No Gene in the Genome Makes Sense Except in the Light of Evolution

Wilfried Haerty and Chris P. Ponting

MRC Functional Genomics Unit, Department of Physiology, Anatomy, and Genetics, University of Oxford, Oxford OX1 3PT, United Kingdom;
email: wilfried.haerty@dpag.ox.ac.uk, chris.ponting@dpag.ox.ac.uk

Annu. Rev. Genomics Hum. Genet. 2014. 15:71–92

First published online as a Review in Advance on April 24, 2014

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi:
10.1146/annurev-genom-090413-025621

Keywords

selection, neutral evolution, noncoding, regulatory element, molecular function

Abstract

Evolutionary conservation has been an accurate predictor of functional elements across the first decade of metazoan genomics. More recently, there has been a move to define functional elements instead from biochemical annotations. Evolutionary methods are, however, more comprehensive than biochemical approaches can be and can assess quantitatively, especially for subtle effects, how biologically important—how injurious after mutation—different types of elements are. Evolutionary methods are thus critical for understanding the large fraction (up to 10%) of the human genome that does not encode proteins and yet might convey function. These methods can also capture the ephemeral nature of much noncoding functional sequence, with large numbers of functional elements having been gained and lost rapidly along each mammalian lineage. Here, we review how different strengths of purifying selection have impacted on protein-coding and non-protein-coding loci and on transcription factor binding sites in mammalian and fruit fly genomes.

1. INTRODUCTION

> Nothing makes sense in biology except in the light of evolution.
>
> —Theodosius Dobzhansky (25)

Functional regions—coding exons, regulatory elements, and others—are sparsely distributed in mammalian genomes, scattered widely in a sea of apparently inert sequence. We understand much about the functions of protein-coding genes but little about the molecular mechanisms, locations, and properties of mammalian non-protein-coding functional elements. The approach adopted by the Encyclopedia of DNA Elements (ENCODE) project to shed light on this genomic "dark matter" detected and categorized regions that participate in one or more biochemical processes (32). However, this approach could not be comprehensive because it investigated only a limited subset of cell types and could not consider all developmental stages. Moreover, it could not effectively distinguish between functional sites and inconsequential sites, such as those involved in low-occupancy transcription factor interactions (39), and it conflated epiphenomena (such as random transcription events) with primary causes (biological function). Current experimental approaches thus are unable to predict comprehensively or accurately how important particular regulatory elements are to organismal biology. Experimental targeting of mutations in vivo can reveal deleterious effects, but formally it is still necessary to demonstrate that these have effects in natural settings, something that current protocols rarely cover.

An alternative approach to predict the biological importance and locations of noncoding functional elements is to recognize the telltale signatures written over time into genomes by the sieving of mutations by natural selection. This approach has the advantages of being comprehensive (because the scrutiny of selection extends to any element that is functional in any cell type and at any developmental stage) and inexpensive (because of the availability of genome sequences and of cheap sequencing technologies). Selection also will not have regarded as functional any sequence that, despite being bound by a factor or being transcribed, remains of no biological consequence. Finally, unlike biochemical assays, evolutionary analyses can predict an element's biological importance by estimating the strength of selection that acted on its sequence. This approach can indicate the extent to which fitness has been affected by ancient mutations that landed in single elements, and more generally in classes of functional elements, and thus can rank elements by their proposed biological importance.

Here, we review the relative contributions to human biology of different classes of functional elements inferred from comparisons of recently sequenced mammalian and human genomes. Owing to space limitations, we mostly restrict the discussion to classes of functional elements rather than individual examples, and we focus on conservation and constraint as opposed to positive selection and adaptation, which have been reviewed elsewhere (41). As a consequence, the review focuses on the approximately 10% of the human genome that is under significant selective constraints and not on the vast majority that evolves neutrally (101, 103). The review is structured around three counterpoints (**Figure 1**): (*a*) selection that acted on either non-protein-coding or protein-coding sequence; (*b*) selection that acted on either the genomes of humans (or other mammals) or those of other, more distantly related metazoans, such as *Drosophila* (fruit flies); and (*c*) selection that was either ancient or more contemporary.

2. MUTATION AND SELECTION, CONSERVATION AND CONSTRAINT

Imagine if we could train a time-lapse camera on a single nucleotide position in your genome and, by winding back time, watch how it changed by chance mutations as it was passed back through the generations (and along the germline) over hundreds of millions of years. If this nucleotide

