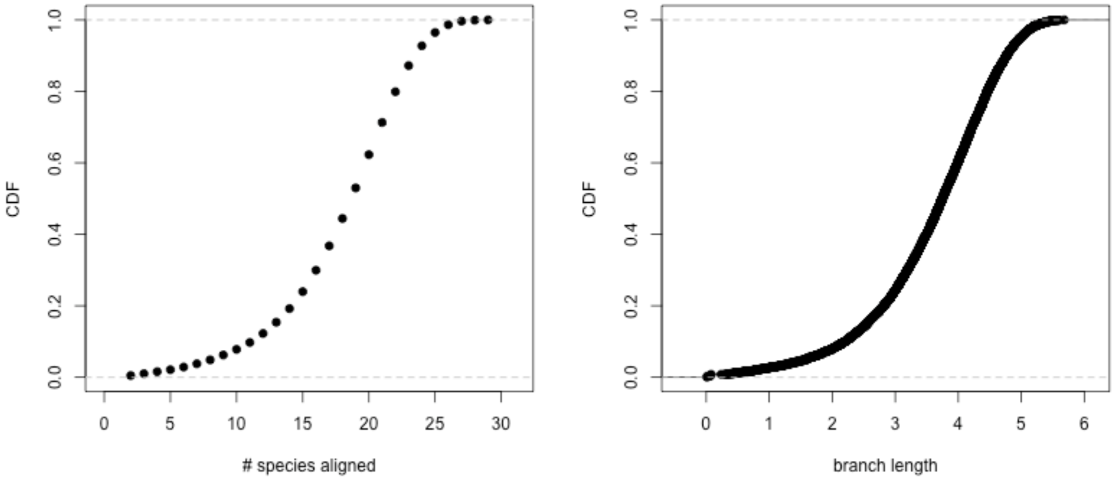
Supplementary Materials for Lin *et al.* (2010)

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

Table of Contents

S1. Alignment coverage for CCDS ORFs



Supplementary Figure 1. Alignment coverage for CCDS ORFs. Each plot shows the cumulative proportion of 16,939 CCDS ORFs with a given number of species aligned at 80% of nucleotide positions or more (left), including human, or the total branch length of the aligned species (right), in codon substitutions per codon site.

S2. Phylogenetic codon model parameterization and estimation

During the initial design of this study, we explored two ways to parameterize the phylogenetic codon models used as null models, which differ in how they determine the entries of the codon rate matrix Q.

1. M0 uses two parameters: κ, representing the transition/transversion rate ratio, and ω, representing the nonsynonymous/synonymous rate ratio. Only rate matrix entries corresponding to single-nucleotide substitutions have nonzero rates (Yang *et al.* 2000, Goldman and Yang 1994).
2. The empirical codon model (ECM) uses 1,830 parameters representing independent estimates for all symmetric 'exchangeabilities' among the 61 codons, with possibly nonzero values for all rate matrix entries (Kosiol *et al.* 2007).

Additionally, both parameterizations include 60 parameters for the codon equilibrium frequencies and 57 branch length parameters for the 29 mammals.

We estimated the parameters with an expectation-maximization algorithm (Siepel and Haussler 2004). In each M-step, we updated the ECM exchangeability parameters (if applicable) using a spectral approximation method (Arvestad and Bruno 1997), and all other parameters by gradient ascent on the expected log-likelihood function.

For the ORF-specific models, there is insufficient data to fully re-estimate the ECM. Instead, we used the ORFeome-wide exchangeability parameters, adjusted with a single scale factor on all non-synonymous rates, as suggested by the "ECM+ω" parameterization of Kosiol *et al.* (2007), optimized by gradient ascent in the M-step for each ORF. The chromosome X null model is also an ECM+ω model using the ORFeome-wide exchangeabilities, with ω re-estimated based on a sample of sites on that chromosome.

We obtained maximum likelihood estimates of λ_s and λ_n in any individual window using cyclic coordinate ascent on the likelihood function directly.

S3. Additional benchmarks of the combined window testing procedure

We performed several empirical benchmarks to ensure that our combined window testing procedure achieves high specificity. In particular, we wished to ensure that an unexpectedly high false discovery rate could not result from (1) violations of the asymptotic assumptions underlying the LRT method (Whelan and Goldman 1999), (2) violations of the null model assumptions, such as random site-to-site rate variation (Pond and Muse 2005), or (3) failure of the Benjamini-Hochberg procedure (BH) to properly correct for multiple testing given the strong correlations between sliding windows (where consecutive windows overlap by two-thirds of their length). While the BH procedure accommodates such positive dependencies (Benjamini and Yekutieli 2001, Storey and Tibshirani 2003), the correlations in our data arising from overlapping sliding windows are stronger than in typical previous applications of the procedure.

Simulated alignment. First, we generated an alignment of 50,000 codon sites according to the ORFeome-wide null model. The simulated alignment had a randomly chosen subset of 19 of the 29 mammalian species (the median number of aligned species for CCDS ORFs). We then ran the full sliding window testing procedure on this alignment at 15-codon resolution, sliding by one-third of the window length. Since the alignment is simulated according to the null model, the procedure should not call any windows significant. Without multiple testing correction, 119 of 9,358 (1.3%) of windows in the simulated alignment had significantly reduced synonymous rate estimates at $P < 0.01$. As expected, BH adjusted all of these to non-significance. This means it is unlikely that problems with the sliding window procedure or assumptions underlying the LRT framework lead to a much higher-than-nominal false positive rate for our approach.

Shuffled ORF alignments. Second, we ran the full procedure on shuffled versions of the chromosome 17 ORFs. (Due to computational requirements, we used chromosome 17 as a small-scale benchmark of interest since it includes both *BRCA1* and the HoxB cluster.) For each CCDS ORF on chromosome 17, we took its alignment and shuffled its codon sites in random order (reordering nucleotide columns in atomic groups of 3, in the correct reading frame). Thus, for each ORF we had a shuffled version that preserves its sequence composition and site-to-site evolutionary rate variation, but randomizes out the immediate adjacency of slowly evolving sites expected in an overlapping functional element. We then applied the full testing procedure used in the real analysis (with criteria on λ_s^{ome} , λ_s^{ORF} , and λ_n^{ome}, followed by multiple testing corrections).

In these shuffled alignments, 0.086% of permuted 15-codon windows were called significant, compared to 1.48% of windows in the real chromosome 17 ORFs (a 17-fold difference). Note that while this permutation experiment leads to an empirically estimated FDR of ~6%, this does not mean that BH failed to control FDR at 1%, since the underlying hypothesis tests are for deviation from a null model that does not incorporate the site-to-site rate variation seen in the real data (which is preserved by the shuffling). This means it is unlikely that (unlocalized) site-to-site rate variation leads to a much higher-than-nominal false positive rate for our approach.

Non-overlapping windows. Lastly, we ran BH on non-overlapping windows in the real alignments (every third window from the raw data). The resulting proportion of significant windows is *higher* in this case than with overlapping windows (2.09% vs. 1.72% at 15-codon resolution), with somewhat higher median λ_s^{ome} (0.2787 vs. 0.2572), and significant windows falling in a smaller proportion of CCDS genes (30% vs. 36%). This means it is unlikely that the BH procedure is badly led astray by the correlations arising from overlapping sliding windows; instead, it suggests that BH properly corrects for multiple testing in both cases, but the resolution and sensitivity of the combined testing procedure are predictably reduced without the overlapping sliding windows.

Overall, these benchmarks strongly confirm that our combined testing procedure achieves high specificity in identifying windows with localized reductions in estimated synonymous substitution rates.

S4. Permutation tests for *BRCA1*

In their criticism of an earlier study on putative synonymous constraint in *BRCA1*, Schmid and Yang (2008) specifically raised concerns about statistical significance in sliding window procedures. The above benchmarks argue that our approach effectively addresses such concerns for our overall dataset. To further ensure the robustness of our results for *BRCA1*, we performed additional permutation tests specifically for this gene.

We generated 1,000 shuffled versions of the *BRCA1* ORF, as described above, and applied our full sliding window testing procedure to each one, which called significant windows in only 2 of the 1,000 shuffled ORFs. Furthermore, none of the 359,000 permuted windows tested in this experiment received a higher likelihood ratio for synonymous rate reduction (based on the ORF-specific null model) than the most significant window in the real data. This strongly confirms that the region identified in the real data indeed represents a highly significant, localized reduction in the synonymous substitution rate, and not merely an artifact of the sliding window procedure.

S5. Proportion of significant windows/genes on each chromosome

	Windows		Genes	
	#	% Significant	#	% Significant
1	186,745	1.69%	1,848	35.2%
2	116,855	2.04%	1,039	41.6%
3	100,350	1.92%	919	41.3%
4	67,223	2.01%	626	37.4%
5	80,360	2.08%	731	43.8%
6	93,079	1.63%	945	34.4%
7	74,745	1.67%	737	39.5%
8	54,884	2.02%	552	33.7%
9	71,993	2.00%	700	36.4%
10	72,337	2.09%	661	42.8%
11	100,348	1.37%	1,099	28.1%
12	90,393	1.84%	893	39.2%
13	33,338	2.10%	285	44.2%
14	57,376	2.03%	530	42.3%
15	57,376	1.60%	471	41.2%
16	72,087	1.53%	699	32.2%
17	99,335	1.48%	977	36.1%
18	26,318	2.28%	231	46.3%
19	105,059	0.73%	1,162	20.3%
20	46,254	1.78%	496	35.7%
21	18,367	1.76%	201	26.4%
22	35,151	1.35%	385	29.4%
X	69,853	1.54%	724	28.9%

Supplementary Table 1. Percentage of windows/genes on each chromosome in which our method identified significant synonymous constraint.

S6. 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 carefully examined the sequence composition of the regions reported as significant. They do show certain subtle biases in sequence composition compared to other CCDS protein-coding sequences. For example, they have 3.4% lower G+C content at first and third codon positions (but 1.6% higher at second). CpG dinucleotides span 3.2% of second and third codon positions, compared to 2.4% in other coding regions, but they are slightly depleted spanning codon positions (1,2) and (3,1). Encoded serine residues are more frequent (9.2% of sites, compared to 8.2%) and tryptophan is less frequent (1.0% vs. 1.2%). These and numerous other nucleotide, dinucleotide and amino acid biases are statistically significant (Supplementary Table 2, below), but none represent more than a 1.4-fold enrichment or depletion, and most far less. To put this in perspective, human ORFs show comparable or greater compositional variation from chromosome to chromosome: for example, ORFs on chromosome 3 are 1.4-fold enriched for ApA dinucleotides across codon positions (3,1), and chromosome 19 ORFs are twofold depleted for TpA's spanning codon positions (2,3).

Other compositional properties of the SCEs allowed us to rule out certain possible artifactual explanations for their low divergence across mammals, at least as predominant effects. First, short tandem and microsatellite repeats within coding regions, which can be maintained through evolution by non-selective processes (Richard *et al.* 2008), might resemble codons with conserved synonymous sites, but the SCEs are depleted for such repeats relative to CCDS coding regions (0.87% vs. 2.03% of positions identified by TRF or the appropriate subset of RepeatMasker annotations). Second, biases in DNA mismatch repair can lead to a reduced apparent neutral substitution rate within some genomic isochores (Chamary *et al.* 2006), but the lower G+C content at third codon positions in SCEs is contrary to the trend in regions that have experienced such biased conversions (Galtier and Duret 2007). Third, codon usage biases related to translational efficiency or background compositional effects can reduce synonymous divergence without additional overlapping function, and while the relative usage frequencies of synonymous codons tend to differ between SCEs and other coding regions (Supplementary Table 3), the overall codon usage in SCEs is actually slightly less biased than average: the Effective Number of Codons (ENC, Wright 1990, Fuglsang 2006) in all SCEs taken together is 56.8, higher than in non-significant regions (54.4). Similarly, the ENC in a typical ORF containing an SCE (median 51.2) is slightly higher than in other ORFs (median 50.1). In contrast, the ENC is markedly reduced in ORFs with strong codon bias owing to isochore-specific G+C composition (Chamary *et al.* 2006).

