternative splice form of *BRCA1* (Fig. 3B), immediately suggesting a hypothesis for an overlapping biological function—namely, a role in regulating translation initiation in this splice form. This highly suggestive positional association may have been much less clear at the 100-codon resolution used by both previous studies. Indeed, the synonymous constraint is not significant at the 30-codon resolution according to our analysis, possibly corroborating the statistical concerns raised by Schmid and Yang (2008), while nonetheless confirming and extending the main conclusion of Hurst and Pál (2001).

Enrichment of miRNA target sequences in SCEs

We next sought evidence that SCEs capture embedded miRNA target sites, since previous studies have demonstrated trends toward their preferential conservation in mammalian ORFs (Lewis et al. 2005; Hurst 2006; Kural et al. 2009). Although the main sequence determinant of miRNA targeting is a "seed" of only ~7 nt, which is much shorter than our present resolution, it is still reasonable to expect SCEs to capture at least some of these sites, due to additional flanking positions influential in targeting, or to multiple closely spaced target sites acting synergistically (Grimson et al. 2007).

Indeed, SCEs show a small but significant enrichment for the 7-nt target seed sequences of known human miRNAs: 8.9% of positions in SCEs start with one of these 7-mers, compared to 8.6% in all coding regions ($P < 10^{-7}$; nine-codon resolution). They are even more frequent in SCEs lying within the last coding exon of each ORF (9.3% of positions; $P < 10^{-5}$), consistent with the trend for miRNA target sites to appear toward the 3′ end of animal transcripts (Lewis et al. 2005). Matched random control motifs showed weaker or no enrichment, indicating that the observed enrichment is not explained by correlated sequence composition effects (Supplemental Table 5). Overall, while the excess of miRNA seeds in SCEs provides further evidence that many mammalian ORFs encode conserved target sites, greater power and resolution will be needed to precisely annotate them.