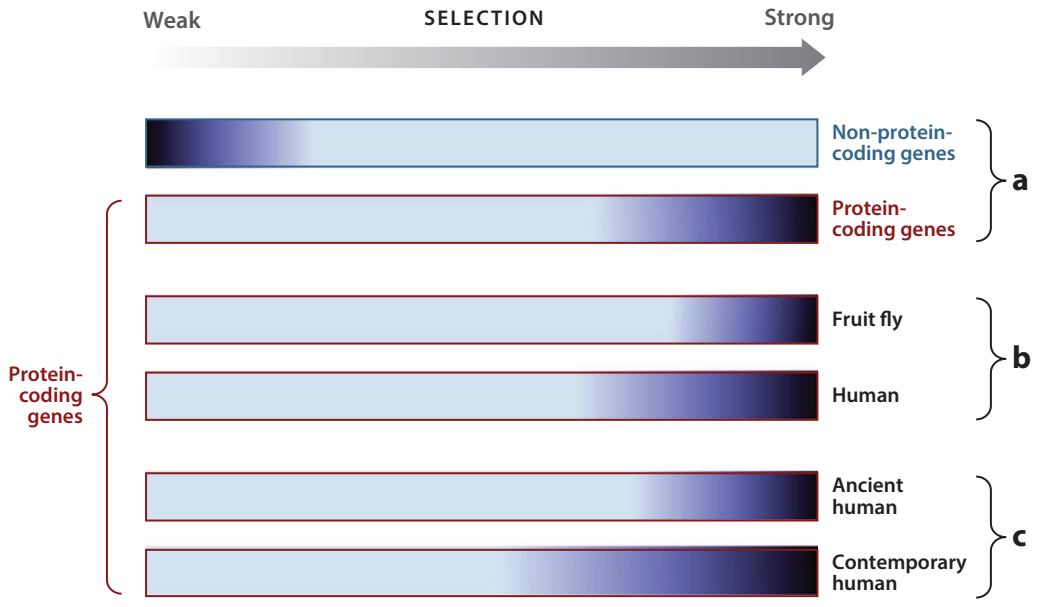

Figure 1

Schematic illustration of the differing extents to which selection has acted on exonic sequences of genes. These genes have been grouped according to the threefold organizational principle of the review: Comparisons are made between (*a*) selection acting on non-protein-coding versus protein-coding genes; (*b*) selection acting on protein-coding genes in species such as *Drosophila* (fruit flies) versus those in humans; and (*c*) selection acting on human protein-coding genes during ancient versus more contemporary evolution. On average, selection has been weakest on transcribed non-protein-coding sequence and strongest on protein-coding sequence from species (such as fruit flies) that have high effective population sizes.

was functional, we would observe it to have changed only very rarely. This is because change would be mostly deleterious and thus would be negatively selected and less likely to have been propagated to subsequent generations. For example, three-fourths of the coding bases of human histone H4, a protein of crucial importance in packaging DNA, have remained unaltered since we last shared an ancestor with plants. However, if the scrutinized nucleotide was not functional, changes would not have been selected against and thus would have occurred more frequently. At the end of these very long time periods, a neutrally evolving sequence would have experienced so many changes that little trace of its ancestral sequence would remain in extant species. After this amount of divergence time, alignment of DNA sequence has often become so inaccurate as to be uninformative.

Comparisons of sequences of extant species whose common ancestors lived in this long-ago time period often reveal short regions that are alignable and well conserved, separated by long stretches that are unalignable and poorly conserved. For example, only 2.5% of nucleotides align in the genomes of chickens and humans, species that last shared a common ancestor a little over 300 million years ago (52). When an accurate alignment can be obtained, it permits an evolutionary change to be assigned, usually unambiguously, to a lineage in a phylogeny. When two such alignments are compared, different numbers of changes assigned to this lineage can indicate differences in these two regions' evolutionary rates. A region whose evolutionary rate significantly exceeds that of another region, chosen specifically because it is believed to have evolved neutrally (**Table 1**), is likely to have experienced episodes of positive selection, perhaps owing to adaptive evolution. A region whose evolutionary rate is significantly lower than that of

Table 1 Genomic elements used as neutrally evolving sequences

Neutral background	Advantages	Disadvantages	References
Fourfold degenerate sites	Interdigitated	Codon usage bias, regulatory elements	22, 33, 71, 74
Small introns	Neutral	Applicable only in *Drosophila* species	96
Ancestral repeats	Neutral	Neutrally assessed only in mouse–human	79
Unannotated nonconserved sites across the genome	Widespread	Background selection, hitchhiking	124
Sampling within flanking sequences	Control for background selection	Not necessarily fully neutral	5, 48

Purifying selection: the process by which deleterious mutations are preferentially removed from the population

the putatively neutrally evolved region has been subject to constraint—in other words, negative (or purifying) selection. Conserved sequence is not necessarily constrained: Human and great ape genomes are highly conserved because of their relatively recent common ancestry, yet most of their sequence is not constrained. Conversely, constrained sequence is not necessarily conserved if its function has arisen only recently and thus is not shared with the other species under consideration. Metazoan genomes differ substantially in their fractions of constrained sequence, depending on their gene content and the size and extent of their regulatory sequence (103) (**Figure 2**).