In summary, the SCEs show significant but not extreme differences in sequence composition and codon usage compared to other coding regions, which probably reflect the sequence-dependent biological nature of the overlapping functional elements they encode. They may also partly reflect biases in our phylogenetic models arising from contextual effects they do not accurately capture, but the observed compositional biases do not suggest crippling shortcomings of our overall approach. In particular, the lack of increased codon usage bias in SCEs indicates that, as intended, our method excluded regions explained by this effect.

Supplementary Table 2 (follows on next page). Nucleotide and dinucleoutide sequence composition biases in SCEs (15-codon resolution). Columns:

- **rf** codon reading frame (0, 1, or 2)
- **kmer** the nucleotide or dinculeotide in question
- **exp** the average (expected) frequency of the nucleotide at the specified codon position in random subsets of CCDS sequences
- **sd** standard deviation of exp across multiple random subsets of CCDS sequences (each with comparable coverage to the real SCEs)
- **obs** frequency observed in SCEs
- **enrich** $\log_2$ fold-enrichment/depletion relative to the random regions
- **Z** Z-score of enrichment/depletion relative to the random regions ([obs – exp]/sd)

rf	kmer	exp	sd	obs	enrich	Z
0	A	26.76%	0.07%	28.46%	0.0888	23.50
1	A	31.54%	0.09%	30.50%	-0.0482	-11.89
2	A	19.44%	0.06%	22.01%	0.1791	39.65
0	C	24.72%	0.09%	23.89%	-0.0489	-9.27
1	C	23.15%	0.08%	25.47%	0.1377	28.52
2	C	29.62%	0.09%	27.10%	-0.1284	-28.31
0	G	31.40%	0.10%	29.78%	-0.0767	-17.02
1	G	18.83%	0.07%	19.16%	0.0254	4.77
2	G	28.34%	0.08%	25.57%	-0.1485	-33.03
0	T	17.12%	0.06%	17.87%	0.0619	11.84
1	T	26.48%	0.08%	24.87%	-0.0908	-20.16
2	T	22.60%	0.07%	25.33%	0.1642	37.07
0	AA	9.45%	0.05%	10.17%	0.1059	13.17
1	AA	6.90%	0.04%	7.49%	0.1200	13.56
2	AA	5.20%	0.04%	6.57%	0.3365	34.84
0	AC	5.27%	0.04%	5.58%	0.0827	8.48
1	AC	7.50%	0.05%	7.16%	-0.0665	-7.46
2	AC	3.89%	0.03%	4.23%	0.1219	11.25
0	AG	5.54%	0.04%	6.37%	0.1997	20.37
1	AG	10.74%	0.06%	8.90%	-0.2717	-33.03
2	AG	7.52%	0.04%	8.08%	0.1030	13.34
0	AT	6.50%	0.04%	6.34%	-0.0349	-3.90
1	AT	6.40%	0.04%	6.95%	0.1182	12.92
2	AT	2.83%	0.03%	3.13%	0.1471	10.81
0	CA	7.40%	0.04%	7.00%	-0.0792	-9.30
1	CA	6.06%	0.04%	6.70%	0.1465	15.53
2	CA	10.21%	0.05%	10.11%	-0.0142	-1.94
0	CC	6.05%	0.05%	7.18%	0.2478	24.88
1	CC	8.28%	0.05%	7.93%	-0.0620	-7.33
2	CC	8.47%	0.05%	7.41%	-0.1924	-20.12
0	CG	3.27%	0.03%	2.96%	-0.1452	-8.96
1	CG	2.37%	0.03%	3.36%	0.5074	33.92
2	CG	4.25%	0.04%	3.48%	-0.2890	-21.83
0	CT	8.00%	0.05%	6.75%	-0.2451	-27.10
1	CT	6.45%	0.04%	7.47%	0.2122	25.22
2	CT	6.69%	0.04%	6.09%	-0.1337	-14.57
0	GA	11.92%	0.07%	10.67%	-0.1606	-18.85
1	GA	3.46%	0.03%	4.14%	0.2575	22.88
2	GA	7.34%	0.04%	6.78%	-0.1144	-12.48
0	GC	6.90%	0.05%	6.80%	-0.0220	-1.95
1	GC	6.42%	0.05%	5.81%	-0.1439	-12.35
2	GC	7.37%	0.05%	6.64%	-0.1506	-15.51
0	GG	6.49%	0.04%	6.92%	0.0937	11.32
1	GG	5.10%	0.04%	4.85%	-0.0716	-5.86
2	GG	9.81%	0.05%	8.49%	-0.2084	-24.17
0	GT	6.09%	0.04%	5.39%	-0.1761	-16.03
1	GT	3.85%	0.03%	4.36%	0.1805	16.12
2	GT	3.82%	0.03%	3.66%	-0.0628	-5.81
0	TA	2.77%	0.03%	2.66%	-0.0559	-4.08
1	TA	3.02%	0.03%	3.67%	0.2796	22.59
2	TA	4.01%	0.03%	5.01%	0.3184	30.86
0	TC	4.93%	0.04%	5.91%	0.2608	27.27
1	TC	7.42%	0.05%	6.20%	-0.2605	-23.11
2	TC	4.99%	0.04%	5.61%	0.1692	15.74
0	TG	3.53%	0.03%	2.92%	-0.2740	-19.19
1	TG	10.14%	0.04%	8.46%	-0.2612	-38.88
2	TG	9.82%	0.06%	9.73%	-0.0133	-1.58
0	TT	5.89%	0.04%	6.38%	0.1151	10.94
1	TT	5.90%	0.04%	6.54%	0.1490	17.22
2	TT	3.78%	0.04%	4.99%	0.3980	34.22

Supplementary Table 3 (follows on next page). Codon usage biases in SCEs (15-codon resolution)**.** Columns:

- **AA.codon** amino acid (IUPAC letter) and codon
- **exp** average (expected) frequency of the codon in random subsets of CCDS sequences, as a fraction of all sites encoding the corresponding amino acid
- **sd** standard deviation of exp across multiple random subsets of CCDS sequences (comparable coverage to the real SCEs)
- **obs** frequency observed in SCEs
- **enrich** $\log_2$ fold-enrichment/depletion relative to the random regions
- **Z** Z-score of enrichment/depletion relative to the random regions ([inc obs – exp]/sd)
- **degen** degeneracy, number of synonymous codons for this amino acid

AA.codon	exp	sd	obs	enrich	Z	degen
A.GCA	23.32%	0.2479%	23.50%	0.01	0.75	4
A.GCC	40.12%	0.3020%	34.39%	-0.22	-18.96	4
A.GCG	9.89%	0.1871%	14.12%	0.51	22.58	4
A.GCT	26.68%	0.2844%	27.99%	0.07	4.63	4
C.TGC	53.29%	0.5960%	50.90%	-0.07	-4.01	2
C.TGT	46.71%	0.5960%	49.10%	0.07	4.01	2
D.GAC	52.86%	0.4267%	50.24%	-0.07	-6.14	2
D.GAT	47.14%	0.4267%	49.76%	0.08	6.14	2
E.GAA	43.37%	0.3318%	49.95%	0.20	19.82	2
E.GAG	56.63%	0.3318%	50.05%	-0.18	-19.82	2
F.TTC	53.13%	0.4211%	46.37%	-0.20	-16.07	2
F.TTT	46.87%	0.4211%	53.63%	0.19	16.07	2
G.GGA	25.38%	0.3470%	28.72%	0.18	9.63	4
G.GGC	33.82%	0.3304%	27.29%	-0.31	-19.79	4
G.GGG	24.48%	0.3060%	24.12%	-0.02	-1.18	4
G.GGT	16.32%	0.2635%	19.87%	0.28	13.49	4
H.CAC	57.59%	0.4544%	53.21%	-0.11	-9.63	2
H.CAT	42.41%	0.4544%	46.79%	0.14	9.63	2
I.ATA	17.07%	0.2947%	20.57%	0.27	11.88	3
I.ATC	46.14%	0.4364%	37.55%	-0.30	-19.67	3
I.ATT	36.80%	0.3844%	41.88%	0.19	13.23	3
K.AAA	44.09%	0.3051%	51.58%	0.23	24.56	2
K.AAG	55.91%	0.3051%	48.42%	-0.21	-24.56	2
L.CTA	7.16%	0.1559%	8.64%	0.27	9.49	6
L.CTC	19.02%	0.2011%	16.95%	-0.17	-10.27	6
L.CTG	39.47%	0.2531%	31.52%	-0.32	-31.41	6
L.CTT	13.40%	0.1638%	15.05%	0.17	10.08	6
L.TTA	7.96%	0.1122%	11.51%	0.53	31.60	6
L.TTG	13.00%	0.1771%	16.34%	0.33	18.85	6
M.ATG	100.00%	0.0000%	100.00%	0.00	N/A	1
N.AAC	52.05%	0.4332%	49.17%	-0.08	-6.64	2
N.AAT	47.95%	0.4332%	50.83%	0.08	6.64	2
P.CCA	27.99%	0.2903%	28.19%	0.01	0.72	4
P.CCC	32.25%	0.3025%	27.51%	-0.23	-15.66	4
P.CCG	10.87%	0.2156%	14.03%	0.37	14.65	4
P.CCT	28.89%	0.2906%	30.27%	0.07	4.72	4
Q.CAA	26.61%	0.3306%	32.42%	0.28	17.58	2
Q.CAG	73.39%	0.3306%	67.58%	-0.12	-17.58	2
R.AGA	21.24%	0.2545%	25.86%	0.28	18.18	6
R.AGG	20.41%	0.2821%	22.26%	0.13	6.57	6
R.CGA	11.21%	0.2428%	11.90%	0.09	2.84	6
R.CGC	18.58%	0.3395%	15.48%	-0.26	-9.12	6
R.CGG	20.54%	0.2967%	16.66%	-0.30	-13.09	6
R.CGT	8.03%	0.1475%	7.83%	-0.04	-1.32	6
S.AGC	23.94%	0.2384%	21.42%	-0.16	-10.58	6
S.AGT	15.48%	0.2304%	16.58%	0.10	4.80	6
S.TCA	15.23%	0.2230%	15.64%	0.04	1.83	6
S.TCC	21.27%	0.2543%	19.69%	-0.11	-6.22	6
S.TCG	5.33%	0.1396%	6.88%	0.37	11.15	6
S.TCT	18.76%	0.2234%	19.79%	0.08	4.63	6
T.ACA	28.74%	0.3174%	28.51%	-0.01	-0.74	4
T.ACC	34.74%	0.3226%	31.24%	-0.15	-10.85	4
T.ACG	11.22%	0.2457%	13.23%	0.24	8.21	4
T.ACT	25.31%	0.2887%	27.02%	0.09	5.96	4
V.GTA	11.94%	0.2130%	16.19%	0.44	19.92	4
V.GTC	23.31%	0.2890%	22.05%	-0.08	-4.35	4
V.GTG	46.27%	0.3526%	38.58%	-0.26	-21.80	4
V.GTT	18.48%	0.2786%	23.18%	0.33	16.88	4
W.TGG	100.00%	0.0000%	100.00%	0.00	N/A	1
Y.TAC	54.82%	0.4607%	51.83%	-0.08	-6.49	2
Y.TAT	45.18%	0.4607%	48.17%	0.09	6.49	2

S7. Characteristics of genes containing SCEs

Gene length. In the main text, we note that the ORFs containing SCEs tend to be longer than other CCDS ORFs, but not significantly longer or shorter than expected based on randomly sampling ORFs weighted by their length. The consistency of the observed ORF length distribution with the random distribution could be taken to suggest a problem with the SCE dataset. In particular, if our method just randomly called one false positive out of every *n* windows tested, we would also expect to see the random ORF length distribution.

However, in addition to our rigorous multiple testing corrections further supported by simulation and permutation benchmarks (Supplement S3, above), we collected a few other statistics that strongly argue against random false positives as a predominant explanation:

- SCEs are depleted in the lengthiest quartile of single-exon ORFs relative to comparable multi-exon ORFs (0.76 vs. 1.21 SCEs per 1,000 codons). Such a depletion would not be expected from random false positives. Instead, this seems to reflect the fact that longer genes also tend to have more exons, and therefore potentially more splice sites regulated by exonic regulatory sequences.
- A linear model predicting the number of SCEs in each gene based on its length and number of introns shows that both variables are significant, but even combined they only provide limited explanatory power ($R^2 = 0.20$). We would expect much higher correlation from random false positives occurring at a fixed rate.
- The individual *introns* in genes containing SCEs also tend to be much longer, which clearly cannot be directly explained by a multiple testing issue, since our method does not examine intronic sequences.