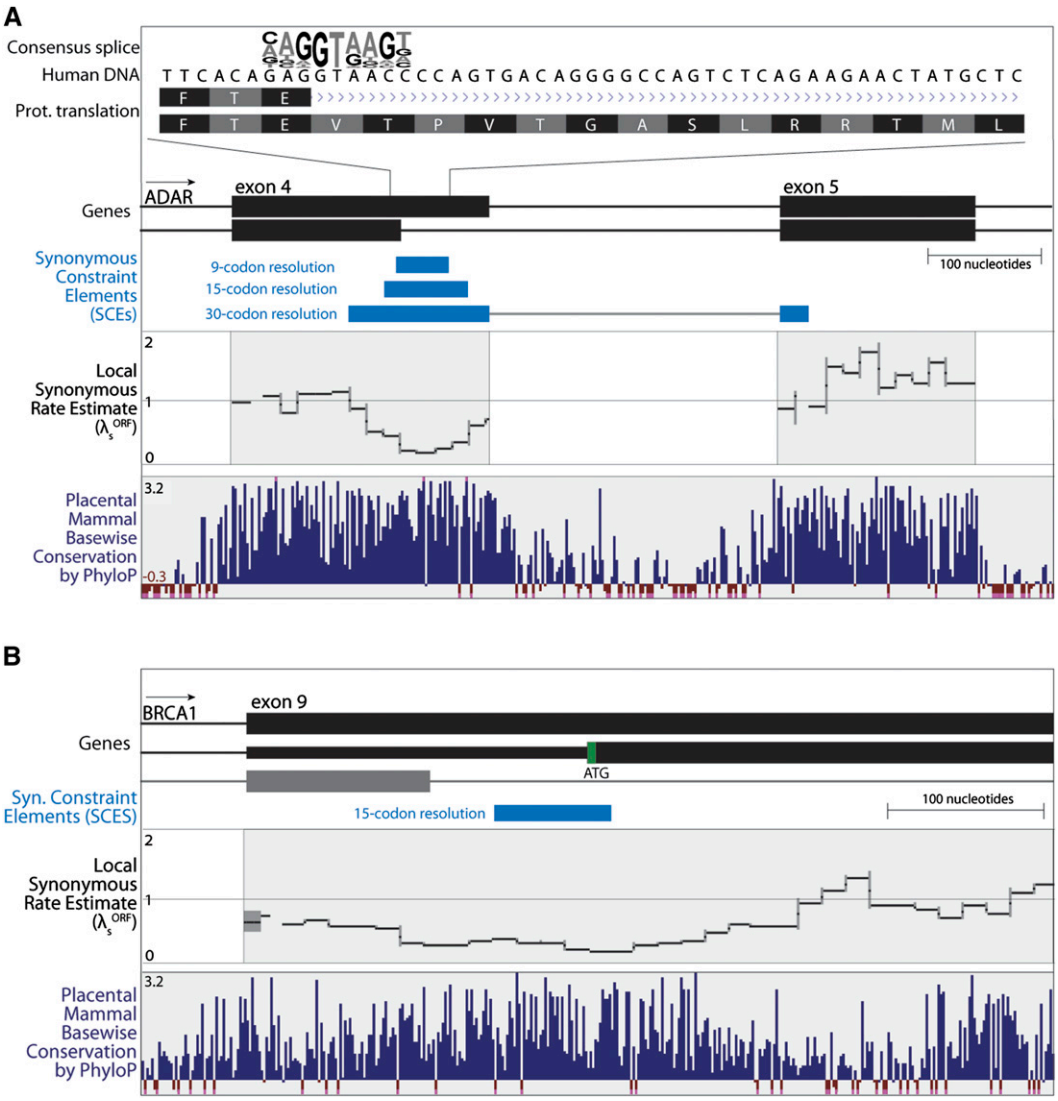

Figure 3. Examples of candidate synonymous constraint elements (SCEs) with likely roles in splicing and translation regulation. (*A*) Predicted SCEs (light blue) overlapping two isoforms of *ADAR* exon 4 (black) arising from an alternative splice donor site encoded within the longer exon variant. With increasing resolution, the SCE is more precisely localized to the region of overlap with the alternative splice site (motif logo for human donor sites rendered by WebLogo) (Crooks et al. 2004). The localization of the synonymous constraint to the splice site is also seen in the local synonymous rate estimate λ_s^{ORF} (relative to the ORF average). Note that the significant reduction in the synonymous rate is not obvious from the nucleotide-level conservation measure (dark blue, *bottom* panel). The extent of the predicted SCE may suggest the presence of additional splicing regulatory elements downstream from the alternative splice site. (*B*) Predicted SCE (light blue) overlapping an alternate translation initiation site (green) in *BRCA1* encoded within exon 9 of a longer isoform. Synonymous constraint ranges from shortly upstream to immediately downstream of the alternate start codon, suggesting this region may be involved in regulating translation initiation at the alternate site. The region just upstream of the predicted SCE also shows a reduced synonymous rate (black curve) overlapping an alternative splice donor site for a third *BRCA1* isoform (gray), although this reduction is not statistically significant and the third isoform is weakly supported. Annotation visualizations in Figures 3 and 4 are based on the UCSC Genome Browser (Kent et al. 2002).

SCEs in known and novel dual-coding genes

Genomic sequences can simultaneously encode different amino acid sequences in multiple reading frames, a common phenomenon in viral genomes but rare in animal genomes. Such "dual-coding" regions can involve ORFs on the same strand, but in an alternate "shifted" reading frame, which can be mediated by alternative splicing, internal translation initiation, or ribosomal frameshifting. Of the six long human dual-coding gene structures for which likely biological functions have been demonstrated (Sharpless and DePinho 1999; Klemke et al. 2001; Yoshida et al. 2001; Hameed et al. 2003; Poulin et al. 2003; Ahmed et al. 2008), all show at least some evidence of overlapping evolutionary constraint in our analysis: *XBP1*, *GNAS*, and the *ANKHD1/EIF4EBP3* fusion transcript contain SCEs in their dual-coding regions; the dual-coding 3′ end *IGF1* narrowly missed our threshold with a 47% reduced synonymous rate; and the dual-coding regions of *CDKN2A* and *LRTOMT* have very low synonymous rates, but were excluded from SCEs because of elevated nonsynonymous rates. Aside from these individually studied examples, CCDS annotates 237 other exons as protein-coding in multiple reading frames of the same strand, 24 (10.1%) of which contain SCEs, compared to six (2.5%) containing random control regions. This lower overlap might

suggest dual-coding regions not under selection across placental mammals.

Alternatively, both strands of the genomic DNA can encode different protein sequences, expressed in "sense" and "antisense" transcription units. At least one such case has been thoroughly studied: an ~200-nt sense/antisense dual-coding sequence of the convergent transcription units for *THRA* and *NR1D1* (Hastings et al. 2000). Indeed, we find synonymous constraints in both ORFs precisely coinciding with the known dual-coding region (Fig. 4A). We similarly detect synonymous constraints in 10 of 44 individual exons that CCDS annotates on both strands (five in random regions).

In addition to these known examples, we found 19 candidate novel dual-coding ORFs in alternate reading frames of known CCDS ORFs, which contain one or more of our 30-codon SCEs, are longer than expected by chance, and also appear to be depleted of