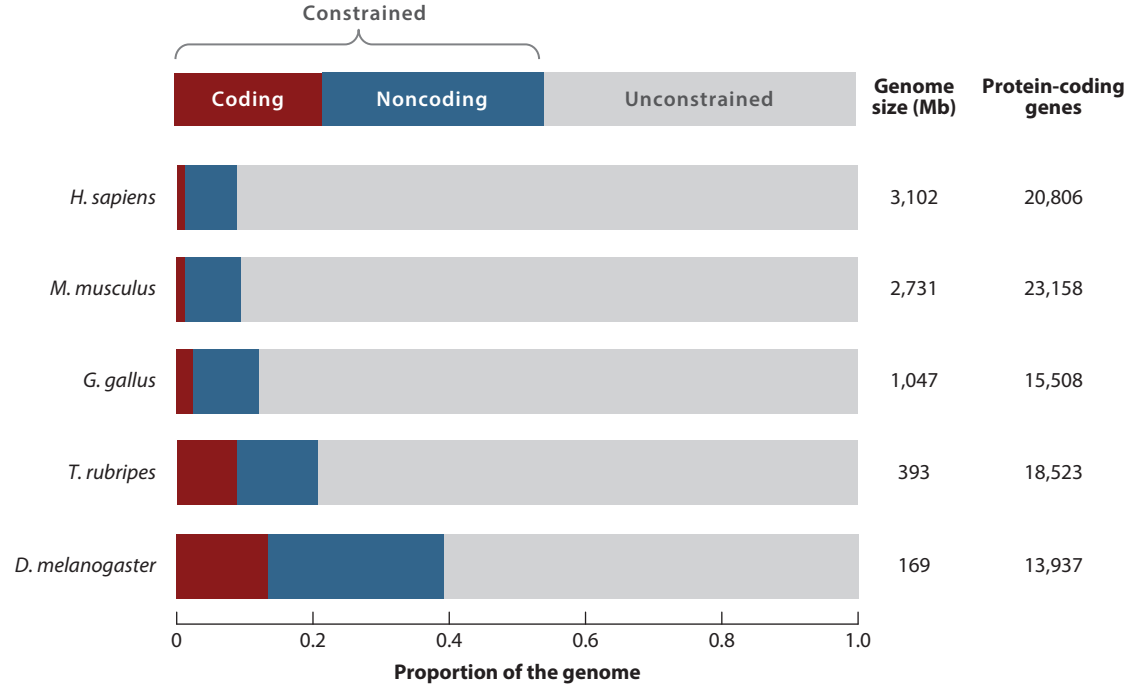

Figure 2

Proportion of the genome identified as being under constraint (purifying selection) when using the neutral indel model (79, 84). Alignments between *Homo sapiens* and *Macaca mulatta*, *Mus musculus* and *Rattus norvegicus*, *Gallus gallus* and *Taeniopygia guttata*, *Takifugu rubripes* and *Tetraodon nigroviridis*, and *Drosophila melanogaster* and *Drosophila simulans* were used to produce the fraction of the genome under constraint. The proportions of sequence under constraint depend on the species pair analyzed and the alignment-quality processing. Data from Meader et al. (84).

Detecting species-specific functional sequence that has only recently become constrained requires comparisons not between species whose lineages have long been separated but rather within a population—for example, that of modern humans. Over the several hundred thousand years since the ancestral population of modern humans arose, mutations have risen and fallen in population frequency: Those that are advantageous tend to rise in frequency because of positive selection, whereas those that are substantially deleterious tend toward extinction because of negative selection. Approaches to identify the imprint of selection thus compare the population frequencies of alleles in the regions of interest against those in putatively neutrally evolved sequence: Higher frequencies might reflect episodes of positive selection, whereas lower frequencies might reflect purifying selection. Importantly, even if a coding sequence region shows a signature of constraint, it is not possible to infer definitively whether this discriminative sieve of selection has acted on DNA or RNA or on protein molecular functions.

d_N: the number of nonsynonymous substitutions per nonsynonymous site

d_S: the number of synonymous substitutions per synonymous site

3. BETWEEN-SPECIES COMPARISONS

3.1. Spectrum of Selection Across Protein-Coding Gene Sequence

Winding back time by approximately 100 million years takes us to the Cretaceous period, just before the separation of lineages that led to either modern primates (including humans) or rodents (such as mice). The protein-coding DNA of humans and mice has been subject mostly to strong purifying selection over this long time period, with approximately 85% of bases remaining the same as they were in their last common ancestor (87, 109). Aligned non-protein-coding DNA, however, has in the meantime changed considerably, with approximately twice as many bases differing between the two species as compared with protein-coding sequence.

The variable impression left across coding sequence from the selective purging of deleterious alleles is evident from a classic image of early vertebrate genomics, which Jim Kent produced by sampling alignments between human and mouse genes (87, figure 25*a*). Instead of using pairwise sequence identity, as he did, we have recalculated the profile of purifying selection across mammalian genes using conservation (phastCons) scores estimated across vertebrate evolution (109) (**Figure 3**). Sequence conservation here directly indicates the relative importance of different nucleotide positions along a generic gene model. Sites in exons, particularly those near exon boundaries (**Figure 3**), tend to be the least changeable without impairing gene function, whereas the bulk of intronic sequence is more accepting of changes because it contains a much lower proportion of functional sequence; mammalian 5′ and 3′ untranslated regions (UTRs) show levels of conservation that are intermediate between these two extremes.

The discriminative impact of selection is most apparent within codons. Owing to the threefold degeneracy of the genetic code, some nucleotide substitutions fail to alter the amino acid and are therefore synonymous. Many such substitutions occur in the third codon position, and there are eight amino acids whose third site—termed a fourfold degenerate (4D) site—is entirely degenerate; substitutions at these sites are always synonymous. Purifying selection thus acts discriminately on these different sites, with stronger selection on the first and second sites than on the third (and thus 4D) site (**Figure 3** insets). Nucleotide substitutions in codons that alter the amino acid (nonsynonymous changes) are usually fixed far less frequently than synonymous changes, and under an assumption—challenged below—that synonymous changes are free from selection, the strength of selection acting on amino acid sites can be inferred. More formally, the ratio of d_N (or K_A, the number of nonsynonymous substitutions per nonsynonymous site) to d_S (or K_S, the number of synonymous substitutions per synonymous site) is expected to be 0 when amino acids are essential, 1 when the codons evolve neutrally, and significantly