In general, gene length (both ORF length, and genomic span including introns) is known to correlate with numerous other relevant characteristics including expression levels, functional categories, conservation, etc., and ultimately it can be difficult to disentangle these effects (Castillo-Davis *et al.* 2002, Urrutia and Hurst 2003, Stanley *et al.* 2006, Pozzoli *et al.* 2007). Overall therefore, while we are highly confident that the observed ORF length distribution does not reflect an artifactual multiple testing issue, the available evidence does not support making more-specific claims about the association between SCEs and gene length.

Supplementary Table 4 (follows on next page). Gene Ontology enrichments for SCEs (15-codon resolution). The enrichments are assessed in the hypergeometric sampling paradigm of drawing from an urn containing white and black balls, where the white balls are of interest, representing the genes annotated with a given GO term. The balls are either genes (left-hand group of colunms) or individual windows (right-hand group), where in the latter case the GO terms associated with each gene are propagated to all windows in its ORF. Columns:

- **w** number of white balls drawn from the urn (number of genes/windows with the corresponding term containing an SCE)
- **W** total number of white balls in the urn before drawing (total number of genes/windows with the corresponding term)
- **B** total number of black balls in the urn before drawing (total number of genes – W)
- **w+b** number of balls drawn from the urn
- **prop w** proportion of drawn balls that are white (w/[w+b])
- **prop** W, proportion of all white balls that were drawn (w/W)
- **enrich** fold enrichment of white balls in the drawn set compared to originally in the urn (prop w /[W/(W+B)])
- **bonfP** Bonferroni-corrected *P*-value of enrichment computed from the hypergeometric distribution.

The purpose of the window-level analysis (right-hand group of columns) is to control for the varying ORF lengths, which can otherwise be a major confounding factor when certain GO terms are enriched in longer or shorter genes. It is somewhat incorrect to use the hypergeometric distribution in the window-level analysis because of the overlapping sliding windows (balls are not drawn independently), but the numbers involved are sufficiently large that this is unlikely to be a serious problem (i.e. the hypergeometric has essentially converged to a binomial in this regime, and e.g. dividing all the counts by three, or performing the analysis only on non-overlapping windows, still leads to highly significant *P*-values).