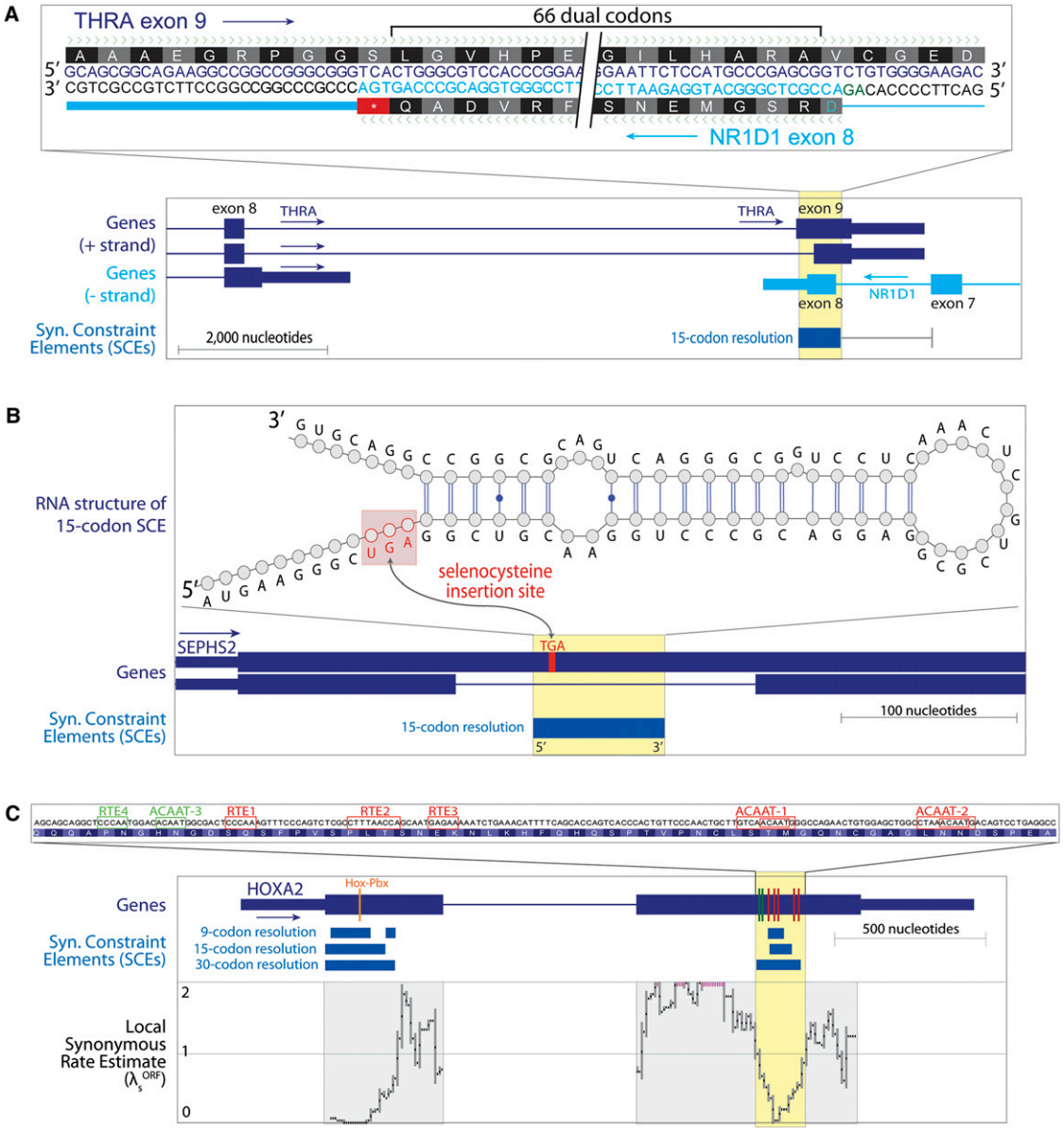

Figure 4. Synonymous constraint elements (SCEs) corresponding to dual-coding, selenocysteine insertion, and expression enhancer functions. (*A*) A large SCE (blue) fully encompasses a 66-codon sense/antisense dual-coding region in the convergent transcripts of *THRA* and *NR1D1*. The SCE is specifically localized to the overlapping exons, while upstream exons of each gene are excluded. (*B*) A predicted SCE in the selenoprotein-encoding gene *SEPHS2* encompasses the selenocysteine insertion site (red) and a predicted RNA hairpin structure (minimum free energy fold rendered by VARNA) (Darty et al. 2009) immediately downstream from the selenocysteine codon. Inferred structure is similar to a hairpin known to stimulate selenocysteine recoding in *SEPN1* (Howard et al. 2005). (*C*) Two SCEs are found within the *HOXA2* ORF, each corresponding to a different enhancer element regulating expression of the mouse ortholog in distinct segments of the developing hindbrain. The 5′ element encodes a HOX-PBX responsive element and drives expression in rhombomere 4 (Lampe et al. 2008), and the 3′ element encodes SOX2 binding sites and drives expression in rhombomere 2 (Tümpel et al. 2008). The 3′ element includes several RTE and ACAAT motif instances that were investigated by site-directed mutagenesis in the previous study (red), as well as two additional upstream instances (green). SCEs are also found within most other *HOX* genes.

stop codons in the other mammals. Twelve of these 19 are encoded on the same strand as the CCDS gene structure, some spanning multiple exons, and the remaining seven are found antisense to individual CCDS exons. A few are further supported by transcript cDNA evidence or similarity to known proteins (Supplemental Material S9). This preliminary assessment suggests that the SCEs probably capture several additional long dual-coding mammalian gene structures, although specialized methods for detecting the evolutionary signatures unique to dual-coding regions (Chung et al. 2007; Ribrioux et al. 2008) would probably have more power applied to the new set of 29 species.

SCEs capture most known A-to-I recoding sites

A-to-I editing is a recoding mechanism in which certain adenosine (A) bases in RNA transcripts are edited to inosine (I), which is read as a guanosine (G) (Bass 2002). This mechanism is essential for normal development of the mammalian nervous system, and, because the enzymes that mediate the reaction target double-stranded RNA, known A-to-I recoding sites conserved between human and mouse transcripts typically show extensive conservation of flanking sequence, presumably reflecting the interlocking constraints of encoding amino acids and pairing with another portion of the transcript (Aruscavage and Bass 2000; Hoopengardner et al. 2003).

Indeed, 10 of 14 known human A-to-I recoding sites in CCDS ORFs lie within SCEs (15-codon; none within random regions) (Supplemental Table 7), although this is not surprising since exceptional conservation was one signature originally used to identify many of the known sites. A recent human-specific study used a high-throughput sequencing approach not biased for highly conserved regions (Li et al. 2009) to identify 40 new edited sites within CCDS ORFs, only three of which also lie within our SCEs (two amino acid changing sites in *CADPS*, *FLNB*, and a synonymously edited site in *GRIA2*) (Supplemental Table 8). Therefore, consistent with that study's report that the 37 remaining experimentally identified sites lack extensive nucleotide-level conservation, we do not find specific evidence for synonymous constraints in other mammals.

SCEs and RNA secondary structures

In addition to the paired structures associated with A-to-I editing sites, the SCEs include several other striking examples of RNA secondary structures embedded within mammalian ORFs. For example, the selenoprotein-encoding gene *SEPHS2* contains a hairpin immediately downstream from its selenocysteine insertion site (Fig. 4B). This structure is likely to stimulate selenocysteine recoding, based on similar known structures in other genes (Howard et al. 2005; Pedersen et al. 2006). An SCE within *TTN*, the human gene with the most exons (313) encoding a protein with numerous functions in striated muscle and associated with several diseases, contains a hairpin showing both compensatory double substitutions and a compensatory deletion that preserve the paired structure in other vertebrates (Supplemental Fig. 2), strong evidence of selection on the RNA secondary structure. An SCE within *QKI*, which encodes an RNA-binding protein, appears to contain an instance of the protein's own binding motif followed by a hairpin showing a compensatory insertion, perhaps suggesting an autoregulatory mechanism (Supplemental Fig. 3).