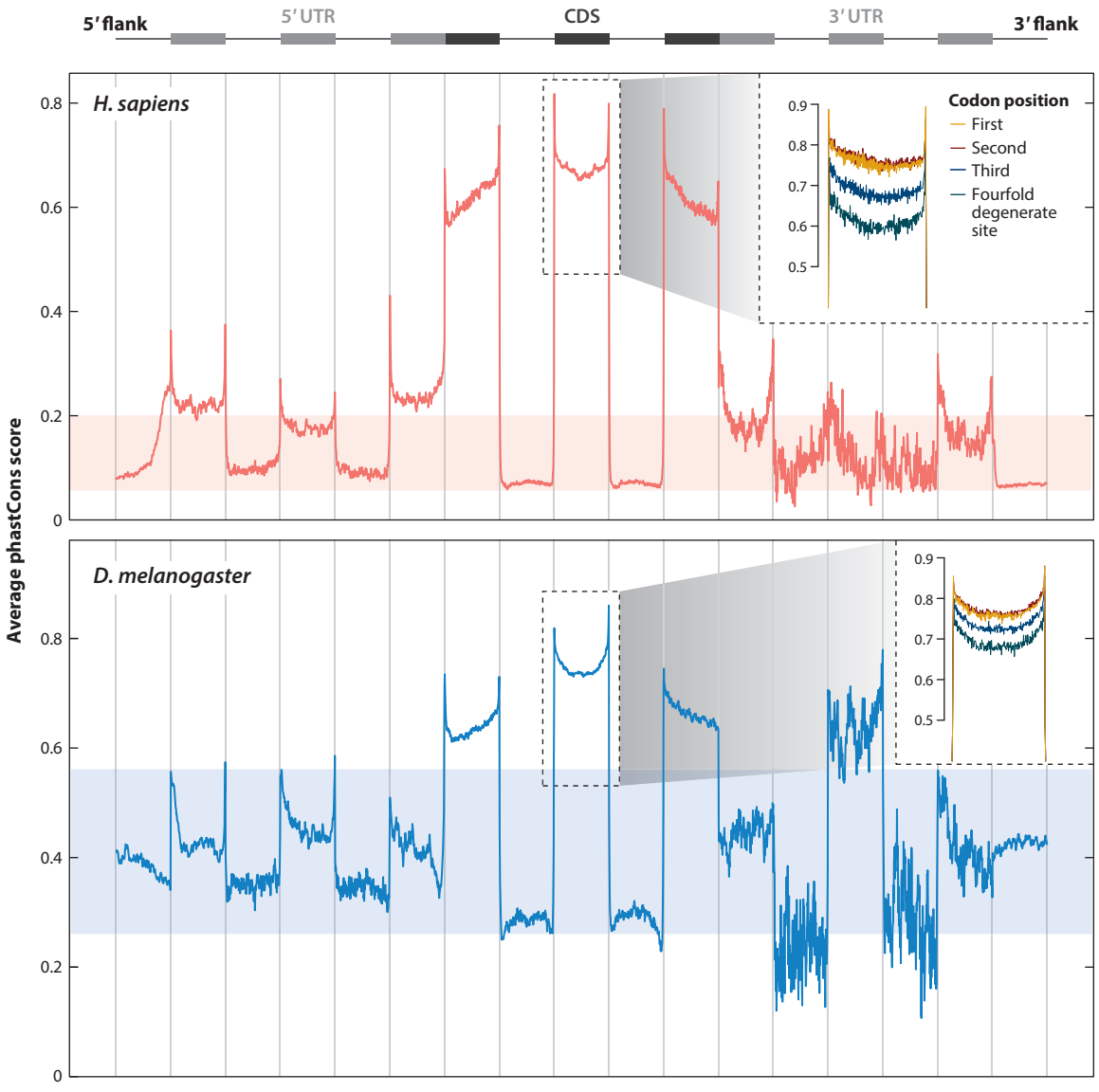

Figure 3

Average nucleotide conservation (phastCons) score sampled across 5′ untranslated region (UTR), coding DNA sequence (CDS), and 3′ UTR exons and introns in *Homo sapiens* and *Drosophila melanogaster* at the first (*orange*), second (*dark red*), and third (*dark blue*) codon positions and fourfold degenerate sites (*dark aqua*). Only exons and introns that were at least 200 nucleotides long and did not overlap other annotated features were used. The high conservation values observed for the intermediate 3′ UTR exons in *D. melanogaster* are likely stochastic noise, a consequence of their low number (20 exons larger than 200 nucleotides). The red and light blue shaded areas represent the range of nucleotide conservation across long noncoding RNA gene models in *H. sapiens* and *D. melanogaster*, respectively.