		balls = genes								balls = windows							
GO term	description	w	W	B	w+b	prop w	prop W	enrich	bonfP	w	W	B	w+b	prop w	prop W	enrich	bonfP
GO:0016568	chromatin modification	94	134	16777	6033	1.56%	70.15%	1.97	7.87E-13	949	25223	1701979	29731	3.19%	3.76%	2.19	8.12E-102
GO:0005249	voltage-gated potassium channel activity	65	97	16814	6033	1.08%	67.01%	1.88	7.14E-07	330	10184	1717018	29731	1.11%	3.24%	1.88	1.37E-22
GO:0022843	voltage-gated cation channel activity	78	117	16794	6033	1.29%	66.67%	1.87	1.87E-08	392	12702	1714500	29731	1.32%	3.09%	1.79	2.95E-23
GO:0034703	cation channel complex	65	98	16813	6033	1.08%	66.33%	1.86	1.43E-06	357	10301	1716901	29731	1.20%	3.47%	2.01	6.84E-30
GO:0004842	ubiquitin-protein ligase activity	66	101	16810	6033	1.09%	65.35%	1.83	2.85E-06	468	15026	1712176	29731	1.57%	3.11%	1.81	4.93E-29
GO:0005267	potassium channel activity	77	121	16790	6033	1.28%	63.64%	1.78	8.65E-07	381	11963	1715239	29731	1.28%	3.18%	1.85	5.31E-25
GO:0006813	potassium ion transport	91	143	16768	6033	1.51%	63.64%	1.78	2.19E-08	414	15029	1712173	29731	1.39%	2.75%	1.60	6.31E-16
GO:0019787	small conjugating protein ligase activity	74	117	16794	6033	1.23%	63.25%	1.77	2.91E-06	505	16122	1711080	29731	1.70%	3.13%	1.82	4.42E-32
GO:0030955	potassium ion binding	68	110	16801	6033	1.13%	61.82%	1.73	5.64E-05	272	11771	1715431	29731	0.91%	2.31%	1.34	9.41E-03
GO:0034702	ion channel complex	68	110	16801	6033	1.13%	61.82%	1.73	5.64E-05	363	11432	1715770	29731	1.22%	3.18%	1.84	1.69E-23
GO:0005244	voltage-gated ion channel activity	98	159	16752	6033	1.62%	61.64%	1.73	5.92E-08	453	16764	1710438	29731	1.52%	2.70%	1.57	3.67E-16
GO:0022832	voltage-gated channel activity	98	159	16752	6033	1.62%	61.64%	1.73	5.92E-08	453	16764	1710438	29731	1.52%	2.70%	1.57	3.67E-16
GO:0045892	negative regulation of transcription, DNA-dependent	89	148	16763	6033	1.48%	60.14%	1.69	3.20E-06	540	15124	1712078	29731	1.82%	3.57%	2.07	5.45E-50
GO:0051253	negative regulation of RNA metabolic process	90	150	16761	6033	1.49%	60.00%	1.68	3.06E-06	553	15310	1711892	29731	1.86%	3.61%	2.10	9.36E-53
GO:0016564	transcription repressor activity	89	149	16762	6033	1.48%	59.73%	1.67	5.27E-06	629	16040	1711162	29731	2.12%	3.92%	2.28	4.91E-73
GO:0016881	acid-amino acid ligase activity	78	131	16780	6033	1.29%	59.54%	1.67	6.42E-05	519	17796	1709406	29731	1.75%	2.92%	1.69	8.67E-26
GO:0010629	negative regulation of gene expression	129	222	16689	6033	2.14%	58.11%	1.63	2.09E-08	891	25014	1702188	29731	3.00%	3.56%	2.07	1.20E-83
GO:0016481	negative regulation of transcription	122	210	16701	6033	2.02%	58.10%	1.63	7.94E-08	828	23515	1703687	29731	2.78%	3.52%	2.05	2.74E-75
GO:0022836	gated channel activity	140	242	16669	6033	2.32%	57.85%	1.62	4.28E-09	768	28566	1698636	29731	2.58%	2.69%	1.56	2.65E-28
GO:0016879	ligase activity, forming carbon-nitrogen bonds	87	151	16760	6033	1.44%	57.62%	1.62	9.69E-05	541	20471	1706731	29731	1.82%	2.64%	1.54	9.10E-18
GO:0007399	nervous system development	146	256	16655	6033	2.42%	57.03%	1.60	6.76E-09	811	28878	1698324	29731	2.73%	2.81%	1.63	4.81E-36
GO:0016071	mRNA metabolic process	135	239	16672	6033	2.24%	56.49%	1.58	1.18E-07	830	23952	1703250	29731	2.79%	3.47%	2.01	2.03E-72
GO:0006512	ubiquitin cycle	202	358	16553	6033	3.35%	56.42%	1.58	1.86E-12	1013	42940	1684262	29731	3.41%	2.36%	1.37	5.74E-19
GO:0003713	transcription coactivator activity	80	142	16769	6033	1.33%	56.34%	1.58	1.32E-03	550	17356	1709846	29731	1.85%	3.17%	1.84	1.91E-36
GO:0005261	cation channel activity	114	204	16707	6033	1.89%	55.88%	1.57	9.45E-06	585	24428	1702774	29731	1.97%	2.39%	1.39	4.44E-11
GO:0043565	sequence-specific DNA binding	228	408	16503	6033	3.78%	55.88%	1.57	1.13E-13	1224	33856	1693346	29731	4.12%	3.62%	2.10	1.26E-120
GO:0003702	RNA polymerase II transcription factor activity	101	182	16729	6033	1.67%	55.49%	1.56	1.23E-04	744	19988	1707214	29731	2.50%	3.72%	2.16	3.37E-77
GO:0004674	protein serine/threonine kinase activity	180	325	16586	6033	2.98%	55.38%	1.55	7.77E-10	942	44306	1682896	29731	3.17%	2.13%	1.24	5.40E-07
GO:0003700	transcription factor activity	399	722	16189	6033	6.61%	55.26%	1.55	1.69E-24	2204	69219	1657983	29731	7.41%	3.18%	1.85	3.83E-159
GO:0046873	metal ion transmembrane transporter activity	132	239	16672	6033	2.19%	55.23%	1.55	1.59E-06	657	28801	1698401	29731	2.21%	2.28%	1.33	7.52E-09
GO:0016563	transcription activator activity	127	230	16681	6033	2.11%	55.22%	1.55	3.45E-06	865	26907	1700295	29731	2.91%	3.21%	1.87	2.57E-61
GO:0048731	system development	229	415	16496	6033	3.80%	55.18%	1.55	7.16E-13	1218	46666	1680536	29731	4.10%	2.61%	1.52	1.11E-40
GO:0003712	transcription cofactor activity	124	226	16685	6033	2.06%	54.87%	1.54	9.33E-06	853	26168	1701034	29731	2.87%	3.26%	1.89	6.04E-63
GO:0008134	transcription factor binding	162	296	16615	6033	2.69%	54.73%	1.53	4.60E-08	1142	35308	1691894	29731	3.84%	3.23%	1.88	1.82E-83
GO:0045934	negative regulation of nucleobase, nucleoside, nucleotid	128	234	16677	6033	2.12%	54.70%	1.53	6.75E-06	853	26127	1701075	29731	2.87%	3.26%	1.90	3.12E-63
GO:0006397	mRNA processing	114	209	16702	6033	1.89%	54.55%	1.53	6.37E-05	698	20017	1707185	29731	2.35%	3.49%	2.03	1.79E-61
GO:0008380	RNA splicing	96	176	16735	6033	1.59%	54.55%	1.53	8.36E-04	596	16636	1710566	29731	2.00%	3.58%	2.08	6.98E-56
GO:0051172	negative regulation of nitrogen compound metabolic pro	128	235	16676	6033	2.12%	54.47%	1.53	9.73E-06	853	26201	1701001	29731	2.87%	3.26%	1.89	1.03E-62
GO:0006325	chromatin organization	135	248	16663	6033	2.24%	54.44%	1.53	3.80E-06	1118	31560	1695642	29731	3.76%	3.54%	2.06	2.04E-104
GO:0006357	regulation of transcription from RNA polymerase II prom	179	329	16582	6033	2.97%	54.41%	1.53	7.82E-09	991	34259	1692943	29731	3.33%	2.89%	1.68	6.18E-50
GO:0004672	protein kinase activity	262	482	16429	6033	4.34%	54.36%	1.52	6.48E-14	1422	70231	1656971	29731	4.78%	2.02%	1.18	2.98E-06
GO:0016773	phosphotransferase activity, alcohol group as acceptor	306	573	16338	6033	5.07%	53.40%	1.50	3.83E-15	1616	82576	1644626	29731	5.44%	1.96%	1.14	6.71E-04
GO:0016310	phosphorylation	286	540	16371	6033	4.74%	52.96%	1.48	2.40E-13	1557	74870	1652332	29731	5.24%	2.08%	1.21	2.90E-10
GO:0006468	protein amino acid phosphorylation	267	506	16405	6033	4.43%	52.77%	1.48	4.93E-12	1490	70338	1656864	29731	5.01%	2.12%	1.23	4.92E-12
GO:0044451	nucleoplasm part	144	273	16638	6033	2.39%	52.75%	1.48	1.95E-05	833	32028	1695174	29731	2.80%	2.60%	1.51	1.66E-26
GO:0005216	ion channel activity	154	292	16619	6033	2.55%	52.74%	1.48	5.82E-06	828	34875	1692327	29731	2.78%	2.37%	1.38	1.02E-15
GO:0030528	transcription regulator activity	561	1064	15847	6033	9.30%	52.73%	1.48	2.60E-28	3275	1E+05	1618955	29731	11.02%	3.03%	1.76	4.58E-208
GO:0045941	positive regulation of transcription	117	222	16689	6033	1.94%	52.70%	1.48	5.57E-04	765	24942	1702260	29731	2.57%	3.07%	1.78	1.57E-46
GO:0045935	positive regulation of nucleobase, nucleoside, nucleotide	122	232	16679	6033	2.02%	52.59%	1.47	3.59E-04	784	25841	1701361	29731	2.64%	3.03%	1.76	4.62E-46
GO:0015672	monovalent inorganic cation transport	135	257	16654	6033	2.24%	52.53%	1.47	8.23E-05	608	27721	1699481	29731	2.05%	2.19%	1.27	1.80E-05
GO:0031327	negative regulation of cellular biosynthetic process	135	257	16654	6033	2.24%	52.53%	1.47	8.23E-05	856	27637	1699565	29731	2.88%	3.10%	1.80	4.60E-54
GO:0009890	negative regulation of biosynthetic process	135	258	16653	6033	2.24%	52.33%	1.47	1.13E-04	856	27651	1699551	29731	2.88%	3.10%	1.80	5.63E-54
GO:0015267	channel activity	160	308	16603	6033	2.65%	51.95%	1.46	1.21E-05	839	35775	1691427	29731	2.82%	2.35%	1.36	1.12E-14
GO:0022803	passive transmembrane transporter activity	160	308	16603	6033	2.65%	51.95%	1.46	1.21E-05	839	35775	1691427	29731	2.82%	2.35%	1.36	1.12E-14
GO:0022838	substrate specific channel activity	155	299	16612	6033	2.57%	51.84%	1.45	2.57E-05	830	35251	1691951	29731	2.79%	2.35%	1.37	6.69E-15
GO:0010558	negative regulation of macromolecule biosynthetic proce	130	251	16660	6033	2.15%	51.79%	1.45	4.52E-04	845	27032	1700170	29731	2.84%	3.13%	1.82	6.51E-55
GO:0010628	positive regulation of gene expression	118	228	16683	6033	1.96%	51.75%	1.45	1.81E-03	766	25309	1701893	29731	2.58%	3.03%	1.76	1.40E-44
GO:0006793	phosphorus metabolic process	345	668	16243	6033	5.72%	51.65%	1.45	2.36E-14	1829	89535	1637667	29731	6.15%	2.04%	1.19	5.56E-10
GO:0006796	phosphate metabolic process	345	668	16243	6033	5.72%	51.65%	1.45	2.36E-14	1829	89535	1637667	29731	6.15%	2.04%	1.19	5.56E-10
GO:0031324	negative regulation of cellular metabolic process	159	308	16603	6033	2.64%	51.62%	1.45	2.43E-05	989	34706	1692496	29731	3.33%	2.85%	1.66	4.03E-47
GO:0010605	negative regulation of macromolecule metabolic process	159	310	16601	6033	2.64%	51.29%	1.44	4.41E-05	1035	35560	1691642	29731	3.48%	2.91%	1.69	1.80E-53
GO:0051173	positive regulation of nitrogen compound metabolic proc	122	238	16673	6033	2.02%	51.26%	1.44	2.30E-03	784	26388	1700814	29731	2.64%	2.97%	1.73	7.08E-43
GO:0009892	negative regulation of metabolic process	166	324	16587	6033	2.75%	51.23%	1.44	2.33E-05	1052	36379	1690823	29731	3.54%	2.89%	1.68	3.91E-53
GO:0016301	kinase activity	329	643	16268	6033	5.45%	51.17%	1.43	8.44E-13	1743	87838	1639364	29731	5.86%	1.98%	1.15	7.01E-06
GO:0051276	chromosome organization	153	300	16611	6033	2.54%	51.00%	1.43	1.36E-04	1182	40303	1686899	29731	3.98%	2.93%	1.70	3.24E-63
GO:0043687	post-translational protein modification	408	801	16110	6033	6.76%	50.94%	1.43	4.57E-16	2422	1E+05	1622647	29731	8.15%	2.32%	1.35	1.84E-44
GO:0009653	anatomical structure morphogenesis	167	333	16578	6033	2.77%	50.15%	1.41	1.53E-04	823	35747	1691455	29731	2.77%	2.30%	1.34	1.54E-12
GO:0003723	RNA binding	267	538	16373	6033	4.43%	49.63%	1.39	5.37E-08	1913	53892	1673310	29731	6.43%	3.55%	2.06	8.52E-184
GO:0006464	protein modification process	524	1061	15850	6033	8.69%	49.39%	1.38	1.10E-17	2945	1E+05	1593708	29731	9.91%	2.21%	1.28	8.34E-39
GO:0043412	biopolymer modification	538	1100	15811	6033	8.92%	48.91%	1.37	4.48E-17	2992	1E+05	1590103	29731	10.06%	2.18%	1.27	3.86E-36
GO:0007275	multicellular organismal development	352	721	16190	6033	5.83%	48.82%	1.37	4.76E-10	1892	81123	1646079	29731	6.36%	2.33%	1.35	2.44E-35
GO:0048856	anatomical structure development	385	800	16111	6033	6.38%	48.13%	1.35	4.61E-10	2026	86940	1640262	29731	6.81%	2.33%	1.35	6.40E-38
GO:0030154	cell differentiation	250	523	16388	6033	4.14%	47.80%	1.34	2.52E-05	1283	55660	1671542	29731	4.32%	2.31%	1.34	3.93E-21
GO:0010468	regulation of gene expression	891	1901	15010	6033	14.77%	46.87%	1.31	5.86E-23	5781	2E+05	1519970	29731	19.44%	2.79%	1.62	1.88E-297
GO:0051252	regulation of RNA metabolic process	788	1689	15222	6033	13.06%	46.65%	1.31	4.81E-19	5224	2E+05	1540428	29731	17.57%	2.80%	1.62	5.09E-267
GO:0006355	regulation of transcription, DNA-dependent	781	1678	15233	6033	12.95%	46.54%	1.30	1.78E-18	5177	2E+05	1541494	29731	17.41%	2.79%	1.62	5.64E-261
GO:0006350	transcription	676	1453	15458	6033	11.21%	46.52%	1.30	2.07E-15	4670	2E+05	1561275	29731	15.71%	2.81%	1.64	2.12E-240
GO:0006996	organelle organization	331	712	16199	6033	5.49%	46.49%	1.30	4.76E-06	2126	91046	1636156	29731	7.15%	2.34%	1.36	1.55E-40
GO:0045449	regulation of transcription	826	1781	15130	6033	13.69%	46.38%	1.30	3.13E-19	5379	2E+05	1530336	29731	18.09%	2.73%	1.59	1.99E-252
GO:0044428	nuclear part	349	756	16155	6033	5.78%	46.16%	1.29	4.57E-06	2068	88972	1638230	29731	6.96%	2.32%	1.35	3.87E-38
GO:0048869	cellular developmental process	298	647	16264	6033	4.94%	46.06%	1.29	1.07E-04	1561	69087	1658115	29731	5.25%	2.26%	1.31	8.89E-23
GO:0080090	regulation of primary metabolic process	969	2111	14800	6033	16.06%	45.90%	1.29	1.61E-21	6066	2E+05	1498267	29731	20.40%	2.65%	1.54	2.38E-256
GO:0060255	regulation of macromolecule metabolic process	967	2107	14804	6033	16.03%	45.89%	1.29	1.96E-21	6117	2E+05	1498501	29731	20.57%	2.67%	1.55	2.53E-269
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nuc	861	1881	15030	6033	14.27%	45.77%	1.28	4.01E-18	5500	2E+05	1520197	29731	18.50%	2.66%	1.54	4.72E-231
GO:0016070	RNA metabolic process	297	649	16262	6033	4.92%	45.76%	1.28	2.46E-04	1809	67942	1659260	29731	6.08%	2.66%	1.55	4.63E-68
GO:0010556	regulation of macromolecule biosynthetic process	872	1906	15005	6033	14.45%	45.75%	1.28	2.48E-18	5653	2E+05	1518832	29731	19.01%	2.71%	1.58	5.05E-260
GO:0051171	regulation of nitrogen compound metabolic process	862	1892	15019	6033	14.29%	45.56%	1.28	2.03E-17	5501	2E+05	1519283	29731	18.50%	2.65%	1.54	5.50E-227
GO:0019222	regulation of metabolic process	1051	2308	14603	6033	17.42%	45.54%	1.28	2.93E-22	6438	2E+05	1480425	29731	21.65%	2.61%	1.52	2.12E-257
GO:0031323	regulation of cellular metabolic process	1011	2222	14689	6033	16.76%	45.50%	1.28	4.82E-21	6255	2E+05	1486607	29731	21.04%	2.60%	1.51	7.73E-245
GO:0032502	developmental process	767	1686	15225	6033	12.71%	45.49%	1.28	7.79E-15	4084	2E+05	1542417	29731	13.74%	2.21%	1.28	2.14E-57
GO:0031326	regulation of cellular biosynthetic process	895	1975	14936	6033	14.84%	45.32%	1.27	2.10E-17	5722	2E+05	1512612	29731	19.25%	2.67%	1.55	1.04E-245
GO:0009889	regulation of biosynthetic process	895	1979	14932	6033	14.84%	45.22%	1.27	4.38E-17	5722	2E+05	1512473	29731	19.25%	2.66%	1.55	4.80E-245
GO:0003677	DNA binding	772	1708	15203	6033	12.80%	45.20%	1.27	4.38E-14	4654	2E+05	1541626	29731	15.65%	2.51%	1.46	4.42E-146
GO:0005634	nucleus	1523	3382	13529	6033	25.24%	45.03%	1.26	1.24E-32	8934	4E+05	1353886	29731	30.05%	2.39%	1.39	1.03E-253
GO:0016043	cellular component organization	567	1262	15649	6033	9.40%	44.93%	1.26	6.97E-09	3381	2E+05	1565035	29731	11.37%	2.08%	1.21	5.65E-27
GO:0003676	nucleic acid binding	1127	2513	14398	6033	18.68%	44.85%	1.26	3.23E-21	6918	3E+05	1456529	29731	23.27%	2.56%	1.48	5.15E-256
GO:0044267	cellular protein metabolic process	691	1544	15367	6033	11.45%	44.75%	1.25	4.49E-11	3604	2E+05	1553648	29731	12.12%	2.08%	1.21	7.11E-28
GO:0034960	cellular biopolymer metabolic process	1577	3565	13346	6033	26.14%	44.24%	1.24	4.77E-29	9096	4E+05	1328381	29731	30.59%	2.28%	1.32	6.09E-194
GO:0044260	cellular macromolecule metabolic process	1596	3619	13292	6033	26.45%	44.10%	1.24	1.14E-28	9157	4E+05	1323694	29731	30.80%	2.27%	1.32	2.90E-189
GO:0008270	zinc ion binding	739	1694	15217	6033	12.25%	43.62%	1.22	4.00E-09	4598	2E+05	1518541	29731	15.47%	2.20%	1.28	1.93E-64
GO:0043284	biopolymer biosynthetic process	799	1841	15070	6033	13.24%	43.40%	1.22	1.54E-09	5093	2E+05	1528020	29731	17.13%	2.56%	1.49	4.67E-178
GO:0034961	cellular biopolymer biosynthetic process	792	1826	15085	6033	13.13%	43.37%	1.22	2.36E-09	5063	2E+05	1529464	29731	17.03%	2.56%	1.49	5.62E-178
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid met	1006	2326	14585	6033	16.67%	43.25%	1.21	1.62E-12	6182	3E+05	1468268	29731	20.79%	2.39%	1.39	4.37E-157
GO:0009059	macromolecule biosynthetic process	806	1865	15046	6033	13.36%	43.22%	1.21	3.67E-09	5109	2E+05	1526089	29731	17.18%	2.54%	1.48	2.79E-173
GO:0034645	cellular macromolecule biosynthetic process	798	1847	15064	6033	13.23%	43.21%	1.21	5.28E-09	5078	2E+05	1527848	29731	17.08%	2.55%	1.48	2.77E-174
GO:0005515	protein binding	2310	5360	11551	6033	38.29%	43.10%	1.21	8.15E-39	11938	6E+05	1128759	29731	40.15%	1.99%	1.16	1.40E-84
GO:0005737	cytoplasm	1114	2610	14301	6033	18.47%	42.68%	1.20	3.09E-12	5589	3E+05	1434532	29731	18.80%	1.91%	1.11	9.08E-14
GO:0043283	biopolymer metabolic process	1706	4002	12909	6033	28.28%	42.63%	1.19	6.79E-22	9631	4E+05	1281161	29731	32.39%	2.16%	1.25	1.06E-139
GO:0043170	macromolecule metabolic process	1713	4039	12872	6033	28.39%	42.41%	1.19	1.04E-20	9645	4E+05	1278529	29731	32.44%	2.15%	1.25	7.58E-135
GO:0046872	metal ion binding	1276	3045	13866	6033	21.15%	41.90%	1.17	1.43E-11	7058	4E+05	1357285	29731	23.74%	1.91%	1.11	9.31E-19
GO:0043167	ion binding	1297	3108	13803	6033	21.50%	41.73%	1.17	3.74E-11	7119	4E+05	1350015	29731	23.94%	1.89%	1.10	6.98E-15
GO:0043169	cation binding	1282	3073	13838	6033	21.25%	41.72%	1.17	6.46E-11	7078	4E+05	1353807	29731	23.81%	1.90%	1.10	2.47E-16
GO:0019538	protein metabolic process	805	1931	14980	6033	13.34%	41.69%	1.17	2.36E-05	4076	2E+05	1511499	29731	13.71%	1.89%	1.10	9.24E-07
GO:0006807	nitrogen compound metabolic process	1082	2597	14314	6033	17.93%	41.66%	1.17	2.46E-08	6413	3E+05	1443423	29731	21.57%	2.26%	1.31	4.31E-116
GO:0043227	membrane-bounded organelle	2165	5263	11648	6033	35.89%	41.14%	1.15	1.31E-19	11134	5E+05	1187811	29731	37.45%	2.06%	1.20	2.87E-113
GO:0043231	intracellular membrane-bounded organelle	2165	5263	11648	6033	35.89%	41.14%	1.15	1.31E-19	11134	5E+05	1187811	29731	37.45%	2.06%	1.20	2.87E-113
GO:0043229	intracellular organelle	2363	5755	11156	6033	39.17%	41.06%	1.15	5.08E-22	12110	6E+05	1131586	29731	40.73%	2.03%	1.18	2.49E-109
GO:0043226	organelle	2363	5757	11154	6033	39.17%	41.05%	1.15	6.61E-22	12110	6E+05	1131304	29731	40.73%	2.03%	1.18	9.78E-109
GO:0046914	transition metal ion binding	849	2070	14841	6033	14.07%	41.01%	1.15	2.97E-04	4943	2E+05	1483367	29731	16.63%	2.03%	1.18	6.61E-31
GO:0044237	cellular metabolic process	1911	4666	12245	6033	31.68%	40.96%	1.15	5.09E-15	10244	5E+05	1232295	29731	34.46%	2.07%	1.20	5.65E-103
GO:0044249	cellular biosynthetic process	962	2364	14547	6033	15.95%	40.69%	1.14	1.89E-04	5687	2E+05	1480872	29731	19.13%	2.31%	1.34	9.07E-116
GO:0009058	biosynthetic process	996	2449	14462	6033	16.51%	40.67%	1.14	1.15E-04	5795	3E+05	1471749	29731	19.49%	2.27%	1.32	2.62E-105
GO:0005488	binding	3764	9279	7632	6033	62.39%	40.56%	1.14	2.50E-45	19233	1E+06	700799	29731	64.69%	1.87%	1.09	3.33E-75
GO:0044238	primary metabolic process	1991	4925	11986	6033	33.00%	40.43%	1.13	7.13E-13	10576	5E+05	1195260	29731	35.57%	1.99%	1.16	5.74E-67
GO:0044424	intracellular part	2934	7302	9609	6033	48.63%	40.18%	1.13	6.48E-23	14840	8E+05	967777	29731	49.91%	1.95%	1.14	3.72E-92
GO:0008152	metabolic process	2123	5432	11479	6033	35.19%	39.08%	1.10	8.61E-07	11134	6E+05	1152279	29731	37.45%	1.94%	1.13	1.01E-48
GO:0050794	regulation of cellular process	1887	4859	12052	6033	31.28%	38.84%	1.09	2.21E-04	10477	5E+05	1216935	29731	35.24%	2.05%	1.19	1.15E-97
GO:0050789	regulation of biological process	1938	5000	11911	6033	32.12%	38.76%	1.09	2.48E-04	10677	5E+05	1204313	29731	35.91%	2.04%	1.19	1.76E-94
GO:0065007	biological regulation	2033	5275	11636	6033	33.70%	38.54%	1.08	6.92E-04	11018	6E+05	1172006	29731	37.06%	1.98%	1.15	1.58E-69
GO:0009987	cellular process	3160	8315	8596	6033	52.38%	38.00%	1.07	1.96E-06	15948	9E+05	852035	29731	53.64%	1.82%	1.06	1.82E-21
GO:0044464	cell part	4060	10772	6139	6033	67.30%	37.69%	1.06	1.34E-09	20233	1E+06	606876	29731	68.05%	1.81%	1.05	4.73E-28
GO:0003674	molecular_function	4450	11835	5076	6033	73.76%	37.60%	1.05	3.66E-12	22032	1E+06	479926	29731	74.10%	1.77%	1.03	5.86E-10
GO:0008150	biological_process	3998	10775	6136	6033	66.27%	37.10%	1.04	1.00E-03	19661	1E+06	609400	29731	66.13%	1.76%	1.02	1.05E-03