We also attempted to study overall mutual enrichments between SCEs and computationally predicted RNA secondary structures using EvoFold (Pedersen et al. 2006; Parker et al. 2011) and RNAz (Gruber et al. 2010), but unfortunately this was confounded by correlations in sequence composition and conservation influencing both types of predictions (Supplemental Material S11). Careful further investigation of this topic is warranted, given previous studies suggesting that many additional RNA structures (Chamary and Hurst 2005; Shabalina et al. 2006; Tuller et al. 2010) and perhaps RNA–RNA interaction sites (Wang et al. 2008) are embedded within protein-coding regions.

Possible roles in exclusion of nucleosomes from certain exons

It was recently shown that nucleosomes preferentially localize within human exons compared to surrounding intronic regions (Schwartz et al. 2009; Tilgner et al. 2009). Two compositional properties of the SCEs suggest a possible relationship with this phenomenon. First, the general enrichment of nucleosomes within exons was shown to positively correlate with G+C content, while the SCEs have slightly lower G+C content than other coding regions (50.3% vs. 52.0%). Second, it is known that nucleosomes especially avoid contiguous stretches of adenine:thymine base pairs, and synonymous codon usage in many species is biased to avoid such "poly(dA:dT) tracts" (Cohanim and Haran 2009; Segal and Widom 2009). The SCEs have a 20% higher frequency of poly(dA:dT) tracts of 5 bp or longer compared to other coding regions.

Based on these properties, we would predict that exons containing SCEs tend to have lower nucleosome occupancy than other exons. We analyzed a high-throughput sequencing data set for nucleosome occupancy in human CD4+ T-cells (Schones et al. 2008) and found that the exons containing SCEs are, indeed, depleted for reads of nucleosome-bound DNA compared to other CCDS exons ($P = 1.0 \times 10^{-11}$, Mann-Whitney U test) (Supplemental Material S12).

Since the biological significance of nucleosome positioning within exons is not yet well understood, we cannot exclude the possibility that this relative depletion could just be a side effect of compositional biases in the SCEs. It is also possible, however, that some SCEs are under selection for sequence-dependent roles in excluding nucleosomes from certain exons, not unlike the "nucleosome-free regions" thought to facilitate chromatin access for regulatory factors near promoters (Schones et al. 2008; Warnecke et al. 2008; Washietl et al. 2008).

SCEs lie within most *HOX* genes and include two known developmental enhancers

Many of the lengthiest and most strikingly conserved SCEs are found within 27 of the 40 genes in the four *HOX* clusters. For example, the first 60 codons of *HOXB5* exhibit absolutely no synonymous substitutions in any of the species in our alignment of its ORF. More generally, we identify SCEs in eight of the 11 genes in the *HOXA* cluster, seven of nine *HOXB* genes, seven of nine *HOXC* genes, and three of nine *HOXD* genes, as well as the *EVX1* and *EVX2* homeobox-encoding genes adjacent to the *HOXA* and *HOXD* clusters. Lin et al. (2008) also noted striking regions of synonymous constraint in *HOX* genes, which they defined as stretches of at least 40 codons without any synonymous substitutions in pairwise comparisons. Our results confirm their findings while also providing much greater power and resolution, locating several additional shorter and/or less extremely conserved SCEs.

Remarkably, the two SCEs found within *HOXA2* correspond to known tissue-specific enhancers that regulate expression in distinct segments of the developing mouse hindbrain (Fig. 4C). A lengthier region (~200 bp) in the upstream exon encodes a HOX-PBX

responsive element and drives *Hoxa2* expression in rhombomere 4 (Lampe et al. 2008), and a shorter region (~75 bp) in the downstream exon encodes SOX2 binding sites and drives expression in rhombomere 2 (Tümpel et al. 2008). Considering these two examples, the SCEs could suggest the existence of a largely unknown regulatory network relying on nucleotide sequence elements embedded within the ORFs of most of these key developmental genes (Woltering and Duboule 2009).

Discussion

In this study, we showed that thousands of human protein-coding genes contain short regions with conspicuously low estimated synonymous rates across placental mammals. We located between 9000 and 12,000 SCEs, depending on the resolution chosen, within one-third of all CCDS ORFs (Table 1), or well over one-quarter of all human protein-coding genes. Our preliminary results implicate many of these regions in biological roles including tissue-specific developmental enhancers, translation regulatory elements, RNA structures involved in A-to-I editing or selenocysteine incorporation, and different classes of post-transcriptional regulatory elements including miRNA targeting and alternative splicing. Still, these provisional explanations do not even account for half of all the SCEs, suggesting that there may be many other overlapping biological roles yet to be elucidated. Therefore, just as our view of nucleotide-level conserved elements throughout the genome has been greatly refined since the first human/mouse comparisons nearly a decade ago, we expect that this initial survey of SCEs can motivate many additional computational and experimental studies to reveal their biological functions.