BACKGROUND SELECTION

The local environment (background) of a variant can affect whether it is purged from a population. For example, neutral variation in a UTR will be reduced because of its linkage with genomically proximal coding sequence that is under negative selection. This is due to neutral changes propagating in a population for an extended period of time only when gametes are free of deleterious alleles (19). Recombination disrupts genetic linkage and thus diminishes background selection.

greater than 1 when positive selection has acted to preferentially fix amino acid changes in the population (56). Selective pressure on synonymous sites, however, is not negligible. This is because for many species there is selection on codon usage and for the preservation of splicing regulatory elements (18, 37). In mammals, approximately 20% of 4D sites are subject to a non-negligible degree of selective constraint (33). Consequently, the use of 4D sites to model neutral evolution in tests for selection can lead to significant biases (**Table 1**). Splicing regulatory elements tend to lie near intron–exon boundaries (within 50 base pairs), and it is their conservation that appears to underlie the peak of conservation scores close to these boundaries (95, 125) (**Figure 3**). The heterogeneity of conservation within and between coding exons (**Figure 3**) indicates that a substantial amount of selection occurs not because of changes in protein function but because of RNA and/or DNA function. This is important because ignoring selection on nucleotide function in coding sequence leads to underestimation of neutral rates, which then leads to higher rates of false predictions of positive selection and more false negatives for purifying selection.

As we have observed, noncoding exonic bases of UTRs are better conserved than noncoding intronic bases (**Figure 3**). This higher degree of conservation is due to its content of various types of functional elements, such as G-quadruplex sequences, internal ribosome entry sites, and upstream open reading frames (7, 33, 54, 94, 109); conservation may also reflect background selection (see sidebar, Background Selection). Many mammalian 5′ UTRs also contain spliced exons whose intronic boundaries, similarly to those in coding sequence, show elevated nucleotide conservation (107) (**Figure 3**), presumably reflecting the high density of splicing regulatory elements lying near exon junctions. The increased sequence conservation of 3′ UTRs over introns is attributable at least in part to purifying selection acting to preserve or avoid microRNA (miRNA) binding sites, also called miRNA response elements (MREs) (20, 112, 129). Mammalian miRNA seed sequences bind via base-pairing with imperfectly complementary MREs within mRNAs, leading to translational repression or target degradation (8). Based on the conservation of at least 45,000 miRNA target sites within human 3′ UTRs, the majority of human genes appear to have been under selective pressure to maintain pairing to miRNAs (40). Mutations that create MREs can also be under negative selection because of their deleterious effect on gene expression regulation (20, 112).

Some investigators consider intronic sequences to have evolved neutrally and thus to provide a useful neutral proxy against which selection on coding sequences can be assessed (**Table 1**). However, in addition to sequences flanking splice acceptor and donor sites and other splicing regulatory motifs that occur within 200 base pairs of exons, introns contain diverse and numerous functional elements that all contribute to lower evolutionary rates (17, 81). Kim & Pritchard (63) reported that approximately 37,000 conserved noncoding elements fall within introns, representing approximately 4.6 Mb, although a more recent estimate of intronic functional elements numbers these at nearly 1.5 million by counting all DNase I–hypersensitive sites (118). Evolutionary rates

MRE: microRNA response element

Neutral proxy: sequence regions that are considered to be neutrally evolving and that are used as the null expectation when testing for selection that acted on sequences of interest

also vary depending on the position of an intron within a gene (45). For example, relative to other introns within a gene, first introns tend to be better conserved, are on average twice as long, and differ in their nucleotide composition (17, 45).

lncRNA: long noncoding RNA

3.2. Spectrum of Selection Across Noncoding RNA Loci

In mammals, thousands of multi-exonic long noncoding RNA (lncRNA) genes have been identified (12, 24, 51) for which cellular functions and molecular mechanisms remain largely unknown (104, 120). Sequence conservation levels of mammalian intergenic lncRNAs are low, much lower than those of protein-coding sequences and only marginally higher than those of putatively neutral sequences (82, 100). The variable impression that purifying selection has made on these exons and introns of noncoding loci, however, is qualitatively similar to that on protein-coding genes (**Figures 3** and **4**). Nevertheless, across mammalian evolution the strength of this selection on mature lncRNA sequences has been only modest, and sequence conservation only marginally exceeds that seen for intronic and untranscribed intergenic sequences (**Figure 4**). Indeed, based on evidence for purifying selection on inserted or deleted sequences, only 5% of all bases in mouse RNA transcripts are estimated to be functional (100).

This low level of conservation might imply that such lncRNAs are rarely functional. Instead, they could be associated with transcriptional noise (114), perhaps emanating from transcription at a neighboring locus (27). If biologically relevant, the lack of sequence conservation of transcribed intergenic noncoding loci is likely to reflect the presence in RNA transcripts of only short patches of functional sequence involved in base-pairing or protein interactions, similar to the limited MREs in 3′ UTRs of protein-coding genes (13, 16). The observed low nucleotide conservation of lncRNAs could also be associated with functional redundancy (24, 108). In contrast, rapid divergence could result from compensatory mutations that are rapidly fixed when they mitigate the deleterious effect of a second mutation at a functional site. This mechanism has been proposed to explain both the accumulation of potentially deleterious mutations within protein-coding sequences or regulatory elements (67, 130) and the maintenance of RNA secondary structures (98).