S8. Enrichment of exonic splicing enhancer and miRNA target motifs in SCEs

We analyzed the enrichment of short sequence motifs corresponding to exonic splicing enhancers (ESEs) and miRNA target seeds by comparing their frequencies in the 9-codon SCEs compared to other coding regions.

The motifs of interest include:

- 238 hexamers identified as likely ESEs in Fairbrother *et al.* (2002)
- 1,062 7mers corresponding to the reverse complement of positions 2-8 and T+2-7 of all human miRNA mature sequences (Lewis *et al.* 2005) in miRBase v13.0 (Griffiths-Jones *et al.* 2008)

We estimated *P* values for observed enrichments by treating the instances of each motif within SCEs as a hypergeometric sample of all its instances within coding regions. To control for sequence composition effects, we additionally evaluated equal-sized sets of shuffled and reverse-complemented versions of the motifs of interest. To control for frame-specific compositional effects, we also evaluated motifs and controls in the reading frame with the fewest instances of each hexamer/7mer (Stark *et al.* 2007).

		Total in all three frames			Frame with fewest instances		
Motifs	SCEs	Freq in SCEs	Freq in CCDS	Enrich. *P*-value	Freq* in SCEs	Freq* in CCDS	Enrich. *P*-value
ESEs	All	10.47%	10.59%	0.99	9.84%	8.91%	$<10^{-17}$
Rev. comp. ESEs	“	7.38%	6.79%	$<10^{-51}$	8.23%	7.76%	$<10^{-5}$
Shuffled ESEs	“	8.27%	7.99%	$<10^{-11}$	8.19%	7.68%	$<10^{-6}$
ESEs	Spanning exon-exon junctions only	11.62%	10.59%	$<10^{-26}$	12.01%	8.91%	$<10^{-38}$
Rev. comp. ESEs	“	7.74%	6.79%	$<10^{-31}$	8.98%	7.76%	$<10^{-7}$
Shuffled ESEs	“	9.15%	7.99%	$<10^{-40}$	9.92%	7.68%	$<10^{-24}$
miRNA seeds	All	8.88%	8.63%	$<10^{-7}$	9.86%	9.23%	$<10^{-6}$
Rev. comp. seeds	“	7.99%	8.21%	1	8.26%	8.25%	0.45
Shuffled seeds	“	6.81%	6.68%	$<10^{-3}$	7.05%	6.86%	0.04
miRNA seeds	Last coding exon only	9.32%	8.63%	$<10^{-5}$	12.40%	9.23%	$<10^{-5}$
Rev. comp. seeds	“	7.99%	8.21%	0.94	8.99%	8.25%	0.12
Shuffled seeds	“	7.06%	6.68%	0.003	8.06%	6.86%	0.02

Supplementary Table 5. Enrichment of short sequence motifs and controls in SCEs. * as fraction of total instances of all hexamers/7mers in their respective reading frame with the fewest instances.

The miRNA seeds consistently show stronger enrichments than the matched controls, indicating that their increased frequency is not explained by correlated sequence composition biases. The ESEs also show enrichment in SCEs spanning exon-exon junctions, but it is only stronger than the controls for the frame-invariant statistic, making it more difficult to unambiguously distinguish from correlated composition biases.

S9. Preliminary assessment of novel dual-coding ORFs associated with SCEs

We searched for ORFs in alternative reading frames of known CCDS ORFs that (1) are longer than expected based on nucleotide, dinucleotide, *and* codon shuffles of the CCDS sequence; (2) are open (contain no in-frame stop codons) over at least 80% of the length in at least 15 of the 29 species; and (3) overlap 30-codon SCEs. We required ORFs in antisense reading frames to lie contiguously in the genome, since it is unlikely that exon structure would be preserved in both sense and antisense transcription units. Alternate reading frames on the sense strand were allowed to span multiple exons of the annotated transcripts.

The lengths were assessed by shuffling the sequence of each complete CCDS ORF at the nucleotide, dinucleotide, or codon levels, and then determining the longest ORF in any alternate frame of the shuffled sequence. Candidate dual-coding ORFs were discarded if they were not longer than the longest alternate ORF in 95% of shuffled sequences for all three shuffling methods.

Overall, this is a highly conservative strategy for this preliminary assessment (particularly due to the ORF length requirements), which did not detect the known dual-coding ORFs mentioned in the main text. We expect that this can be greatly refined in the future using specialized methods.

Known gene	**CCDS ID**	**Orient.**	**Length (AA)**	**Putative alternate translation**	**Comments**
WAC	CCDS7159.1	Sense	188	CMRGNSRDSVMAVTTGGGTRSLTRHLSIHRRVTPVAVITDMKRCETPEILHHQIKCCGDLIVLKTNTVTAQVTVRPKMCILTELERGMVGPVTLHKKIHTTTVLFIVQIHILLIQAITQAKLQMHLMILQMTGLSILALLGKSTTTIVEQKFHNGKNQKSGLKENRDKKKQTRWQSTASQKIGITEER	Predicted in AF116666
PHF7	CCDS2854.1	Sense	171	GNFFRKTISACIISVLSYLVSCLRGASPTEASMDFCLKTSKRRQPGLLGRSALCARKRELLSTARRISASETSICLVAKKGVAFHNFLESTNHFVTNIAQHRTSNMGMWGRKAASYVVKTYPNRVLRTSRARVVVKPSTTASAYRNMPTHQQSISSNVHSVTIEKSFLKKC	Predicted in AF151060
ZNF3	CCDS43619.1	Sense	240	IPTLSHIRDSPWETDPISVMNVARALIELQTLFNIRESTLGKSPMNVMSVGRPSARAHTLFSIRESTLGKNLMNVVIVGKPSAVALPSFCIGGSTRGRNPMNVMSVGRPSAGAPPSPTIRESTLVRNPTPAMNVGRPSAGAQPLFTIRESTLEKNPMNVMNVGKPSARAHTSISTRESTLERSPTNVWNVEESLPTVQALFSIKESTPGRTPMNVVSVGKPSGTARLLFAIREFTLERSL	High similarity to various other ZNFs
FAM98A	CCDS33179.1	Sense	220	WAGCLTEVVDPMKSNLHPQRCHRGRRGKMAPSSKQEAEEEGEVAMNIPHTEDEEVMNKEAGEVDVVAMTMVAEGEEEEISIKEAGQMEGVVEEVATKMVVIEIQVSSQVAIMVATAVVAIKAEVMVASKHLLHIQEVDTRVVATSRTIDTKMAGTMVIVVVVVVGEVVVEAEVVVQAREEAGEEEGARIITKGVNLNSISSMEVISIIILDLDREDITLV	
UBE2E2	CCDS2637.1	Sense	136	CPLRHKELMTVQALVEEVPMEINVKVFSKNQKENKFSPRKRREKYPAKPLLNCQLVLKEFRRNLQKSHWTLLPTVVLDPKETTFMNGGQLYWDPQDLSMKEGCSFLTLPFHQTIRLNPLRLPSEQESITVILTAKV	

