Dispensability of mammalian DNA

Cory McLean and Gill Bejerano

Genome Res. 2008 18: 1743-1751
Access the most recent version at doi:10.1101/gr.080184.108

Letter

Dispensability of mammalian DNA

Cory McLean[1] and Gill Bejerano[1,2,3]

[1]Department of Computer Science, Stanford University, Stanford, California 94305, USA; [2]Department of Developmental Biology, Stanford University, Stanford, California 94305, USA

In the lab, the *cis*–regulatory network seems to exhibit great functional redundancy. Many experiments testing enhancer activity of neighboring *cis*–regulatory elements show largely overlapping expression domains. Of recent interest, mice in which *cis*–regulatory ultraconserved elements were knocked out showed no obvious phenotype, further suggesting functional redundancy. Here, we present a global evolutionary analysis of mammalian conserved nonexonic elements (CNEs), and find strong evidence to the contrary. Given a set of CNEs conserved between several mammals, we characterize functional dispensability as the propensity for the ancestral element to be lost in mammalian species internal to the spanned species tree. We show that ultraconserved-like elements are over 300-fold less likely than neutral DNA to have been lost during rodent evolution. In fact, many thousands of noncoding loci under purifying selection display near uniform indispensability during mammalian evolution, largely irrespective of nucleotide conservation level. These findings suggest that many genomic noncoding elements possess functions that contribute noticeably to organism fitness in naturally evolving populations.

[Supplemental material is available online at www.genome.org.]

Comparison to other mammals finds at least 5% of the human genome evolving under purifying selection (Waterston et al. 2002). Two-thirds of this genomic mass does not code for protein (Cooper et al. 2004). Strong sequence conservation among mammals and conservation to other vertebrates have repeatedly proven good indicators of gene *cis*-regulatory function (Pennacchio et al. 2006). Ultraconserved elements are some of the genomic regions most strongly conserved during mammalian evolution, having resisted all changes between human, mouse, and rat, along contiguous stretches ≥200 base pairs (bp) (Bejerano et al. 2004). Point mutations in these regions are under extreme purifying selection in the human population (Katzman et al. 2007). And yet, transgenic mice homozygous for four independent deletions of *cis*-regulatory ultraconserved elements are viable and fertile in the lab, and display no obvious phenotypic abnormalities (Ahituv et al. 2007).

There are two compatible hypotheses that may explain this lack of phenotype. One suggests that ultraconserved element deletions do not affect organism viability because of functional redundancy among the large number of mammalian *cis*-regulatory loci (Visel et al. 2008). The other posits that ultraconserved element deletions most likely lead to a selective disadvantage that either does not manifest itself under lab conditions (Barbaric et al. 2007), or is too subtle to be observed in the lab but is large enough that evolutionary pressures eliminate the mutation from the population (Garcia-Dorado et al. 2003). We set out to investigate the extent to which the loss of nonexonic sequence under purifying selection is tolerated in natural populations on an evolutionary timescale. While elements with complete functional redundancy may be dispensable, the loss of elements that provide unique functions may be deleterious and will not fix in natural populations.

By surveying genomic losses during the last 100 million years of eutherian (placental mammal) evolution, we calculate evolutionary loss rate as a function of conservation. We find that ultraconserved-like elements, along with many thousands of additional noncoding loci under purifying selection, are hundreds of fold less likely to be lost than neutrally evolving DNA. These results suggest that the vast majority of *cis*-regulatory sequences possess one or more unique functions whose loss is deleterious. Furthermore, a very weak relationship between sequence conservation and evolutionary loss rate indicates that conservation %id (percentage of identical alignment columns) is a poor measure of the relative importance of conserved nonexonic elements (CNEs).

[3]Corresponding author.
E-mail bejerano@stanford.edu; fax (650) 725-2923.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.080184.108. Freely available online through the *Genome Research* Open Access option.

Results

Ultraconserved-like element loss rates

To analyze the evolutionary pressures against CNE losses, we quantify the extent to which rodents have discarded sequences strongly conserved in multiple mammalian species. The computational pipeline to discover rodent-specific losses of mammalian conserved DNA is divided into three conceptual steps: discovery of strongly conserved mammalian sequence, identification of conserved sequence absent in rodents, and removal of assembly, alignment, and structural RNA migration artifacts (Supplemental Fig. S1). We use whole genome sequences of two rodents (mouse and rat), two primates (human and macaque), and a close carnivore outgroup species (dog). By examining sequences conserved between human, macaque, and dog, with no orthologous DNA in either mouse or rat, we infer by parsimony that the sequence losses were fixed in the rodent lineage prior to the mouse/rat split (Supplemental Fig. S2). Forty-six percent (1327 Mb) of the human genome aligns at 50%id or greater to macaque and dog. To avoid assembly and alignment artifacts, we examine only the 38.4% (1108 Mb) of the human genome alignable to macaque and dog that possesses unique orthologous sequence in both rodents. We rely on the UCSC chaining and netting algorithm to infer orthologous rodent sequences (Kent et al. 2003), and ignore regions where the chains and nets do not provide an unambiguous ortholog in both rodents. By excluding these regions from

consideration we slightly bias against regions with paralogs, though the total effect on our analysis is small (Supplemental Methods S1).

By sliding a 100 bp window over the alignment, we compute the overall level of primate-dog sequence conservation in the form of percentage of identical alignment columns (%id). The choice of 100 bp as the sliding window size provides enough resolution to identify distal functional elements and has proven effective in previous functional studies (Stone et al. 2005). We then find all such windows that are completely lost in both rodents. By requiring complete window losses, we focus the study on losses of longer functional units, to directly compare to the ultraconserved element knockout experiments (Ahituv et al. 2007). Smaller losses would eventually fall in the realm of individual binding site turnover, and potentially be subject to gap placement artifacts. In total, 261 Mb of pan-eutherian sequence have no orthologous rodent DNA. We divide this set into coding and nonexonic (neither coding nor untranslated region [UTR]) windows, and compute the fraction of eutherian conserved bases lost in rodents per %id conservation bin.