In addition to identifying candidate overlapping functional elements, our extensive annotation of synonymous sites under selection in placental mammals can help refine many types of evolutionary and functional analyses that typically assume they are neutral. For example, widely used methods for detecting positive and negative selection on amino acid sites are based on the ratio of nonsynonymous and synonymous substitution rates ($\omega = d_N/d_S$). Since the extreme drop in d_S observed in SCEs would naturally tend to elevate local estimates of ω, any claims of positive selection on the amino acid sites encoded within SCEs ought to be regarded with caution (Xing and Lee 2006; Parmley and Hurst 2007b). Our results can also inform future disease association studies and other types of population genetics analyses, which frequently ignore synonymous SNPs (Chamary et al. 2006). The SCEs we identified represent specific regions in which synonymous SNPs may well have significant consequences and should therefore be included in such analyses.

One major challenge in the design of our study, shared with nucleotide-level constraint detection methods, laid in the estimation of null models. Nucleotide-level methods typically calibrate their null models to presumptively neutral regions such as ancestral repeats or fourfold degenerate sites, but these were obviously not suitable for our purposes because our null models must capture the typical evolutionary rates of both synonymous and non-synonymous sites in coding regions. Therefore, we simply calibrated our null models to the ORFeome-wide or ORF-level background, averaging in any unusually evolving sites. If we assume that purifying selection is much more common than positive selection in synonymous sites (Resch et al. 2007), then this approach leads to a somewhat conservative test for synonymous constraint, providing one of our countermeasures against the possibility of background rate variation leading to spurious inferences of selection. More accurate null model calibration is an important direction for future investigation, perhaps by explicitly modeling statistical distributions of synonymous rates (Pond and Muse 2005; Rodrigue et al. 2008).

Still, despite our prudently conservative null model calibration, we were able to achieve far greater discovery power than previous efforts to locate regions of synonymous constraint in mammalian genes (Schattner and Diekhans 2006; Parmley and Hurst 2007b), which identified at most ~2% of our SCEs, and at much lower resolution. This is attributable both to the many informant species now available and to the rigorous phylogenetic methodology we devised to take advantage of them, based on maximum likelihood estimates of the synonymous substitution rates in short windows and formal statistical tests for their reduction. Naturally, this methodology can accommodate additional sequenced genomes and improved assemblies and alignments as they become available—perhaps eventually enabling systematic resolution of lineage-specific overlapping functional elements and individual binding sites for miRNAs and regulatory factors.

Methods

Genome annotations and alignments

This study was based on ORF annotations from the 2009-03-27 build of CCDS (Pruitt et al. 2009) for NCBI version 36 of the reference human genome assembly. When CCDS annotates multiple isoforms of a single locus (as defined by multiple CCDS IDs in overlapping chromosomal regions sharing the same HGNC gene symbol), only the isoform with the longest coding sequence was analyzed. We extracted the alignments for these ORFs from the MULTIZ whole-genome alignments of 44 vertebrate species, generated by UCSC Genome Bioinformatics and used throughout the initial analysis of the 2× mammals data set. Only the "rows" of the MULTIZ alignments corresponding to the 29 available placental mammals were used, and the whole-genome alignments were "spliced" as necessary to produce an alignment of the complete human ORF.

Estimating codon substitution rates in short windows

Our relative rate estimation procedure uses, as a parameter, any standard phylogenetic codon model $M = \langle T,Q \rangle$ where T specifies the topology and branch lengths of a phylogenetic tree and Q is a reversible 61×61 rate matrix describing codon evolution as a stationary, homogeneous, continuous-time Markov process, such that the transition probability matrix for any branch with length t is given by $P = \exp(Qt)$. Given such a model, the probability of any alignment of extant sequences can be computed using Felsenstein's algorithm, assuming independence of the codon sites and using the equilibrium frequencies of the codons implicit in Q as the prior distribution over the root. Nonaligned, gapped, or stop codons in any informant species are marginalized out so that they are irrelevant to the probability, using standard techniques for statistical phylogenetic models (Felsenstein 2004).

To analyze a given alignment window of several codons, we wish to obtain maximum likelihood estimates of the synonymous (and nonsynonymous) rates relative to the null model M. Our approach is to hold T fixed and estimate a new window-specific rate matrix Q^{wnd} by numerically maximizing the probability of the given alignment window jointly over two nonnegative parameters λ_s and λ_n, where the entries of Q^{wnd} relative to the entries of Q are given by

$$q_{ij}^{\mathrm{wnd}} = \begin{cases} \lambda_s \times q_{ij} \text{ if } i \neq j \text{ and } aa_i = aa_j \\ \lambda_n \times q_{ij} \text{ if } i \neq j \text{ and } aa_i \neq aa_j \\ -\sum_{k \neq i} q_{ik}^{\mathrm{wnd}} \text{ otherwise } (i = j) \end{cases}$$

where aa_i denotes the amino acid translation of codon i. Thus, λ_s represents a scale factor on the synonymous rates specified by Q, and similarly for λ_n on the nonsynonymous rates. Importantly, while Q is typically normalized to unity mean rate of replacement at equilibrium, we do not renormalize Q^{wnd}. Since T is held fixed, this allows λ_s and λ_n to control the absolute synonymous and nonsynonymous rates, respectively. Assuming Q is reversible, it is easy to verify that Q^{wnd} is reversible with the same equilibrium frequencies π_j as Q, by considering the decomposition $q_{ij} = \pi_j \times s_{ij}$ for symmetric "exchangeabilities" s_{ij} and noting that λ_s and λ_n scale the entries symmetrically.

Using this approach, we reduced the parameter estimation problem in each window to a mere two-dimensional optimization by reusing ORF- or ORFeome-wide estimates of many other phylogenetic model parameters, for which we probably could not obtain reliable joint estimates based only on a few codon sites (Anisimova et al. 2001; Suzuki and Nei 2002; Schmid and Yang 2008; Nozawa et al. 2009). This is similar to the approach used by the Sitewise Likelihood Ratio method of Massingham and Goldman (2005), which reuses ORF-wide parameter estimates while estimating $\omega = d_N/d_S$ at individual sites. We reuse common parameter estimates while effectively estimating both d_N and d_S in windows of several sites.