EYA1	CCDS34906.1	Sense	178	TISQVQQLGAVVSAHDQLTSSLHHRFTLPTDHTHIFSLPLPHKLWLHMGKHSLPQECNKLQPMPRTHSQDSRTAFPHMVHCGQASRLKVDCHSLSHLDRQDFSAMAQASVPLNLDRHHTATRCKVAVLQHHQEYIQEIIHSQIPLDLIVHSRTIRLIPVLARVSTHSIITAHRIQHII	
RBMX	CCDS14661.1	Sense	242	KEDHHQEVGVLLLRDLHLQDQFAVAVEWEEELLYHVEEIVMEVHLEGNRCPLVEMFICPQEMMGILLKTAIQAEITQVLVILEIMHHHHEIILTVIMVIPVHVMTIHQEDIAIEMDMVVIVTIQIIQVEVPTEIHMRVMVTHVVLHLHEGPRHLMVEAVAMMITAAHVTDMVEVETVTQAAEVISTQVVVIGLADKKEGFPLLWKGGTLLHVIPTAVQAAEHQEVVAVEEADLIEGEAEADT	Supported by NM_001164803
ALG9	CCDS41714.1	Sense	138	WISEFQCSLCFGSPSPTTDFSYGIPAAEISCSEFRPPVLAYLGSNVYLVYNFLHPASQRGEISFPCVSTYMSLWRCGSLCTSAQFSVLPEMLPLCVSTISPGALYCDIELAGIRNCLPVWALVIFSLCGTVQRISRAP	
SLC35F3	CCDS1600.1	Sense	242	GGSEMVALRGGSHQRAGGAHVHRGADPAHHWLLWLPALGSELQKGGAPPGLPGPGGGPGTGRGGGRRESESPLLDVLPGATQEDLLGRGGRAVRVLLVGGLHAARQADLQEVRRALHPHVVCHQLELFILPVVLRGARLQVHREAVCEAAIQGMLSIFWRQWLDFEGVFYQGSTLWCSLDTHKLPVLTCNKENKHYGCLRVVLLQQSFCVLALMDRSQGQIHGSEDCGRHPRHRWHCDDDLR	
SRCAP	CCDS10689.2	Sense	421	SCVLYSLSSIVIFTPHLCPHHTSCPSLGSTHHPHLSPLDCFCFGPSSVDQCDSTIGTCCPSGSWTSLLGTIWCFPVSISLDSRFGHSSIPVFISDTWSPSVVGSHLFTCSRVELNRGPSMLTCPGASFGSGQSFSVSTKSSSSSGFPSGSSIFCISGSSHPSGSYGGSTDSNSGSFSSSSSGSSSGPGTIARCCSCPGFITDSGSSYGSIVYSRNLFSLSFTGTSSNPCVGSIINSNYATSPGSVTSPEPGFYADTGPSPSFSTHSWRLISISDTLFGNGEPPGTLSNSDIVINSSIIPGTNSSPDTVFGTRTTTGSNSDAVSGSSTPSGSSFSSGPSPSSHADFGSSIVICFTPGPSFSADTDLEPCPSSYPGPGRSSDLGAGPSLHTVPSFPGIFPCGFGIWCRSLACHHGIPAACFQG	
FOXP2	CCDS43635.1	Sense	186	WASSQHICSACVSGHDDSPGDHPSANAADPSATSPVSSAATSPSPTTAGCHAAAATTTRVLQETARAVTSSAFAAAAATAAAATTAATTAAATTTTTAATTAAAAATAAAAATASWKASERAAAAAAAATAIGSPAACLPAAASPDATTPAAAASAQPSASGTHLHSTWPGSTSCPIAASSWLKSC	
SMG7	CCDS1355.1	Sense	188	LPVHPHSSPWSLPSSSQQARVSAPNICYPPACGIFYGLRLHLPSWCFCPRNLSSAYSSLSSRKPGASWETVPHSLQPATALWTRANEPGTSTITATFPATPYIFTSSANSTVYKPAAGSSSNSATTIPYKSCAGFGEKPASPLWIPAVSTGRCLQTAVESPSGSRPIRENYACETALLPSDPRPHKTV	
KLHL31	CCDS34478.1	Antisense	208	LVVLRPLVELGAEALDALLVLLLALVPAPHQVGAAMQRRDARSAHSHLQRRRVAPLAGRGAVALHGEHVHALPARPQLAAAHHVHSVAQRDRAVPAPGCAEVRQLLPRVAGRVVGAHRARVGVGYVASGHQHAAVGDRASVAAARHLQGRLRLPLIGGGHVALQRGQASFCVAAARGVHEPVEHAQAEVRALLVHAGQVYPGVEAGIV	
TRPC3	CCDS3725.1	Antisense	210	VAEGVVPLLLHAVPALAEIVVRPLDPGTLHQQHVHHFVFLAVRRQDDGGDVRREARAVLVVSVEVVVLQLLLTGAQSETLAGREARVVEDGLYDAHVALADGEQQGVANARQVLLLEQQLGHLQVLVAHSQLQRVLAHVVHAVDVQRLGLLQHLAHHWDVAVLGGVEEALLLGGEAGAAVIEHEGRAPDSLAPALLPHHCHASQGWASLH	
ALKBH5	CCDS42272.1	Antisense	256	EHSHASPAHRQGKHWFRHPNRLELELAAEAQRRVAKEGHGHDGALEDVDGVHVRHDAAAGLVVVDDGAVDEALGDDAVLHQLLDHQLVHPLGDLVHVARRVEALLAGPALLQLGAVSVAFAEEVLVAQWGPVHRVLVVQALLSAGHHLVNAGLDLGALVLAEEAHLADAALHLARLLLLLQLLLLIVAALGVRVLLILALGPGHRVRFGGSGGCGCGGYGGCGSGGGLPAARLIVVPGRHGLELLTQVRVAAGGRH	Evidence for antisense transcription

RSBN1	CCDS862.1	Antisense	234	SLGGFDQRHPSSPLAVFLLVLVLEIGSAGAGITIGAAVVMVLVPLLVALMLGPGRAHRQRREGEEGRGAWGRERGSDSGPGGCQRRWRRRRQEKRLNRAWDSWGCIRWRQRPRMLAAATPPVRGSERGLLARPPPLLLRPPTARSALPFNSPRGDPSMRFLLPLFVRLLLLRRHRPYYAHRSDLRSHFTHKHTFKWPDRPPVRASRKRRPRTARWALESLLRPPLVGRSSSRDEH	
SON	CCDS13629.1	Antisense	384	TEISSATVATGCPGNSGAPGCPGNSSTPVARGCPKSSSAPAAPDCPENSSAPAAMSGPGNSSAPVATGCPDNSSALVAEGSPGNSGTPVTEGCPGNSSAVVATGCPGNSSTIVASGCPSNPVAVVTSGCPGCSTVVIPTGCSNSVVVTGCSVNSIATVTAGCSGNSTAAVTKGSPGNSCAITGGSGTSCGGSNPMDGAGSPGNSCDNCGTGVTEGPGNSGTAVAGGPGNSGTGVERGPGNSGTGVAEGPGNSSTGVTGGPCNSGMEVAGGPGNSIAEATDDSCSSNAVVLGSSGNSCGLVMDESAISDGTSTGCSGNLSASIADDSGKSIVVDVLGSGYTSVGVSDDTTRVSEGSNTFATGGSTNMIFDGSGDVLSTDFSTDFI	Hit to PRK07003 (DNA PolIII subunit) domain, weak similarity to collagen proteins
TNKS	CCDS5974.1	Antisense	209	AAKCRGEDFRPAMSFAFTFAASTSLFTRDTSPLRQASSSSRRAPLTAGTPVPGPAAPGPSGAVLLTPAASGLSAKLDPGEDGEEDEEEVGEEEDDGELLLPLPAGLGATPAADEVETAGTTGAAATVQIVLVVLQQLVPSTGSGDRGLSGGSRLPSPSGSARPCRGEAKGARPLAVGEAGVVPGARPGLSGGGGGGGGAEAPGAGWSCC	
FAM135A	CCDS34481.1	Antisense	197	VSLSTVGISEYICVRFNTSLFFFIVLSSLTSVLLEMEPEMTQGVPKQISFEFNVIAVPVELLDLSGLWKPLHIVSKLGLELQLLFTSLEVSSSEKTNCKGSLMLEFKECIATVSLSKFPENTTLCCCKVILDFLVSSHFGKSVVPRFNTVLLSFRWHSGTTDFLEFSSGDSSSSEGREFMDVKDDCISLIALVLSVQ	

Supplementary Table 6. Candidate novel dual-coding ORFs associated with SCEs.

S10. A to I editing sites in SCEs

Table S1 in the supplement of Li et al. (2009) lists fourteen previously known A-to-I recoding sites in human genes. All of these are found within CCDS ORFs, and ten lie within the 15-codon SCEs:

Gene name	Genomic position (hg18)	Amino acid change	SCE?
GRIA2	chr4:158477325	Q>R	√
GRIA2	chr4:158500744	R>G	√
IGFBP7	chr4:57670991	K>R	
IGFBP7	chr4:57671043	R>G	
CYFIP2	chr5:156669386	K>E	
GRIA4	chr11:105309904	R>G	√
KCNA1	chr12:4892003	I>V	√
BLCAP	chr20:35580947	K>R	√
BLCAP	chr20:35580977	Q>R	√
BLCAP	chr20:35580986	Y>C	√
GRIK1	chr21:29875621	Q>R	√
GRIA3	chrX:122426643	R>G	
GABRA	chrX:151108975	I>M	√
FLNA	chrX:153233144	Q>R	√

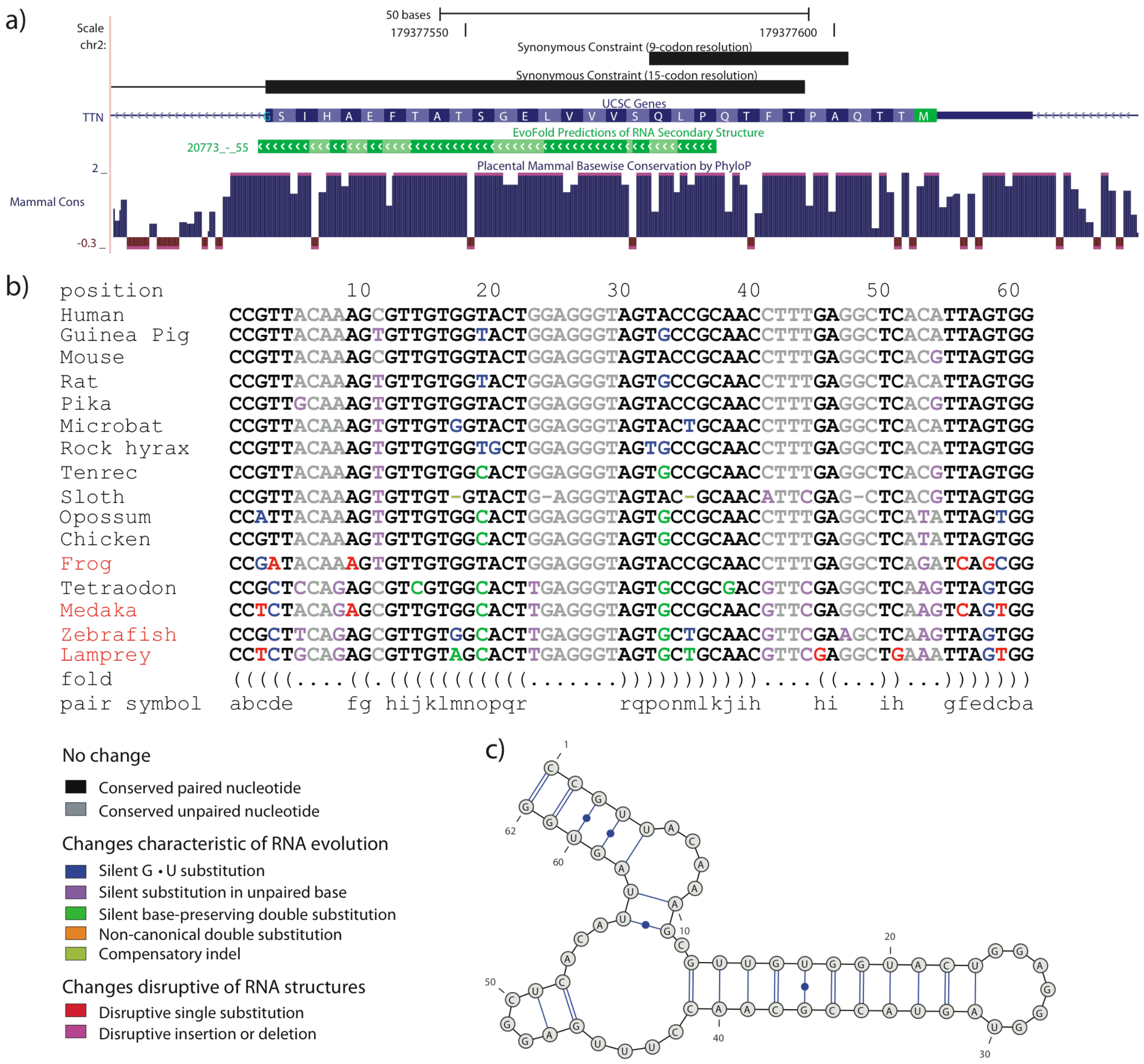
Supplementary Table 7. Fourteen known A-to-I recoding sites from Table S1 of Li et al. (2009)**, ten of which lie within SCEs.**