The nonexonic alignment, capturing 95.1% of all alignable bases, includes virtually all mammalian *cis*-regulatory elements characterized to date. We define ultraconserved-like sequences as contiguous regions of 100 bp or more perfectly conserved between the primates and dog. The nonexonic ultraconserved-like sequences encompass 1,691,090 bp of the human genome in regions up to 1016 bp in length. Of these regions, only 1447 bp are identified as lost in rodents, in stretches no longer than 184 bp (Supplemental Fig. S3). Thus, the rodent loss rate of nonexonic ultraconserved-like elements is $1447/1{,}691{,}090 \approx 0.086\%$.

At 50%id–65%id, most aligning sequences are expected to be neutrally evolving (Margulies et al. 2005). Roughly a quarter of all neutrally evolving primate-dog nonexonic sequence (135,415,463 of 515,843,585 bp at 60%id) is lost in rodents (Fig. 1A,E), making nonexonic ultraconserved-like elements over 300-fold more indispensable than neutral sequence.

Dispensability of neutral DNA

While we expect the neutrally evolving sequence loss rate to be essentially uniform, we observe a rodent tendency to retain more such sequences the less conserved they are (50%id–65%id range in Fig. 1E). We show that this phenomenon stems from an enrichment of less conserved sequences immediately adjacent to coding exons. We annotate each 100 bp window of the least conserved alignable sequence (50%id–65%id) with its distance to the nearest exon, rounded up to the nearest kilobase. We then compute the fraction of all nonexonic sequences aligned at 50%id–65%id for each distance (truncated at 20 kb). We repeat the analysis for highly-conserved sequences (85%id–100%id). The results (Fig. 2A) show that poorly-conserved sequences are over-represented near exons. Elements within 1 kb of the nearest exon are markedly less likely to be lost in rodents (Fig. 2B), suggesting that exons do shield flanking DNA from many deletions that result in gene structure disruption. In line with theoretical predictions (Petrov 2002), larger deletions are biased away from gene structures, and consequently away from the least conserved alignable sequence. Loss rates for sequences of decreasing maximum size show a monotonic decrease in slope magnitude (Fig. 2C) as the probability to disrupt gene structures is lessened. A restriction to losses of at most 1000 bp shows a nearly constant rodent-specific neutral sequence loss rate (Fig. 2D).

Rodent dispensability under extended branch lengths

As we move from 100% identical primate-dog sequences into less conserved alignments, we expect the mixing of two populations: DNA under purifying selection, and the slowly mutating tail of neutrally evolving sequence (Fig. 1C). This mixing, first observed genome-wide in the seminal mouse genome paper (Fig. 28 in Waterston et al. 2002), must at least partially account for the increase in rodent loss rate we observe the further we move from 100% primate-dog identity (Fig. 1E). Interestingly, intronic and intergenic sequences face similar evolutionary pressures against rodent-specific loss (Supplemental Fig. S4).

To separate the loss rate of DNA under purifying selection from slowly-mutating neutrally evolving loci, we search for eutherian tree topologies that will increase the cumulative branch lengths between the three aligning species. The cumulative branch lengths of a human-dog-horse alignment are longer than the branch lengths of a human-macaque-dog alignment (Supplemental Fig. S2a), which suggests that neutrally-evolving sequence will align at a lower conservation %id in the human-dog-horse alignment.

Forty-one percent (1182 Mb) of the human genome aligns at 50% or more to dog and horse; 35% (1009 Mb) has a unique homolog in both mouse and rat. Nested elements lost in rodents comprise 7.25% (209 Mb) of the human genome.

The abundance of nonexonic sequence in the human-dog-horse alignment (Supplemental Fig. S5B) shows a moderate shift in neutrally-evolving sequence toward lower conservation %id values when compared to its human-macaque-dog counterpart (Supplemental Fig. S5A). The rodent-specific loss rates of highly-conserved sequence are lower in the human-dog-horse alignment (compare Supplemental Fig. S5D,E), indicating that the ratio of functional sequence to highly-conserved neutral sequence is larger in the human-dog-horse alignment than in the human-macaque-dog alignment. However, the slowly evolving tail of neutral sequences may still contribute to correlate loss events with sequence conservation in Supplemental Figure S5E.

Dispensability of DNA under purifying selection

To investigate whether a correlation exists between conservation %id and dispensability for noncoding sequence under purifying selection, we must ensure the neutral rate of evolution is sufficiently large to separate sequence under purifying selection from the slowly mutating tail of neutrally evolving sequence (Waterston et al. 2002). To increase the neutral rate sufficiently we focus on mouse-rat-dog conserved elements deleted in the primate lineage. While the divergence time between human and macaque is comparable to that of mouse and rat, the two lineages exhibit markedly different substitution rates (Supplemental Fig. S2). Rodents have evolved much faster than primates, both prior to and following their respective splits (Gaffney and Keightley 2006). Consequently, only 22.3% of the mouse genome (573 Mb) is spanned by regions alignable to rat and dog, and 20.5% (527 Mb) has a unique homolog in both human and macaque. Only 0.598% (15.3 Mb) of 100 bp alignment windows is additionally lost in both primates.

By nearly doubling the neutral branch lengths of the three genomes we align, we expect the dominant curve of neutrally evolving eutherian DNA to shift dramatically toward lower conservation values. Indeed, the peak of the nonexonic abundance curve shifts from 65%id for primate-dog to well below the 50%id marker of reliable alignability for rodent-dog (Fig. 1A–D).

Based on the hypothesis that most losses of functional elements will not fix, we expect the curve of sequences under purifying selection to roughly retain its overall volume under increased branch lengths. Furthermore, the mass of sequence should remain at similar conservation values, as most tolerated mutations are the relative few that confer little to no fitness change. To test these predictions we examine the preservation of coding exons, the largest well-annotated genomic functional set, between our three tree configurations. With increasing branch lengths, we observe both the predicted volume preservation and mild skew (Fig. 3).