3.3. Spectrum of Selection Across Other Noncoding Sequence

Protein-coding sequence is thus the principal substrate of negative selection. However, a significant minority of intergenic noncoding sequence is also subject to purifying selection (**Figure 2**). When the mouse genome was first sequenced and compared with the human genome, investigators realized that most of the constrained sequence lay outside of the protein-coding sequence. In addition to the 1.2% of the human genome that encodes proteins, nearly 4% was estimated to be constrained (78, 87, 109) (**Figure 2**). An analysis of 29 placental mammal genomes showed that approximately 39% of the 3.6 million constrained elements lie within intergenic sequences, more than 2 kb away from annotated gene models (78). Experiments in zebrafish and mice indicate that many of these sequences are biologically functional, with potential roles as enhancers, insulators, or promoters (78, 97, 122, 128).

Purifying selection on noncoding sequence acts most stringently on bases adjacent to transcription start sites and rapidly declines in strength for approximately 200 nucleotides farther upstream (14, 115) (**Figure 3**). These upstream regions of protein-coding genes have experienced unusually high levels of nucleotide substitutions, short insertions and deletions, and transposable element insertions during primate evolution (115, 126). Rather than being the result of pervasive positive selection, this likely reflects elevated mutation rates in sequence that is made accessible and more mutable by the act of transcription initiation. A higher mutability for the upstream regions of

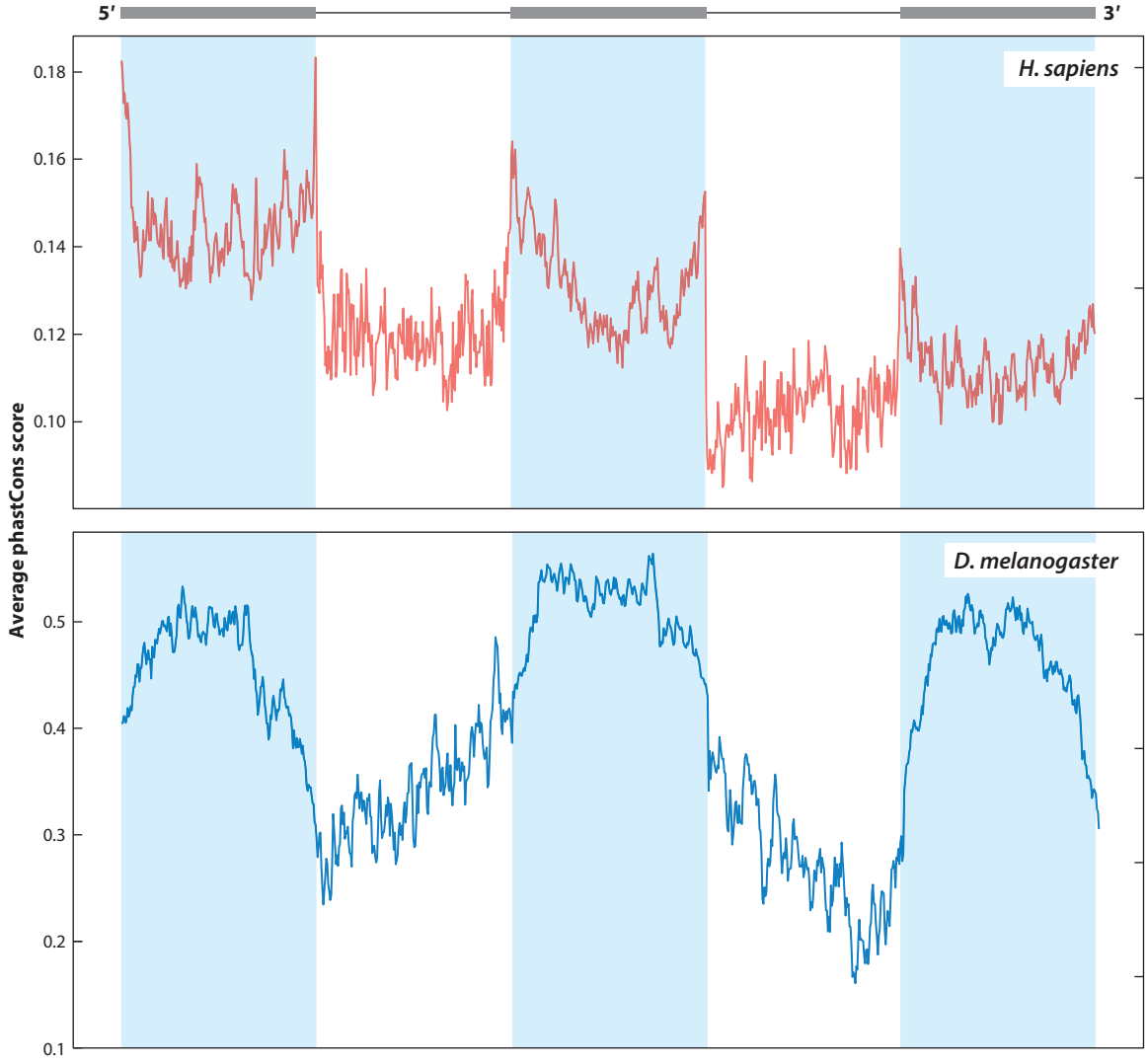

Figure 4

Average nucleotide conservation (phastCons) scores sampled across intergenic long noncoding RNA loci in *Homo sapiens* and *Drosophila melanogaster*.

genes that are frequently transcribed in the germline also likely explains the otherwise curious observation that promoters of non-protein-coding RNA genes tend to be better conserved than promoters of protein-coding genes (3, 82, 100). At the 3′ termini of genes, sequence in the vicinity of the polyadenylation site (motif AWUAAA) also exhibits a trend for more stringent purifying selection, presumably owing to its requirement by cleavage and polyadenylation specificity factors (23, 90).