Null models

The relative rate estimation procedure does not make any assumptions about how the null model was originally estimated (except for reversibility of Q). In fact, we explored two different ways to estimate Q before providing it to this procedure. The first uses a parameterization equivalent to the M0 model of PAML, based on estimates of ω and κ, the transition/transversion rate ratio (Goldman and Yang 1994; Yang et al. 2000). The second is an empirical codon model (ECM) that essentially amounts to an independent estimate for every entry in the 61×61 rate matrix (under the reversibility constraint), not restricted to single-nucleotide instantaneous substitutions (Kosiol et al. 2007). Comparing the two approaches, we found the ECM parameterization to be clearly superior to M0 for our purposes: It achieved better fit to the training data based on BIC and AIC scores, led to tighter distributions of λ_s and λ_n, and resulted in a more conservative overall test for synonymous constraint (slightly fewer rejections of the null hypothesis at any significance level). The ECM approach also has the advantage that it accounts for the CpG hypermutability effect and any other sequence-specific rate biases, to the extent possible under the assumption of independence between codon sites. The results described in the main text are all based on the ECM parameterization. The parameterizations and estimation procedures are described in greater detail in Supplemental Material S2.

With both parameterizations, the ORFeome-wide null model was fit to a random sample of 5% of the codon sites in autosomal CCDS ORFs. A separate chromosome-specific null model was used for genes on chromosome X, based on all codon sites on that chromosome. The ORF-specific null models were estimated from the complete alignment of each ORF. The topology of the mammalian species tree proposed in the 2× mammals analysis (Lindblad-Toh et al. 2011) was used in all models, while the branch lengths were re-estimated for each ORF in the ORF-specific models.

Likelihood ratio tests

To test the significance of the adjusted rate estimates in any alignment window, we evaluated the likelihoods of several models:

1. The null model M ($\lambda_s = 1$; $\lambda_n = 1$)
2. $\lambda_s = 1$; λ_n estimated by maximum likelihood
3. λ_s and λ_n jointly estimated by maximum likelihood ($0 \leq \lambda_s \leq 1$)
4. $\lambda_s = 1$; λ_n estimated by maximum likelihood ($\lambda_n > 1$)

The likelihood ratios of nested models can then be used to perform the different significance tests we described. For example, the primary test for $\lambda_s < 1$ compares model 3 to model 2, and the test for $\lambda_n > 1$ compares model 4 to model 1.

To formally compare two of these models, we follow the standard frequentist approach for phylogenetic model comparison by computing the log-likelihood ratio (lods) and assuming that, when the null model holds, the statistic $-2 \times$ lods converges in distribution to the χ^2 distribution with one degree of freedom. We then report a P-value for each window by halving the χ^2 distribution tail probability corresponding to lods (Ota et al. 2000).

The test statistic exhibits some artifactual discretization behaviors in windows with a small number of substitutions (reflected in the stripes seen in Fig. 2E), likely violating the asymptotic assumptions justifying these significance estimates (Whelan and Goldman 1999). To ensure the robustness of the test, we performed additional benchmarks with simulated and permuted data, which are described in Supplemental Material S3.

Sliding windows and multiple testing correction

We applied LRTs to windows of a designated length in every CCDS ORF, beginning at the first site following the start codon and sliding by one-third of the window length. This yielded a list of millions of P-values for each window size and stringency threshold, which has a complex internal dependency structure owing to the overlapping windows and to latent rate correlations between nearby windows. We corrected the primary tests on λ_s^{ome} using the Benjamini and Hochberg false discovery rate (FDR)-controlling method (Benjamini and Hochberg 1995), which tolerates local positive correlations (Benjamini and Yekutieli 2001; Storey and Tibshirani 2003), requiring estimated FDR < 0.01. The secondary tests on λ_s^{ORF} were Bonferroni-corrected for the number of windows tested in each ORF, requiring corrected $P < 0.01$. Again, benchmarks with simulated and permuted data confirmed the robustness of this strategy (Supplemental Material S3).

Acknowledgments

We thank Matt Rasmussen, Loyal Goff, Nick Goldman, Katie Pollard, Kerstin Linblad-Toh, and the anonymous reviewers for helpful advice and discussions. Funding for this work was provided by the National Institutes of Health (U54 HG004555-01) and the National Science Foundation (DBI 0644282).

References

Ahmed ZM, Masmoudi S, Kalay E, Belyantseva IA, Mosrati MA, Collin RWJ, Riazuddin S, Hmani-Aifa M, Venselaar H, Kawar MN, et al. 2008. Mutations of LRTOMT, a fusion gene with alternative reading frames, cause nonsyndromic deafness in humans. *Nat Genet* **40:** 1335–1340.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol* **26:** 255–271.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18:** 1585–1592.

Aruscavage PJ, Bass BL. 2000. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA* **6:** 257–269.

Baek D, Green P. 2005. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc Natl Acad Sci* **102:** 12813–12818.

Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* **71:** 817–846.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304:** 1321–1325.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser A Stat Soc* **57:** 289–300.

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* **29:** 1165–1188.

Castillo-Davis C, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet* **31:** 415–418.

Castresana J. 2002. Genes on human chromosome 19 show extreme divergence from the mouse orthologs and a high GC content. *Nucleic Acids Res* **30:** 1751–1756.

Chamary JV, Hurst LD. 2005. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol* **6:** R75. doi: 10.1186/gb-2005-6-9-r75.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7:** 98–108.