Thus, we conclude that most of the nonexonic 42.8 Mb at 80%id–100%id between rodents and dog in Figure 1B are under purifying selection. Of the 195,163 genomic loci in this set, only 301 (0.154%) are lost in the primate lineage (Fig. 1F). By analyzing the relationship between conservation %id and loss rate, we see that the loss rate (nonexonic bp lost in primates/nonexonic bp conserved in rodent-dog) at 100%id is 0, at 80%id is 0.00122, and the least-squares best-fit line of the data has a slope magnitude of only 0.00006 (Fig. 4A). The minuscule slope magnitude suggests pan-mammalian indispensability of nearly all nonexonic DNA under purifying selection, largely irrespective of substitution rate. Though we are restricted to sequence of 80%id or greater when examining the loss rate of nonexonic DNA under purifying selection to ensure the elements are functional, coding exon sequences conserved at lower %id can be analyzed. The primate loss rate of nonexonic sequence is also comparable to the primate loss rate of coding regions (0 at 100%id, 0.0000861 at

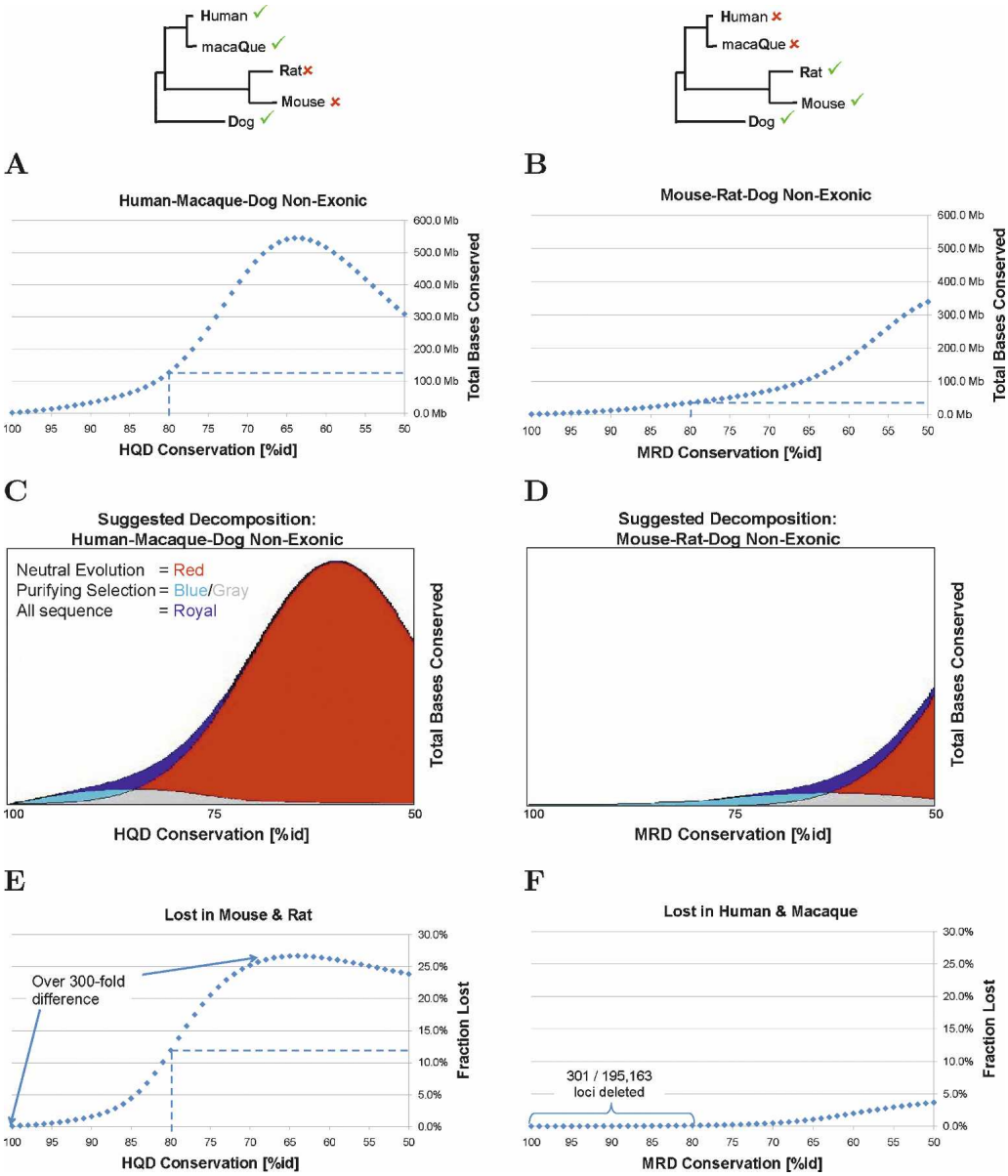

Figure 1. Indispensability of mammalian DNA under purifying selection. (*A,B*) Abundance of primate-dog, rodent-dog nonexonic DNA, respectively. (*C,D*) Suggested decomposition of above curves. Color scheme after Waterston et al. (2002). (*E*) Fraction of primate-dog nonexonic DNA lost in rodents, where ultraconserved-like elements are over 300-fold more indispensable than neutral DNA. (*F*) Fraction of rodent-dog nonexonic DNA lost in primates.