The effects of mutations in noncoding sequence can sometimes be predicted only by the scrutiny of purifying selection and are not corroborated by the scrutiny of laboratory experiments. For example, the deletion of long intergenic sequences containing more than 1,000 noncoding

elements, each conserved between human and rodents (>70% sequence identity over 100 base pairs), in laboratory mice has failed to reveal phenotypic effects with respect to litter size, body weight, or longevity (3). Similarly, the deletion of elements that are completely conserved in sequence between humans, mice, and rats or the deletion of a conserved enhancer region involved in limb patterning leads to viable and fertile mice lacking overt phenotypes (3, 23). Knockout mutant mice for three intergenic lncRNA loci (*Hotair*, *Malat1*, and *Neat1*) also show no overt phenotype, even though these loci are among the most highly expressed and most highly conserved of all intergenic lncRNAs (29, 89, 90, 108, 132).

Effective population size (N_e): the number of individuals in an idealized population that show a level of loss of heterozygosity due to genetic drift equivalent to that of the actual population

TFBS: transcription factor binding site

How can the absence of overt phenotypes be reconciled with the deep phylogenetic conservation of these elements? Different hypotheses have been proposed to resolve this conundrum. First, phenotypes not observable under laboratory conditions may be revealed under more natural conditions. Knockout mutants for the *BC1* locus, for example, showed no phenotypic changes except when assayed in natural conditions (75). Second, changes may be missed owing to shallow phenotyping. Fewer than one-fifth of genes in either *Caenorhabditis elegans* or *Saccharomyces cerevisiae* initially yielded phenotypic effects when disrupted, yet upon more extensive investigation, larger proportions (42–60% and 97% of genes, respectively) were found to yield significant effects (55, 105). Natural selection is thus better able to discriminate deleterious from neutral variants than are laboratory experiments.

3.4. Evolutionary Turnover of Functional Sequence

Most analyses described thus far successfully associate deep phylogenetic conservation with sequence functionality (78, 87). However, methods that identify genomic regions by their reduced rates of substitution sacrifice sensitivity for specificity, leading to a relatively high rate of false-negative predictions. Sequences of regulatory regions that have been subject to weak selective pressure or lineage-specific evolution will diverge rapidly (9, 38, 83). This may explain, in part, why the pilot ENCODE project found that nearly half of all annotated functional elements were not conserved across species (30).

A sequence may be constrained in one species yet not conserved in others if it has gained functionality only recently; another sequence may be relatively well conserved yet be without constraint if it has recently lost its functionality. These events represent turnover, occurring either when environmental changes or sequence mutations alter the functionality of the locus, or when a functional element is lost through genetic drift or gained when (for example) such an element is acquired that compensates for prior losses. Turnover will occur most frequently for elements that are subject to low constraint and for species or sequences that are associated with relatively low effective population sizes. New functional elements will emerge most frequently from previously functionally inert sequence when their lengths are short (57, 66). Consequently, relatively long and highly constrained protein-coding sequence is only rarely turned over, whereas short noncoding elements, such as transcription factor binding sites (TFBSs) and enhancers, turn over rapidly.

Half of all functional noncoding sequence is predicted to have turned over in approximately the past 130 million years of mammalian evolution (106). Experimental studies of TFBSs support these evolutionary predictions, with 41–89% of such events not being conserved between humans and mice for four liver transcription factors (91) and almost half of all intergenic lncRNA loci having been gained or lost in the interval since the last common ancestor of mice and rats (72). Target sites of miRNA have turned over more slowly, with 20% of such sites turning over among diverse mammals (129). A prominent example of a sequence that has been the frequent substrate of functional loss and gain is human HAR1F (99), a lncRNA that has gained 18 substitutions in

118 nucleotides since humans' last common ancestor with chimpanzees. These substitutions are very strongly biased toward G or C bases, indicating that, rather than being adaptive changes, they are the consequences of a mutational process, specifically recombination-associated GC-biased gene conversion (44).

3.5. Selection on Other Metazoan Genomes

The sequencing of the genomes of 12 *Drosophila* species enabled an assessment of whether the signatures of selection found for mammalian genes and genomes are also evident for these invertebrates (21, 47). In the main, the patterns of selection seen for *D. melanogaster* genes mirror well those found for mammalian gene models: Their protein-coding exons are the most conserved, and their introns are the least (21, 86) (**Figure 3**); a similar proportion (22%) of fruit fly genes' 4D sites appear to have been subject to selective constraints (22, 74); and sequences upstream of their transcription start sites are diverged, a likely consequence of a lower degree of selection relative to transcribed regions (80). Additionally, TFBSs, insulator sites, and RNA polymerase II binding sites all exhibit moderately strong conservation patterns across these drosophilids, as (to a lesser extent) do UTRs (86).