The high-throughput sequencing approach in that study identified 54 editing sites in human ORFs, of which 48 lie within CCDS ORFs, of which 40 are novel (not included in the above list). These are listed in their Table S3. Of those forty, the following three lie within SCEs:

Gene name	Genomic position (hg18)	Amino acid change
CADPS	chr3:62398847	E>G
FLNB	chr3:58116841	Q>R
GRIA2	chr4:158477329	None (Q>Q)

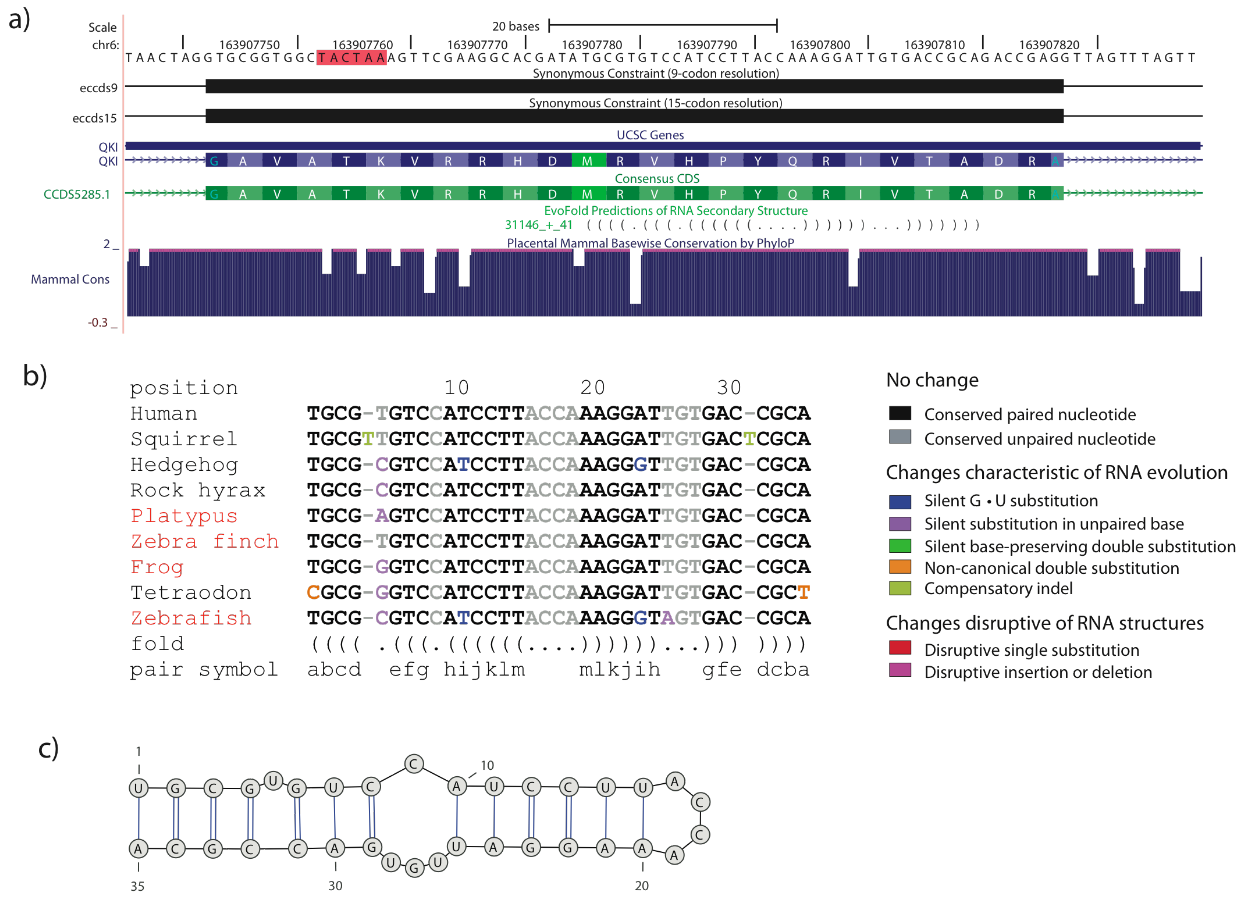
Supplementary Table 8. Three novel A-to-I editing sites within CCDS ORFs identified by Li et al. (2009) **lie within SCEs.**

S11. RNA structures in SCEs


a)
Scale
chr2:
50 bases
179377550
179377600
Synonymous Constraint (9-codon resolution)
Synonymous Constraint (15-codon resolution)
UCSC Genes
TTN
G S I H A E F T A T S G E L V V V S Q L P Q T F T P A Q T T M
EvoFold Predictions of RNA Secondary Structure
20773_-_55
Placental Mammal Basewise Conservation by PhyloP
2 _
Mammal Cons
-0.3 _
b)
No change
■ Conserved paired nucleotide
■ Conserved unpaired nucleotide
Changes characteristic of RNA evolution
■ Silent G • U substitution
■ Silent substitution in unpaired base
■ Silent base-preserving double substitution
■ Non-canonical double substitution
■ Compensatory indel
Changes disruptive of RNA structures
■ Disruptive single substitution
■ Disruptive insertion or deletion
c)
1
10
20
30
40
50
60
62

```
position               10        20        30        40        50        60
Human       CCGTTACAAAGCGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGGCTCACATTAGTGG
Guinea Pig  CCGTTACAAAGTGTTGTGGTACTGGAGGGTAGTGCCGCAACCTTTGAGGCTCACATTAGTGG
Mouse       CCGTTACAAAGCGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGGCTCACGTTAGTGG
Rat         CCGTTACAAAGTGTTGTGGTACTGGAGGGTAGTGCCGCAACCTTTGAGGCTCACATTAGTGG
Pika        CCGTTGCAAAGTGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGGCTCACGTTAGTGG
Microbat    CCGTTACAAAGTGTTGTGGTACTGGAGGGTAGTACTGCAACCTTTGAGGCTCACATTAGTGG
Rock hyrax  CCGTTACAAAGTGTTGTGGTGCTGGAGGGTAGTGCCGCAACCTTTGAGGCTCACATTAGTGG
Tenrec      CCGTTACAAAGTGTTGTGGCACTGGAGGGTAGTGCCGCAACCTTTGAGGCTCACGTTAGTGG
Sloth       CCGTTACAAAGTGTTGT-GTACTG-AGGGTAGTAC-GCAACATTCGAG-CTCACGTTAGTGG
Opossum     CCATTACAAAGTGTTGTGGCACTGGAGGGTAGTGCCGCAACCTTTGAGGCTCATATTAGTGG
Chicken     CCGTTACAAAGTGTTGTGGCACTGGAGGGTAGTGCCGCAACCTTTGAGGCTCATATTAGTGG
Frog        CCGATACAAAGTGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGGCTCAGATCAGCGG
Tetraodon   CCGCTCCAGAGCGTCGTGGCACTTGAGGGTAGTGCCGCGACGTTCGAGGCTCAAGTTAGTGG
Medaka      CCTCTACAGAGCGTTGTGGCACTTGAGGGTAGTGCCGCAACGTTCGAGGCTCAAGTCAGTGG
Zebrafish   CCGCTTCAGAGCGTTGTGGCACTTGAGGGTAGTGCTGCAACGTTCGAAGCTCAAGTTAGTGG
Lamprey     CCTCTGCAGAGCGTTGTAGCACTTGAGGGTAGTGCTGCAACGTTCGAGGCTGAAATTAGTGG
fold        ((((....((.((((((((((.......))))))))))....((...))...)))))))
pair symbol abcde    fg hijklmnopqr       rqponmlkjih    hi   ih   gfedcba
```

Supplementary Figure 2. RNA secondary structure in SCEs of titin (*TTN*). a) SCEs (black) near the start codon in one the isoforms of the gene titin (purple) overlap well-supported RNA secondary structure predicted by EvoFold (green). b) Selected sequences from the genomic alignment are shown together with the predicted secondary structure in parenthesis format. Part of the structure (the stem starting at position 13) is extremely well supported by compensatory substitutions (green). A number of the shown species were not used for structure inference and substitutions in these may provide independent evidence for the structure (e.g., the case for Lamprey). A compensatory deletion has taken place in Sloth, which also provides strong evidence for the selection for maintaining the structure. c) Secondary structure drawing of the structure.


a)
Scale
chr6:
20 bases
163907750
163907760
163907770
163907780
163907790
163907800
163907810
163907820
TAACTAGGTGCGGTGGCTACTAAAGTTCGAAGGCACGATATGCGTGTCCATCCTTACCAAAGGATTGTGACCGCAGACCGAGGTTAGTTTAGTT
eccds9
Synonymous Constraint (9-codon resolution)
eccds15
Synonymous Constraint (15-codon resolution)
UCSC Genes
QKI
QKI
G A V A T K V R R H D M R V H P Y Q R I V T A D R A
Consensus CDS
CCDS5285.1
G A V A T K V R R H D M R V H P Y Q R I V T A D R A
EvoFold Predictions of RNA Secondary Structure
31146_+_41 ((((.(((.((((((....)))))))...)))))))
Placental Mammal Basewise Conservation by PhyloP
2 _
Mammal Cons
-0.3 _
b)
c)
1
10
20
30
35
U G C G U G U C C A U C C U U A C C A A A G G A U U G U G C A

```
position                    10        20        30
Human          TGCG-TGTCCATCCTTACCAAAGGATTGTGAC-CGCA
Squirrel       TGCGTTGTCCATCCTTACCAAAGGATTGTGACTCGCA
Hedgehog       TGCG-CGTCCATCCTTACCAAAGGGTTGTGAC-CGCA
Rock hyrax     TGCG-CGTCCATCCTTACCAAAGGATTGTGAC-CGCA
Platypus       TGCG-AGTCCATCCTTACCAAAGGATTGTGAC-CGCA
Zebra finch    TGCG-TGTCCATCCTTACCAAAGGATTGTGAC-CGCA
Frog           TGCG-GGTCCATCCTTACCAAAGGATTGTGAC-CGCA
Tetraodon      CGCG-GGTCCATCCTTACCAAAGGATTGTGAC-CGCT
Zebrafish      TGCG-CGTCCATCCTTACCAAAGGGTAGTGAC-CGCA
fold            ((((  .(((.((((((....)))))))...))) ))))
pair symbol    abcd  efg hijklm    mlkjih   gfe dcba
```