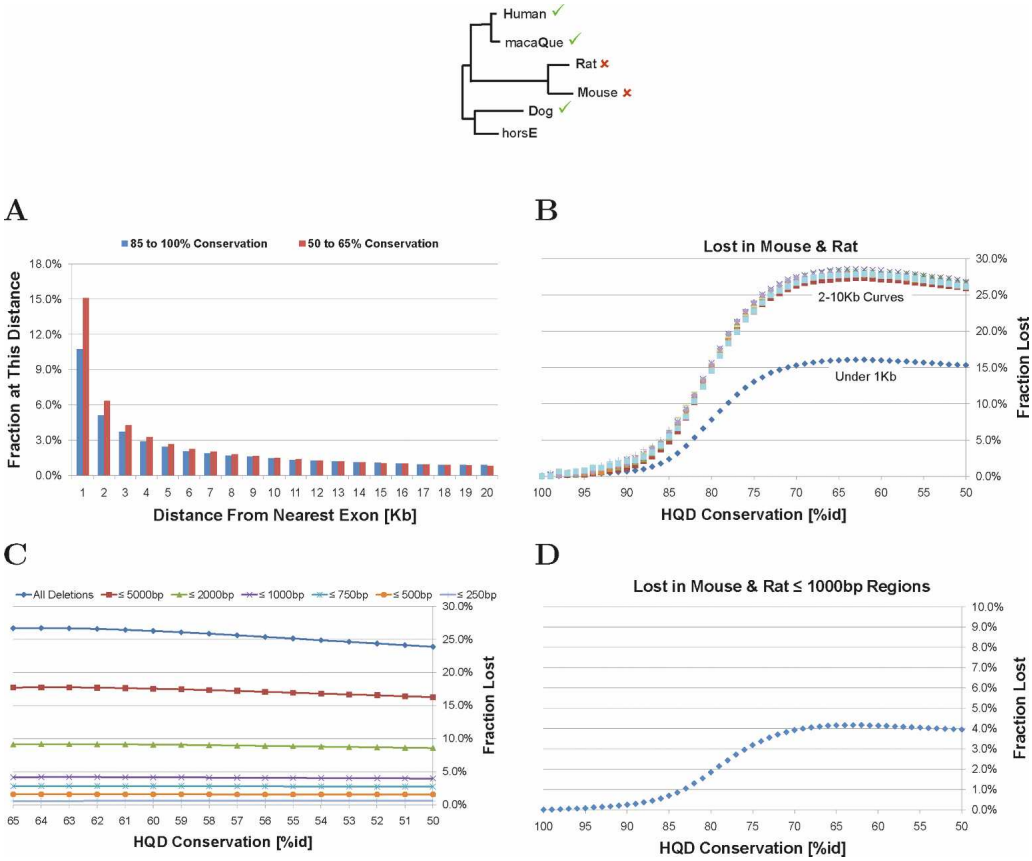

Figure 2. Relationship of neutrally-evolving mammalian DNA loss rates to mammalian genome structure. (*A*) Fraction of conserved nonexonic primate-dog elements as a function of distance from nearest exon. (*B*) Rodent-specific loss rate of nonexonic elements as a function of distance from the nearest exon. Elements within 1 kb of the nearest exon are markedly less likely to be lost at all conservation levels. (*C*) Rodent-specific loss rate of low-conservation nonexonic regions as a function of lost region size. (*D*) Rodent-specific loss rate of nonexonic regions at most 1000 bp in size.

80%id, 0.00276 at 50%id), though the latter seems even more extreme (Fig. 4B), probably, as above, because an exon loss will often lead to gene function loss.

Deep sequence conservation and dispensability

By searching each eutherian element in our abundance curves (Fig. 1A,B) in opossum, platypus, chicken, frog, and teleost fish, we trace the most ancient species divergence within the Euteleostomi (bony vertebrate) clade at which sequence homology of the element is retained (Miller et al. 2007). By tracing the depth of sequence conservation in the loss curves a clear dispensability ranking emerges (Fig. 5). Namely, the more deeply in the vertebrate tree a nonexonic element possesses sequence conservation, the less likely it is to be lost in either primates or rodents, at all levels of high conservation. Thus, depth of sequence conservation provides a stronger measure of indispensability than conservation %id alone. This finding reinforces the utility in sequencing additional nonmammalian species (Editorial 2007).

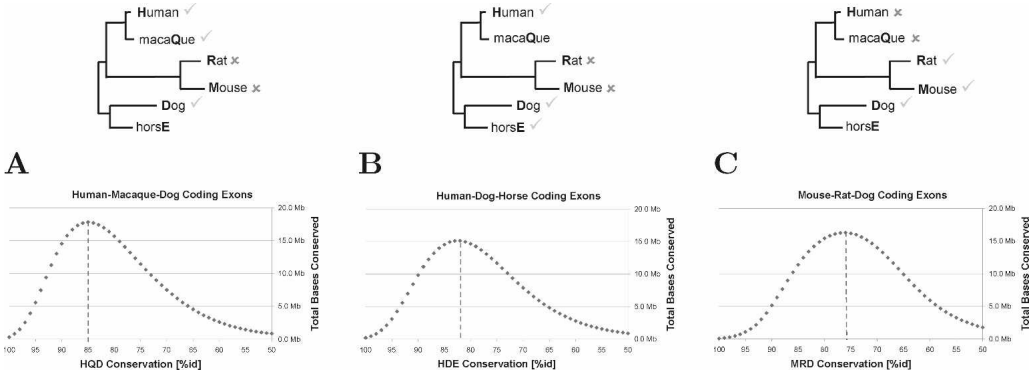

Figure 3. Abundance of coding exon sequences for all three alignment topologies. (*A*) Abundance of primate-dog coding exon DNA. (*B*) Abundance of human-dog-horse coding exon DNA. (*C*) Abundance of rodent-dog coding exon DNA.

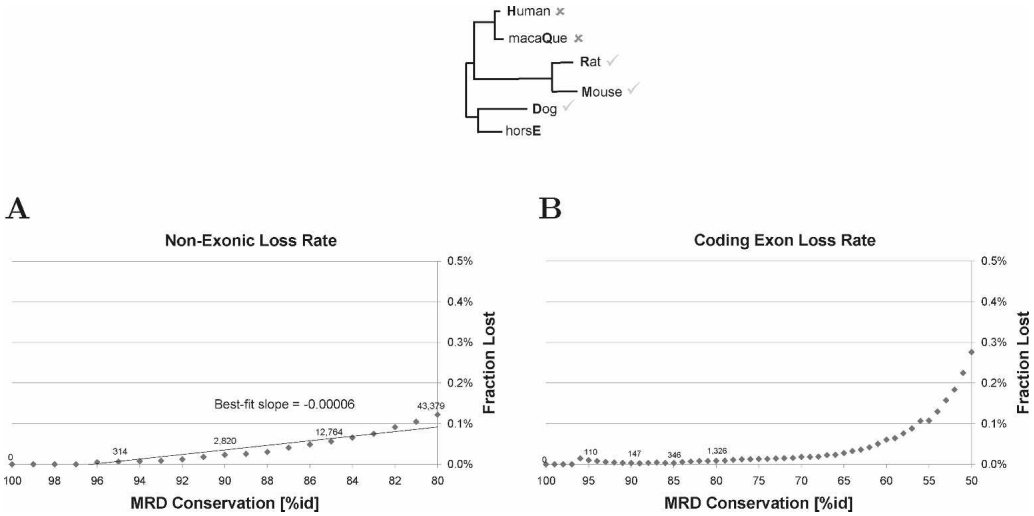

Figure 4. Functional DNA loss rates. (*A*) Fraction of rodent-dog nonexonic DNA lost in primates. (*B*) Fraction of rodent-dog coding exon DNA lost in primates. Total base pairs lost at 100%id, 95%id, 90%id, 85%id, and 80%id displayed above corresponding loss rate points. Note the scale change on the *x*-axis.