Chen H, Blanchette M. 2007. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol* **7:** S9. doi: 10.1186/1471-2148-7-S1-S9.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: Insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10:** 741–754.

Chen F, Wang S, Chen C, Li W, Chuang T. 2006. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol* **23:** 675–682.

Chung WY, Wadhawan S, Szklarczyk R, Pond SK, Nekrutenko A. 2007. A first look at ARFome: Dual-coding genes in mammalian genomes. *PLoS Comput Biol* **3:** e91. doi: 10.1371/journal.pcbi.0030091.

Cohanim AB, Haran TE. 2009. The coexistence of the nucleosome positioning code with the genetic code on eukaryotic genomes. *Nucleic Acids Res* **37:** 6466–6476.

Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Res* **14:** 1188–1190.

Darty K, Denise A, Ponty Y. 2009. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25:** 1974–1975.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform* **10:** 97–109.

Dong X, Navratilova P, Fredman D, Drivenes O, Becker TS, Lenhard B. 2010. Exonic remnants of whole-genome duplication reveal *cis*-regulatory function of coding exons. *Nucleic Acids Res* **38:** 1071–1085.

Down T, Leong B, Hubbard TJ. 2006. A machine learning strategy to identify candidate binding sites in human protein-coding sequence. *BMC Bioinformatics* **7:** 419. doi: 10.1186/1471-2105-7-419.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007–1013.

Felsenstein J. 2004. *Inferring phylogenies*. Sinauer, Sunderland, MA.

Fox AK, Tuch BB, Chuang JH. 2008. Measuring the prevalence of regional mutation rates: An analysis of silent substitutions in mammals, fungi, and insects. *BMC Evol Biol* **8:** 186. doi: 10.1186/1471-2148-8-186.

Fuglsang A. 2006. Estimating the "effective number of codons": The Wright way of determining codon homozygosity leads to superior estimates. *Genetics* **172:** 1301–1307.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25:** i54–i62.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11:** 725–736.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. *Mol Cell* **22:** 769–781.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell* **27:** 91–105.

Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M, et al. 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428:** 529–535.

Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. 2010. Rnaz 2.0: Improved noncoding RNA detection. *Pac Symp Biocomput* **15:** 69–79.

Hameed M, Orrell RW, Cobbold M, Goldspink G, Harridge S. 2003. Expression of IGF-I splice variants in young and old human skeletal muscle after high resistance exercise. *J Physiol* **547:** 247–254.

Han J, Pedersen JS, Kwon SC, Belair CD, Kim YK, Yeom KH, Yang WY, Haussler D, Blelloch R, Kim VN. 2009. Posttranscriptional crossregulation between drosha and DGCR8. *Cell* **136:** 75–84.

Hastings ML, Ingle HA, Lazar MA, Munroe SH. 2000. Post-transcriptional regulation of thyroid hormone receptor expression by *cis*-acting sequences and a naturally occurring antisense RNA. *J Biol Chem* **275:** 11507–11513.

Hoopengardner B, Bhalla T, Staber C, Reenan R. 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301:** 832–836.

Howard MT, Aggarwal G, Anderson CB, Khatri S, Flanigan KM, Atkins JF. 2005. Recoding elements located adjacent to a subset of eukaryal selenocysteine-specifying UGA codons. *EMBO J* **24:** 1596–1607.

Hughes JF, Skaletsky H, Pyntikova T, Minx PJ, Graves T, Rozen S, Wilson RK, Page DC. 2005. Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437:** 100–103.

Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, Minx PJ, Fulton RS, McGrath SD, Locke DP, Friedman C, et al. 2010. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463:** 536–539.

Hurst LD. 2006. Preliminary assessment of the impact of microRNA-mediated regulation on coding sequence evolution in mammals. *J Mol Evol* **63:** 174–182.

Hurst LD, Pál C. 2001. Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet* **17:** 62–65.

Itzkovitz S, Alon U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res* **17:** 405–412.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12:** 996–1006.

Klemke M, Kehlenbach RH, Huttner WB. 2001. Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *EMBO J* **20:** 3849–3860.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol Biol Evol* **24:** 1464–1479.

Kural D, Ding Y, Wu J, Korpi A, Chuang J. 2009. COMIT: Identification of noncoding motifs under selection in coding sequences. *Genome Biol* **10:** R133. doi: 10.1186/gb-2009-10-11-r133.

Kuroki Y, Toyoda A, Noguchi H, Taylor TD, Itoh T, Kim D, Kim D, Choi S, Kim I, Choi HH, et al. 2006. Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat Genet* **38:** 158–167.

Lampe X, Samad OA, Guiguen A, Matis C, Remacle S, Picard JJ, Rijli FM, Rezsohazy R. 2008. An ultraconserved hox-pbx responsive element resides in the coding sequence of *Hoxa2* and is active in rhombomere 4. *Nucleic Acids Res* **36:** 3214–3225.

Lang G, Gombert WM, Gould HJ. 2005. A transcriptional regulatory element in the coding sequence of the human *bcl-2* gene. *Immunology* **114:** 25–36.

Lercher MJ, Williams EJB, Hurst LD. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol* **18:** 2032–2039.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15–20.

Li JB, Levanon EY, Yoon J, Aach J, Xie B, LeProust E, Zhang K, Gao Y, Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210–1213.

Lin Z, Ma H, Nei M. 2008. Ultraconserved coding regions outside the homeobox of mammalian hox genes. *BMC Evol Biol* **8:** 260. doi: 10.1186/1471-2148-8-260.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. Evolutionary constraint in the human genome based on 29 eutherian mammals. *Nature* **477**. doi: 10.1038/nature10530.

Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17:** 760–774.

Massingham T, Goldman N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169:** 1753–1762.