No change

- ■ Conserved paired nucleotide
- ■ Conserved unpaired nucleotide

Changes characteristic of RNA evolution

- ■ Silent G • U substitution
- ■ Silent substitution in unpaired base
- ■ Silent base-preserving double substitution
- ■ Non-canonical double substitution
- ■ Compensatory indel

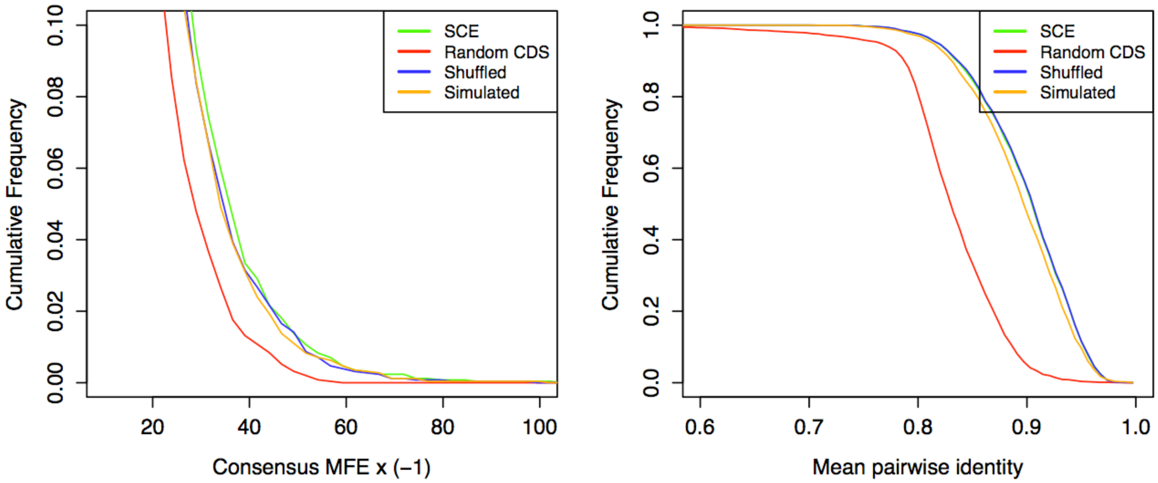
Changes disruptive of RNA structures

- ■ Disruptive single substitution
- ■ Disruptive insertion or deletion

Supplementary Figure 3. Hairpin in SCEs of RNA binding gene. a) SCEs (black) overlap an EvoFold hairpin prediction (green) in the RNA-binding gene *QKI* (purple and green). The protein product of QKI binds to the 5'-NACUAAY-N(1,20)-UAAY-3' motif, part of which (NCAUAA) is found just upstream of the hairpin (highlighted in red) (Galarneau and Richard 2005)**. b) Unique sequences from the genomic alignment are shown together with the predicted hairpin in parenthesis format. Substitutions are color-coded according to their effect on the structure. Insertions and deletions (indels) are ignored by EvoFold and the double insertion in squirrel therefore strongly supports the structure. Species not used for structure inference are shown with red names. c) Secondary structure drawing.**

EvoFold mutual enrichment. We initially studied the general association between SCEs and RNA secondary structures by intersecting with EvoFold predictions based on the 29 mammals and other species (Parker, Moltke, and Pedersen, in prep., Pedersen *et al.* 2006). We found that 8.2% of the SCEs (9-codon resolution) overlap predictions of conserved RNA structures by EvoFold, compared to only 2.6% of corresponding random coding regions (3.2-fold enrichment). The SCEs conversely overlap 11% of the ~9,000 EvoFold predictions lying within CCDS ORFs (3.2-fold enrichment), providing additional evidence for overlapping functions in those regions. However, some mutual enrichment is expected since EvoFold is influenced by conservation and sequence composition, both of which are obviously different in SCEs compared to other coding regions. Therefore, we also undertook a more carefully controlled thermodynamic stability analysis.

RNA thermodynamic stability analysis. We used RNAz (Gruber *et al.* 2010) to calculate the "consensus minimum free energy" (consensus MFE) in the SCEs and random controls. RNAz is based on the RNAalifold algorithm (Hofacker *et al.* 2002) that predicts the optimal folding energy by incorporating phylogenetic information (such as consistent or compensatory mutations) into the classical energy model. Following the convention for MFEs of single sequences, lower (i.e. more negative) consensus MFE indicate a more stable/more conserved RNA secondary structure in an alignment. As control sets we (i) chose random coding regions in the genome with comparable length distribution and total number, (ii) shuffled the codons in the alignments by extending the algorithm introduced in Washietl & Hofacker (2004), and (iii) simulated random alignments with the same average dinucleotide content (Gesell and Washietl 2008).



Supplementary Figure 4. Consensus minimum free energy (MFE) and mean pairwise identity distributions for 15-codon SCEs and control sets. The SCEs were extended by flanking regions of 15 nucleotides since the boundaries of the predicted constrained regions are unlikely to correspond exactly to potential RNA structures. Similar analysis with the other SCE prediction sets and flanking regions of 30 nucleotides gave comparable results (not shown).

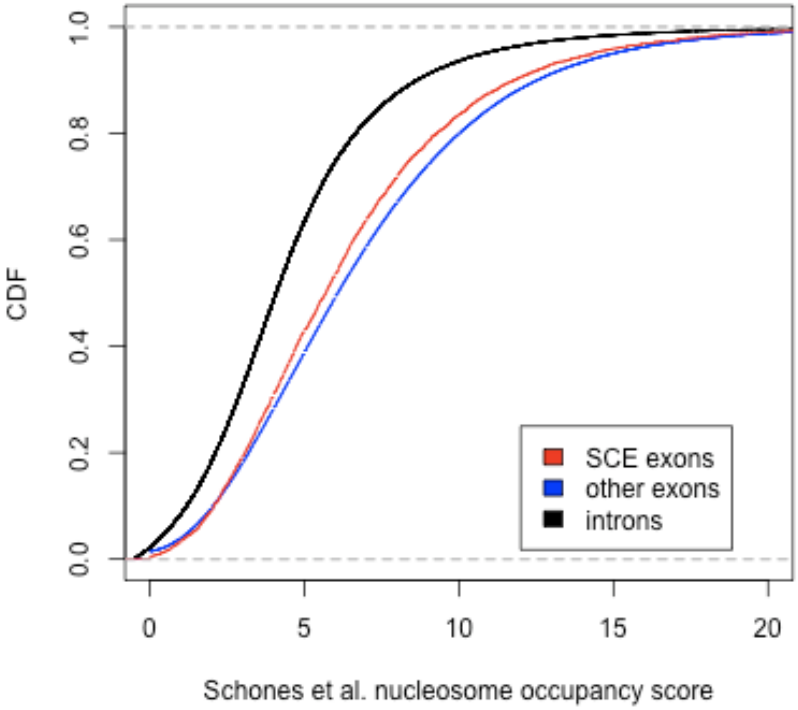
Although the consensus MFE distribution for SCEs is better than random coding regions, it is not distinguishable from shuffled and simulated controls. Therefore, the difference between the SCEs and random coding regions is probably largely explained by the difference in sequence composition and conservation levels to which the consensus MFE is relatively sensitive. From these results it is unlikely that the SCEs contain a high fraction of extremely stable RNA secondary structures. Structural clustering of the EvoFold predictions within SCEs also did not reveal compelling families comparable to those discovered in non-coding regions.

It should be noted that the power of this analysis is severely limited by the high conservation levels (around 90% mean pairwise identity) and the corresponding lack of information from mutational patterns. For many highly conserved alignments the analysis is effectively the same as single sequence analysis, which is generally not significant enough to detect most of the known functional RNA structures (Washietl and Hofacker 2004).

S12. Nucleosome positioning in exons containing SCEs

To assess nucleosome positioning within exons containing SCEs, we downloaded the "Nucleosome scores profile" from
http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx
which is postprocessed from high-throughput sequencing of nucleosome-bound DNA in resting human CD4+ T-cells (Schones *et al.* 2008). This data provides a score for each 10bp window in the genome, indicating the total number of reads mapping at an appropriate distance upstream or downstream to indicate nucleosome occupancy.

For each CCDS exon, we computed the mean score of all the 10bp windows of which at least one-half are within the exon, and compared the distribution of this mean score for exons containing at least one-half of an SCE and other exons. For reference, we also computed the mean score in flanking introns.



Consistent with recent reports, exons have higher nucleosome occupancy than flanking introns. However, exons containing SCEs (red, 15-codon resolution) tend to have lower nucleosome occupancy than other exons (blue, $P = 1.0\times10^{-11}$, Mann-Whitney U).

S13. BED files for synonymous constraint elements

The BED files specifying the locations of the synonymous constraint elements are available from the following web site:

http://compbio.mit.edu/SCE/

S14. References

Arvestad L and Bruno WJ. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. *J. Mol. Evol.* **45:** 696-703.

Benjamini Y and Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29:** 1165-1188.

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

Chamary JV, Parmley JL, and Hurst LD. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7:** 98-108. doi:10.1038/nrg1770.

Fairbrother WG, Yeh RF, Sharp PA, and Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007-1013. doi:10.1126/science.1073774.

Fuglsang A. 2006. Estimating the "effective number of codons": The wright way of determining codon homozygosity leads to superior estimates. *Genetics* **172:** 1301-1307. doi:10.1534/genetics.105.049643.

Galarneau A and Richard S. 2005. Target RNA motif and target mRNAs of the quaking STAR protein. *Nat. Struct. Mol. Biol.* **12:** 691-698. doi:10.1038/nsmb963.

Galtier N and Duret L. 2007. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends in Genetics* **23:** 273-277. doi:DOI: 10.1016/j.tig.2007.03.011.

Gesell T and Washietl S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9:** 248. doi:10.1186/1471-2105-9-248.

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

Griffiths-Jones S, Saini HK, van Dongen S, and Enright AJ. 2008. miRBase: Tools for microRNA genomics. *Nucl. Acids Res.* **36:** D154-158. doi:10.1093/nar/gkm952.

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

Hofacker IL, Fekete M, and Stadler PF. 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319:** 1059-1066. doi:10.1016/S0022-2836(02)00308-X.

Kosiol C, Holmes I, and Goldman N. 2007. An empirical codon model for protein sequence evolution. *Mol. Biol. Evol.* **24:** 1464-1479. doi:10.1093/molbev/msm064.

Lewis BP, Burge CB, and Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20. doi:10.1016/j.cell.2004.12.035.

Li JB, Levanon EY, Yoon J, Aach J, Xie B, LeProust E, Zhang K, Gao Y, and Church GM. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324:** 1210-1213. doi:10.1126/science.1170995.

Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, and 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.

Pond SK and Muse SV. 2005. Site-to-site variation of synonymous substitution rates. *Mol. Biol. Evol.* **22:** 2375-2385. doi:10.1093/molbev/msi232.

Pozzoli U, Menozzi G, Comi GP, Cagliani R, Bresolin N, and Sironi M. 2007. Intron size in mammals: Complexity comes to terms with economy. *Trends in Genetics* **23:** 20-24. doi:DOI: 10.1016/j.tig.2006.10.003.

Richard GF, Kerrest A, and Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* **72:** 686-727. doi:10.1128/MMBR.00011-08.

Schmid K and 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, and Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132:** 887-898. doi:DOI: 10.1016/j.cell.2008.02.022.

Siepel A and Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21:** 468-488.

Stanley S, Bailey T, and Mattick J. 2006. GONOME: Measuring correlations between GO terms and genomic positions. *BMC Bioinformatics* **7:** 94.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature* **450:** 219-232. doi:10.1038/nature06340.

Storey JD and Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **100:** 9440-9445. doi:10.1073/pnas.1530509100.

Urrutia AO and Hurst LD. 2003. The signature of selection mediated by expression on human genes. *Genome Research* **13:** 2260-2264. doi:10.1101/gr.641103.

Washietl S and Hofacker IL. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* **342:** 19-30. doi:10.1016/j.jmb.2004.07.018.

Whelan S and Goldman N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **16:** 1292-1299.

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87:** 23-29. doi:DOI: 10.1016/0378-1119(90)90491-9.

Yang Z, Nielsen R, Goldman N, and Pedersen AK. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431-449.