Interestingly, the rodent-dog conserved regions, which at 83%id–100%id are presumably all functional, are dominated by regions aligning to chicken (amniote node, Supplemental Fig. S6). The primate-dog conserved regions are qualitatively similar for conservation thresholds of 90%id–100%id. In both cases, the less-conserved sequence is predominantly conserved only to dog, further supporting the notion that the bulk of moderately conserved sequences is neutrally evolving (Waterston et al. 2002, Prabhakar et al. 2006).

Repeating the analysis adding the lizard genome draft (anoCar1) at the amniote node yields qualitatively similar results (data not shown).

Functional annotations of rodent losses

By assigning coding exons to the genes in which they reside, and highly conserved nonexonic elements to the nearest gene locus they may regulate, we calculate gene ontology annotation enrichments for rodent-specific losses (Ashburner et al. 2000). Enrichment analysis of coding losses do not yield statistically significant associations (Supplemental Table S1), while the analysis of nonexonic losses at different conservation thresholds suggest mild enrichments for genes involved in DNA binding, transcription regulation, and various aspects of development (Supplemental Table S2, S3). The larger set of all eutherian conserved nonexonic elements is known to preferentially congregate near genes having these same annotations (Lindblad-Toh et al. 2005). Interestingly, rodent deletions show further preference for these functional categories even against the background of all such eutherian elements. Similar enrichment tests applied to primate losses yield no significant associations for either coding or functional nonexonic sequence (data not shown).

Robustness of loss event predictions

Mutational hotspots and positive selection can cause orthologous regions to be mislabeled as lost sequence. It is currently hard to estimate the magnitude of this phenomenon in nonexonic

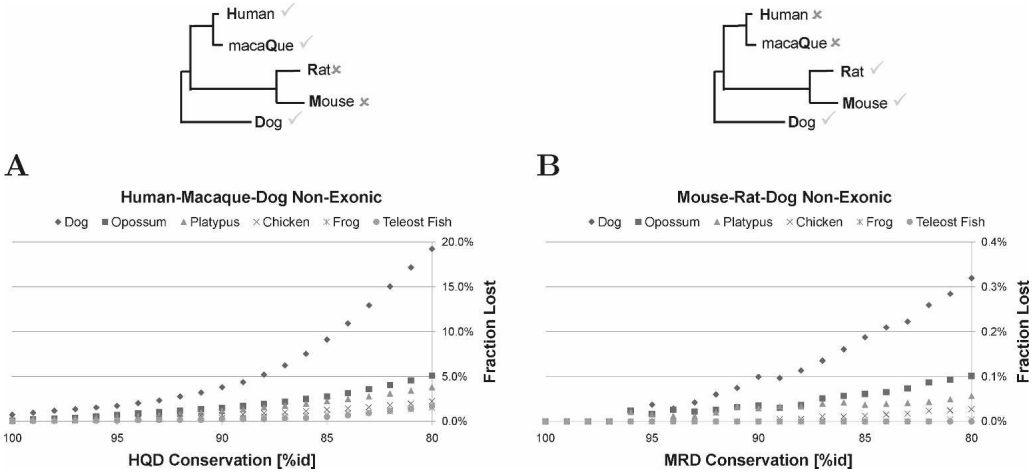

Figure 5. Deep sequence conservation and indispensability of mammalian DNA. (*A*) Fraction of primate-dog nonexonic DNA lost in rodents decomposed by conservation depth. (*B*) Fraction of rodent-dog nonexonic DNA lost in primates decomposed by conservation depth. Note the scale change on the *y*-axis.

sequences, as functional nonexonic elements are identified largely via comparative methods. Coding sequences, the largest well-annotated set of functional elements in the genome, are much better characterized. Consequently, we examine mislabeled coding exon losses to estimate the false positive rate.

Rodent exons that preserve gene structure but are unalignable to their human ortholog cause mislabeled coding exon losses (e.g., Supplemental Fig. S7). For each element labeled as a coding exon loss we examine the orthologous genomic region in both mouse and rat for evidence of an exon annotated by the UCSC Genome Browser knownGene table of the species. Only 4.8% (27/567) of rodent-specific coding exon losses possess an unalignable rodent exon that preserves gene structure. Interestingly, nearly half (12/25) of the genes harboring these exons have been associated with recent positive selection in human or rodents (Supplemental Table S4).

Discussion

Recent lab experiments in mice (Nobrega et al. 2004; Ahituv et al. 2007; Werner et al. 2007; Cretekos et al. 2008) and early theoretical predictions (Zuckerkandl 1992) suggest great redundancy and flux among regulatory genomic elements. We examined dispensability on an evolutionary timescale to assess the importance of regulatory elements in a manner much more sensitive to small fitness effects not measurable in the lab. Sequences perfectly conserved in multiple species show a strong selective pressure against being lost in rodents. We conclude that in all likelihood the four ultraconserved element knockouts (Ahituv et al. 2007) also suffer an evolutionarily noticeable fitness loss that either is too small to be detected (e.g., 1% change), or does not manifest itself under lab conditions (where air is filtered, predators are absent, and water, food, and mate are served).

The regions immediately flanking exons are enriched for poorly-conserved sequence. Though *cis*-regulatory modules have been shown to cluster preferentially near transcription start sites (Blanchette et al. 2006), it is possible that distal regulatory elements preferentially occur away from exons to provide modularity and flexibility for regulatory elements and exons to evolve under their different evolutionary pressures. Alternatively, the presence of nearby strongly-conserved exons may encourage a disproportionate fraction of poorly-conserved sequence to align (Pollard et al. 2004).