Nguyen MQ, Zhou Z, Marks CA, Ryba NJP, Belluscio L. 2007. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell* **131:** 1009–1017.

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci* **106:** 6700–6705.
Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol Biol Evol* **17:** 798–803.
Parker BJ, Moltke I, Roth A, Washietl S, Wen J, Kellis M, Breaker R, Pedersen JS. 2011. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res* (in press).
Parmley JL, Hurst LD. 2007a. Exonic splicing regulatory elements skew synonymous codon usage near intron–exon boundaries in mammals. *Mol Biol Evol* **24:** 1600–1603.
Parmley JL, Hurst LD. 2007b. How common are intragene windows with KA > KS owing to purifying selection on synonymous mutations? *J Mol Evol* **64:** 646–655.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **23:** 301–309.
Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol* **5:** e14. doi: 10.1371/journal.pbio.0050014.
Pedersen JS, Forsberg R, Meyer IM, Hein J. 2004a. An evolutionary model for protein-coding regions with conserved RNA structure. *Mol Biol Evol* **21:** 1913–1922.
Pedersen JS, Meyer IM, Forsberg R, Simmonds P, Hein J. 2004b. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res* **32:** 4925–4936.
Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2:** e33. doi: 10.1371/journal.pcbi.0020033.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.
Pond SK, Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* **22:** 2375–2385.
Poulin F, Brueschke A, Sonenberg N. 2003. Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. *J Biol Chem* **278:** 52290–52297.
Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, Sironi M. 2007. Intron size in mammals: Complexity comes to terms with economy. *Trends Genet* **23:** 20–24.
Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, et al. 2009. The Consensus Coding Sequence (CCDS) Project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19:** 1316–1323.
Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol* **24:** 1821–1831.
Ribrioux S, Brungger A, Baumgarten B, Seuwen K, John M. 2008. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics* **9:** 122. doi: 10.1186/1471-2164-9-122.
Robins H, Krasnitz M, Levine AJ. 2008. The computational detection of functional nucleotide sequence motifs in the coding regions of organisms. *Exp Biol Med* **233:** 665–673.
Rodrigue N, Lartillot N, Philippe H. 2008. Bayesian comparisons of codon substitution models. *Genetics* **180:** 1579–1591.
Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399:** 75–80.
Schattner P, Diekhans M. 2006. Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Res* **34:** 1700–1710.
Schmid K, Yang Z. 2008. The trouble with sliding windows and the selective pressure in BRCA1. *PLoS ONE* **3:** e3746. doi: 10.1371/journal.pone.0003746.
Schones DE, Cui K, Cuddapah S, Roh T, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132:** 887–898.
Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon–intron structure. *Nat Struct Mol Biol* **16:** 990–995.
Segal E, Widom J. 2009. Poly(dA:DT) tracts: Major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19:** 65–71.
Shabalina SA, Ogurtsov AY, Spiridonov NA. 2006. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res* **34:** 2428–2437.
Sharpless NE, DePinho RA. 1999. The *INK4A/ARF* locus and its two gene products. *Curr Opin Genet Dev* **9:** 22–30.
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.
Smith NGC, Hurst LD. 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. *Genetics* **153:** 1395–1402.
Stanley S, Bailey T, Mattick J. 2006. GONOME: Measuring correlations between GO terms and genomic positions. *BMC Bioinformatics* **7:** 94. doi: 10.1186/1471-2105-7-94.
Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci* **100:** 9440–9445.
Suzuki Y, Nei M. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* **19:** 1865–1869.
Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16:** 996–1001.
Tuller T, Waldman YY, Kupiec M, Ruppin E. 2010. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci* **107:** 3645–3650.
Tümpel S, Cambronero F, Sims C, Krumlauf R, Wiedemann LM. 2008. A regulatory module embedded in the coding region of *Hoxa2* controls expression in rhombomere 2. *Proc Natl Acad Sci* **105:** 20077–20082.
Urrutia AO, Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Res* **13:** 2260–2264.
Wang P, Yin S, Zhang Z, Xin D, Hu L, Kong X, Hurst L. 2008. Evidence for common short natural trans sense–antisense pairing between transcripts from protein coding genes. *Genome Biol* **9:** R169. doi: 10.1186/gb-2008-9-12-r169.
Warnecke T, Batada NN, Hurst LD. 2008. The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet* **4:** e1000250. doi: 10.1371/journal.pgen.1000250.
Washietl S, Machné R, Goldman N. 2008. Evolutionary footprints of nucleosome positions in yeast. *Trends Genet* **24:** 583–587.
Whelan S, Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol Biol Evol* **16:** 1292–1299.
Williams EJ, Hurst LD. 2002. Is the synonymous substitution rate in mammals gene-specific? *Mol Biol Evol* **19:** 1395–1398.
Woltering J, Duboule D. 2009. Conserved elements within open reading frames of mammalian Hox genes. *J Biol* **8:** 17. doi: 10.1186/jbiol116.
Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87:** 23–29.
Xing Y, Lee C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci* **102:** 13526–13531.
Xing Y, Lee C. 2006. Can RNA selection pressure distort the measurement of K_a/K_s? *Gene* **370:** 1–5.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15:** 496–503.
Yang Z, Nielsen R, Goldman N, Pedersen AK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11:** 377–394.
Yoshida H, Matsui T, Yamamoto A, Okada T, Mori K. 2001. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell* **107:** 881–891.

Received April 5, 2010; accepted in revised form June 15, 2010.

Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes

Michael F. Lin, Pouya Kheradpour, Stefan Washietl, et al.

Genome Res. 2011 21: 1916-1928 originally published online October 12, 2011
Access the most recent version at doi:10.1101/gr.108753.110