The failure to identify a measurable phenotype in the ultraconserved knockout mice (Ahituv et al. 2007) is unexpected, given that the deleted ultraconserved elements are verified enhancer elements near ancient genes (*DMRT1*, *DMRT2*, *DMRT3*, *PAX6*, *ARX*, *SOX3*, all conserved to teleost fish) that produce marked phenotypes when inactivated or expression levels are significantly altered (Ahituv et al. 2007). However, these results do have an experimental analog in coding sequences. Our limited ability to assay for phenotypic changes in the lab can allow many deleterious effects to go unreported. A significant percentage (6.7%) of publicly documented mouse gene knockouts fail to produce measurable phenotypes, including knockouts of very ancient genes (Supplemental Methods S2). These numbers also constitute an (potentially large) underestimate, as many journal editors hesitate to publish lack of phenotype results. Estimates of the fraction of genes in which knockout experiments fail to identify a measurable phenotype run as high as 20% of all genes (Barbaric et al. 2007). Complete functional redundancy may be quite rare, however, as a study in yeast recently showed a quantifiable mutant phenotype in some environment for nearly all genes that produce no phenotypic consequences in rich medium (Hillenmeyer et al. 2008). Similarly, multiple regulatory elements with an experimentally-verified overlap in functionality (Göttgens et al. 2004) or overlapping expression domains during a particular developmental stage (Uemura et al. 2005; Bejerano et al. 2006) suggest some functional redundancy exists among *cis*-regulatory elements, but does not imply that these elements are completely redundant. The demonstrated resistance of nonexonic sequences under purifying selection to deletion on an evolutionary timescale suggests in fact that complete functional redundancy of *cis*-regulatory elements is rare. In this context of partial redundancy, small fitness losses undetectable by conventional lab assays may be expected.

An initial functional exploration of both more and less highly conserved *cis*-regulatory elements under purifying selection shows that many drive expression in similar precise patterns, irrespective of exact conservation level (Visel et al. 2008). Our evolutionary analysis corroborates the notion of importance of *cis*-regulatory elements under purifying selection, regardless of conservation %id. By looking beyond rodent losses of primate-dog sequence to three mammalian configurations of differing total branch lengths (Fig. 3; Supplemental Fig. S5), and by introducing a cumulative loss rate plot overlaying all three configurations (Supplemental Fig. S8), we observe a correlation shared between rodent and primate losses that strongly suggests the mixing of loci under purifying selection with neutrally evolving ones is mostly responsible for higher loss rates as the conservation %id is lowered. This observation suggests that deletion of the majority of DNA elements under purifying selection leads to an evolutionary noticeable fitness loss. If so, most such deletions, at a wide range of fitness effects, will be swept out of the population over time (Garcia-Dorado et al. 2003). Indeed, our evolutionary analysis shows that when mammalian DNA under purifying selection is isolated using sufficient branch lengths, most loci in this set (of many thousands) are nearly equally indispensable over millions of years of evolution.

In 1977, Allan Wilson and colleagues (Wilson et al. 1977) proposed that the rate of protein evolution depends on a combination of the probability that a substitution will be compatible with protein function and on the dispensability of the protein to the organism. A rich body of protein literature in several organisms has since validated Wilson's claim by showing that loss rate correlates with essentiality more than substitution rate does (Pal et al. 2006) (irrespective of sophistication level used to quantify substitution rate). Our evolutionary analysis suggests a similar scenario in *cis*-regulatory elements under purifying selection. In further analogy to proteins, the dispensability of *cis*-regulatory elements under purifying selection increases minimally as conservation across species is lowered. Nearly all *cis*-regulatory elements under purifying selection that make any meaningful contribution toward organism fitness (large or small) are indispensable, regardless of exact inter-species conservation level. No primate losses occur in nonexonic sequences conserved at 97%id or higher in rodents and dog, which precludes analysis of the exact relationship between conservation %id and dispensability of these sequences. However, the relationship observed in the other species configurations (Supplemental Fig. S5) indicate this is probably a consequence of the small loss rate of primates in general, rather than a unique property of perfectly conserved sequences.

In this study, we rely on sequence similarity methods to identify functional homologs across species. Sequences that possess redundant functionality but lack sequence homology may not be quantified using these methods. In particular, the ability of vertebrate *cis*-regulatory sequences to shuffle individual binding sites (Sanges et al. 2006) and our inability to align many orthologous noncoding RNAs (Torarinsson et al. 2006) may be contributors to false positive losses of nonexonic sequences, and thus may partially account for differences in the loss rates of nonexonic and coding sequences. Since this phenomenon would increase our estimates of the loss rates of nonexonic elements, the small loss rates of nonexonic sequences we report may even be slightly conservative.

By including sequence conservation depth information in the analysis of mammalian *cis*-regulatory elements under purifying selection, we achieve additional predictive power of the dispensability of an element. Regulatory elements that possess sequence homology in distant species are less dispensable than those that are only found in close species, at all levels of conservation %id. While low coverage genomes are useful for extending branch-lengths of individual elements, the forthcoming availability of over a dozen high coverage mammalian genomes (Green 2007) will allow us to deploy the presence/absence measure genome-wide to more accurately highlight indispensable genomic regions evolving at a range of rates.

Our work serves to provide the evolutionary record of mammalian ultraconserved and highly conserved element indispensability. Additional work will be needed to better understand the extremes at which some genomic loci are conserved, as well as the manifested irreducibility of the noncoding portions of our genome.

Methods

Identifying rodent-specific losses of mammalian conserved DNA

The procedure we next describe is summarized in Supplemental Figure S1. Five UCSC genome assemblies were used throughout this study: human (hg18), macaque (rheMac2), mouse (mm8), rat (rn4), and dog (canFam2). The phylogenetic relationships and evolutionary distances between the species are shown in Supplemental Figure S2.

To identify primate-dog conserved DNA, we created a multiple alignment of human, macaque, and dog using MULTIZ (Blanchette et al. 2004) on syntenic human-macaque and human-dog pairwise alignments generated by running the BLASTZ (Schwartz et al. 2003) and UCSC chaining and netting tools (Kent et al. 2003). Each 100 bp window within the multiple alignment was annotated for conservation %id (the number of bases perfectly conserved between the three species), recorded in human coordinates. Overlapping windows of the same conservation %id were merged into single genomic intervals. Alignment reliability concerns limited the set to three-way alignments of 50%id or more. Roughly 46.0% (1327 Mb) of the human genome was categorized as alignable at this threshold.

The pairwise alignment files from human to mouse and from human to rat identified conserved elements potentially lost in the rodent lineage. To guarantee loss of orthologous rodent DNA, we required losses only from pairwise alignments that were one-to-one mappings, based on the UCSC chains and nets (Kent et al. 2003), leaving 1108 Mb of human, macaque, and dog conserved sequence as candidates for rodent loss. We identified rodent losses as the intersection of gaps from human to mouse and from human to rat within the conserved regions defined above. Only elements of 50 bp or larger (corresponding to indel mediated alignment windows of 100 bp at 50%id) were retained.

Rodent sequence assembly gaps have similar signatures to true rodent-specific losses, as they appear in pairwise alignments to human as stretches of human sequence that have no rodent orthologs. We removed all rodent assembly gaps from further consideration.

We ran BLAT (Kent 2002) on all remaining candidate losses to discover and remove false positives arising from alignment artifacts and structural RNA migration since the rodent–primate split. To further ensure that the elements were truly missing from the rodent lineage, we ran the more sensitive BLASTN (Pearson et al. 1997) of the human sequence within candidate rodent-specific losses against a database of rodents that included mouse NCBI Build 36.1, the Celera Genomics whole-genome mouse shotgun assembly, the mouse MGSC v3 assembly, and the rat RGSC v3.4 assembly. We removed elements with significant similarity (E-score $\leq 10^{-10}$) found in any of the four databases scanned. The stringent BLAST significance threshold identifies a very limited amount of additional sequence (0.25%) beyond that discovered by the conservative BLAT screening. Further relaxing the significance level used to exclude sequences runs the risk of preferentially excluding sequences that have diverged paralogs elsewhere in the genome. The resultant elements were classified as lost in rodents, and comprised 9.04% (261 Mb) of the human genome.

We annotated portions of both the primate-dog conserved DNA and the subset lost in rodents as nonexonic. Nonexonic elements at a particular conservation %id c were identified by taking the union of all 100 bp windows of conservation %id c that do not overlap any portion of the reference genome annotated as coding exons or UTRs. Other regions of primate-dog conservation and rodent loss were labeled as coding exon elements. Coding exon elements at a particular conservation %id c consist of the union of all 100 bp windows of conservation %id c that overlap at least 50 bp of a coding exon. All gene annotations used to identify both coding exon and nonexonic elements were based upon the UCSC Genome Browser knownGene table (Hsu et al. 2006) of the human genome assembly.

The loss rate at a particular conservation %id c was calculated as the total number of base pairs in rodent-specific loss elements of conservation %id c divided by the total number of base pairs in primate-dog conserved elements of conservation %id c. Nonexonic and coding exon loss rates were calculated using only elements in the appropriate classes in both the numerator and denominator of the equation.

Repeating our entire analysis using both smaller and larger window sizes of 75 bp and 150 bp yields very similar results (Supplemental Fig. S9).

Losses within the rodent lineage of elements conserved between human, dog, and horse were processed in an identical manner to the primate-dog elements, using the UCSC horse (equCab1) genome assembly in place of macaque.

Identifying primate-specific losses of mammalian conserved DNA

Losses within the primate lineage of elements conserved between mouse, rat, and dog (Supplemental Fig. S2) were processed in a manner analogous to rodent-specific losses. Elements were identified in mouse coordinates. The BLASTN step was run against a database of primates that included the human NCBI Build 36.1, the Celera Genomics whole-genome human shotgun assembly, and the chimpanzee CSAC Build 2 v1.

Deep sequence conservation and dispensability

Pairwise syntenic genome alignments of human sequence conserved in macaque and dog with the following UCSC genomes were generated: opossum (monDom4), platypus (ornAna1), chicken (galGal3), frog (xenTro2), zebrafish (danRer4), tetraodon (tetNig1), fugu (fr2), and stickleback (gasAcu1).

Each nonexonic 100 bp window within a primate-dog conserved region was tested for overlap of 50 bp or more with each pairwise alignment. Window conservation depth was assigned as the most distant species/clade for which the overlap criteria was fulfilled; the distance metric has teleost fish most distant (bony vertebrate ancestry), followed by frog (tetrapod), chicken (amniote), platypus (base of the mammalian lineage), opossum (therian–marsupial and placental–mammals), and dog (placental mammal specific). Windows satisfying the overlap criteria for one or more of the pairwise alignments to zebrafish, tetraodon, fugu, or stickleback were labeled with teleost fish conservation depth. Windows that did not satisfy the overlap criteria for any of the pairwise alignments were labeled with dog (eutherian) conservation depth. All 100 bp windows of the same conservation %id and conservation depth were merged into nonoverlapping elements, and the total abundance of sequence plotted (Supplemental Fig. S6A).

The conservation depth of rodent-dog conserved regions (Supplemental Fig. S6b) was determined using pairwise syntenic alignments from mouse to the same species mentioned above, using the same overlap criteria.

Loss rates (Fig. 5) were calculated by dividing the number of bases lost of a particular conservation depth and conservation %id by the total number of bases of that conservation depth and conservation %id.

Statistical enrichment analyses

Enrichment analyses are based on the hypergeometric test as described in Lowe et al. (2007), and are performed against the Gene Ontology (GO) database (Ashburner et al. 2000). Briefly, for each coding exon element conserved between human, macaque, and dog, if the gene whose exon is conserved has any informative GO annotations, the gene is included in the background set. Each of those coding exon elements that is additionally lost in mouse and rat has its gene included in the foreground set. Independent tests are performed for each GO annotation π, and it is asked whether the number of genes in the foreground set with annotation π, k_π, picked from all foreground genes, n, is significantly enriched for π when compared to the total number of genes in the background set with annotation π, K_π, picked from the background set, N.

The nonexonic enrichment analyses are performed in a similar manner. Each nonexonic element is mapped to the GO-annotated gene whose transcription start site (TSS) is nearest the element. If no GO-annotated gene has its TSS within 1 Mb of the element, the element is ignored. The sets of genes chosen are used in a similar hypergeometric test.

Acknowledgments

We thank David Haussler, Jim Kent, Hiram Clawson, Mark Diekhans, and Craig Lowe for technical help; Dmitri Petrov and members of the Bejerano lab for insightful manuscript comments; David Kingsley and Craig Lowe for valuable discussions; and Nadav Ahituv and Ivan Ovcharenko for sharing results prepublication. C.M. is supported by a Stanford Bio-X graduate fellowship. G.B. is a Sloan research fellow and a Searle scholar. Research partially supported by an Edward Mallinckrodt, Jr. Foundation junior faculty grant.

References

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennachio, L., and Rubin, E. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5:** e234. doi: 10.1371/journal.pbio.0050234.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Barbaric, I., Miller, G., and Dear, T. 2007. Appearances can be deceiving: Phenotypes of knockout mice. *Brief. Funct. Genomic. Proteomic.* **6:** 91–103.

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

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J., and Haussler, D. 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441:** 87–90.

Blanchette, M., Kent, W., Riemer, C., Elnitski, L., Smit, A., Roskin, K., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Blanchette, M., Bataille, A., Chen, X., Poitras, C., Laganire, J., Lefbvre, C., Deblois, G., Gigure, V., Ferretti, V., Bergeron, D., et al. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.* **16:** 656–668.

Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., and Sidow, A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14:** 539–548.

Cretekos, C., Wang, Y., Green, E., Martin, J., Rasweiler, J., and Behringer, R. 2008. Regulatory divergence modifies limb length between mammals. *Genes & Dev.* **22:** 141–151.

Editorial. 2007. Life as we know it. *Nature* **449:** 1.

Gaffney, D.J. and Keightley, P.D. 2006. Genomic selective constraints in murid noncoding DNA. *PLoS Genet.* **2:** e204. doi: 10.1371/journal.pgen.0020204.

Garcia-Dorado, A., Caballero, A., and Crow, J.F. 2003. On the persistence and pervasiveness of a new mutation. *Evolution Int. J. Org. Evolution* **57:** 2644–2646.

Göttgens, B., Broccardo, C., Sanchez, M., Deveaux, S., Murphy, G., Gthert, J., Kotsopoulou, E., Kinston, S., Delaney, L., Piltz, S., et al. 2004. The *scl* +18/19 stem cell enhancer is not required for hematopoiesis: Identification of a 5′ bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. *Mol. Cell. Biol.* **24:** 1870–1883.

Green, P. 2007. 2× genomes—Does depth matter? *Genome Res.* **17:** 1547–1549.

Hillenmeyer, M., Fung, E., Wildenhain, J., Pierce, S., Hoon, S., Lee, W., Proctor, M., St. Onge, R., Tyers, M., Koller, D., et al. 2008. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science* **320:** 362–365.

Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. 2006. The UCSC known genes. *Bioinformatics* **22:** 1036–1046.

Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama, S.R., and Haussler, D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317:** 915.

Kent, W.J. 2002. BLAT—The BLAST-like alignment tool. *Genome Res.* **12:** 656–664.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.

Lindblad-Toh, K., Wade, C., Mikkelsen, T., Karlsson, E., Jaffe, D., Kamal, M., Clamp, M., Chang, J., Kulbokas, E., Zody, M., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.

Lowe, C.B., Bejerano, G., and Haussler, D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci.* **104:** 8005–8010.

Margulies, E.H., Maduro, V.V.B., Thomas, P.J., Tomkins, J.P., Amemiya, C.T., Luo, M., and Green, E.D. 2005. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc. Natl. Acad. Sci.* **102:** 3354–3359.

Miller, W., Rosenbloom, K., Hardison, R., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D., Baertsch, R., Blankenberg, D., et al. 2007. 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* **17:** 1797–1808.

Nobrega, M.A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E.M. 2004. Megabase deletions of gene deserts result in viable mice. *Nature* **431:** 988–993.

Pal, C., Papp, B., and Lercher, M. 2006. An integrated view of protein evolution. *Nat. Rev. Genet.* **7:** 337–348.

Pearson, W.R., Wood, T., Zhang, Z., and Miller, W. 1997. Comparison of DNA sequences with protein sequences. *Genomics* **46:** 24–36.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.

Petrov, D. 2002. Mutational equilibrium model of genome size evolution. *Theor. Popul. Biol.* **61:** 531–544.

Pollard, D., Bergman, C., Stoye, J., Celniker, S., and Eisen, M. 2004. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5:** 6. doi: 10.1186/1471-2105-5-6.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O., and Pennacchio, L.A. 2006. Close sequence comparisons are sufficient to identify human *cis*-regulatory elements. *Genome Res.* **16:** 855–863.

Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F., and Stupka, E. 2006. Shuffling of *cis*-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.* **7:** R56. doi: 10.1186/gb-2006-7-7-r56.

Schwartz, S., Kent, W., Smit, A., Zhang, Z., Baertsch, R., Hardison, R., Haussler, D., and Miller, W. 2003. Human–mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.

Stone, E., Cooper, G., and Sidow, A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6:** 143–164.

Torarinsson, E., Sawera, M., Havgaard, J., Fredholm, M., and Gorodkin, J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* **16:** 885–889.

Uemura, O., Okada, Y., Ando, H., Guedj, M., Higashijima, S.-I., Shimazaki, T., Chino, N., Okano, H., and Okamoto, H. 2005. Comparative functional genomics revealed conservation and diversification of three enhancers of the *isl1* gene for motor and sensory neuron-specific expression. *Dev. Biol.* **278:** 587–606.

Visel, A., Prabhakar, S., Akiyama, J., Shoukry, M., Lewis, K., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E., Pennacchio, L., et al. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* **40:** 158–160.

Waterston, R., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Werner, T., Hammer, A., Wahlbuhl, M., Bosl, M.R., and Wegner, M. 2007. Multiple conserved regulatory elements with overlapping functions determine Sox10 expression in mouse embryogenesis. *Nucleic Acids Res.* **35:** 6526–6538.

Wilson, A., Carlson, S., and White, T. 1977. Biochemical evolution. *Annu. Rev. Biochem.* **46:** 573–639.

Zuckerkandl, E. 1992. Revisiting junk DNA. *J. Mol. Evol.* **34:** 259–271.

Received April 28, 2008; accepted in revised form August 13, 2008.