Nonetheless, several important differences are apparent. In contrast to mammalian coding sequence, codon usage is under selection in fruit flies, which strongly modulates protein sequence evolution (53, 110). Moreover, introns in *Drosophila* species are subject to a substantial degree of selection (46), excepting those that are short (<86 nucleotides) (22, 74) (**Table 1**). Perhaps the most striking difference with mammalian sequence is that fruit fly lncRNA loci are moderately well conserved across fly species (86, 131), particularly in their exons (48) (**Figure 4**). This could reflect a more central functional role of these noncoding genes in fruit fly evolution and biology. Nevertheless, such differences are best explained through the application of the theoretical framework provided by population genetics. Indeed, more generally, the ability to distinguish between transcriptional noise and signal, or between random and consequential binding events, or between ephemeral or long-lasting functional sequences requires methods that go beyond species-level conservation. To achieve these distinctions requires evidence of constraint among individuals of a species.

4. CONTEMPORARY SELECTION: POPULATION-BASED ESTIMATES

Comparisons between mammalian genomes can only infer the strength of selection that occurred over long time periods, often tens of millions of years, that separate these species from their last common ancestor. Moreover, interspecies analyses produce results that average across the phylogenetic scope of the study and may not hold true for individual lineages or for extant species. If the species being compared are too distant phylogenetically, then only the most conserved elements will be detected, at the cost of low sensitivity (high false-negative rate). By contrast, if the selected species are too closely related, regions that are neutral may wrongly be considered to evolve under selective constraints, owing to selection acting at linked sites (see sidebar, Background Selection), leading to a high false-positive discovery rate (10, 28, 113). Finally, sequencing and alignment errors can also significantly inflate the neutral evolutionary rate for closely related species.

The strength of selection can instead be inferred using population genetic approaches that consider a shorter, more recent timescale, such as the past 35,000–45,000 years, which is the average age of a new (i.e., derived) variant in the human population (42). Deleterious mutations are expected to segregate at a lower population frequency than mutations occurring at neutral sites.

DAF: derived allele frequency

Genetic hitchhiking: the process by which positive selection at one site leads to an increased allele frequency at a linked neutral site

The converse is expected for advantageous mutations: They should become fixed or segregate at a higher frequency than neutral mutations. Accordingly, the recent 1000 Genomes Project analysis as well as an analysis of 6,515 exomes demonstrated a negative correlation between the age of a mutation and its deleterious effect, with most deleterious mutations within the human populations having arisen only within the past 5,000 years (2, 42, 62).

Population-based studies initially investigated the density of polymorphic sites in different sequence categories, based on the idea that polymorphisms in functional sites will have been more frequently purged if they are deleterious and more rapidly fixed if they are advantageous. For example, both 5′ and 3′ sequences flanking genes show an excess of low-frequency polymorphisms relative to intergenic sites, presumably owing to a combination of background selection originating from linked protein-coding sequence and their *cis*-regulatory element content (73, 119). Furthermore, within human populations, both 5′ and 3′ UTRs exhibit a significantly lower density of polymorphic sites than do introns or putatively neutral sequences (pseudogenes, ancestral repeats) (73, 88). Although polymorphic sites tend to be evenly distributed across 3′ UTRs, single-nucleotide polymorphism (SNP) density is significantly lower at the UTR boundaries (and, more specifically, at the polyadenylation sites) and within miRNA binding sites (20, 117).

Analyses of the site frequency spectrum, in particular minor allele frequencies or derived allele frequencies (DAFs), can infer selection not for single functional elements but for many such elements when they are considered together as a class. Low frequencies of alleles could reflect the purging of deleterious variants from the population but could also reflect population structure or past demographic events. Consequently, selection is best inferred by comparing allele frequencies in the sequence class of interest to frequencies in a set of presumed neutrally evolving sequences (**Table 1**). The strength of purifying selection acting on recently arisen deleterious variants can be estimated from the excess density of low-frequency derived alleles in a sequence class (e.g., lncRNA exons) compared with such alleles within putatively neutral sequences (**Table 1**).

As expected, the different strengths of selection at various genomic elements, estimated using this DAF test, mirror those derived from cross-species comparisons. In humans, the skew toward rarer derived alleles, and thus stronger selective constraints, is strongest for variants within both nonsynonymous and splice sites. In contrast, lower selective constraints act on variants in UTRs, introns, and intergenic regions (1, 2, 88). Only 6% of synonymous substitutions are predicted to be deleterious in humans (116), a lower proportion than the 20% inferred from cross-species comparisons (33). As it does not rely on interspecies sequence conservation, the DAF test can identify constraint that has arisen recently. For example, it was used to predict that 30–50% of nonconserved MREs are functional when their mRNA and cognate miRNA are coexpressed (20). The method has also been used in an important demonstration that sets of human noncoding sequences are conserved not because they are subject to low mutation rates but rather because mutations within them have been preferentially purged from the human population (6, 26, 58, 119).

The intersection of human sequence variation (2) and functional annotations (32) permits population-based estimation of selection acting on open chromatin, transcription factor binding, or more generally transcribed sequence (5, 48, 118). TFBSs, such as those for POU-, HOX-, or FOX-domain-containing transcription factors, include both a lower density of SNPs and an excess of low-frequency SNPs relative to putatively neutral sites or similar sequences in the genome that are without evidence of binding (5, 111, 121). Using a more rigorous approach that better corrects for selection's effect of distorting the pattern of polymorphism at neutral sites (background selection, genetic hitchhiking), Arbiza et al. (5) estimated that, on average, 33% of the nucleotides in a set of high-confidence TFBSs are under selection. They also reported that patterns of selection identified in TFBSs varied greatly depending on the transcription factor

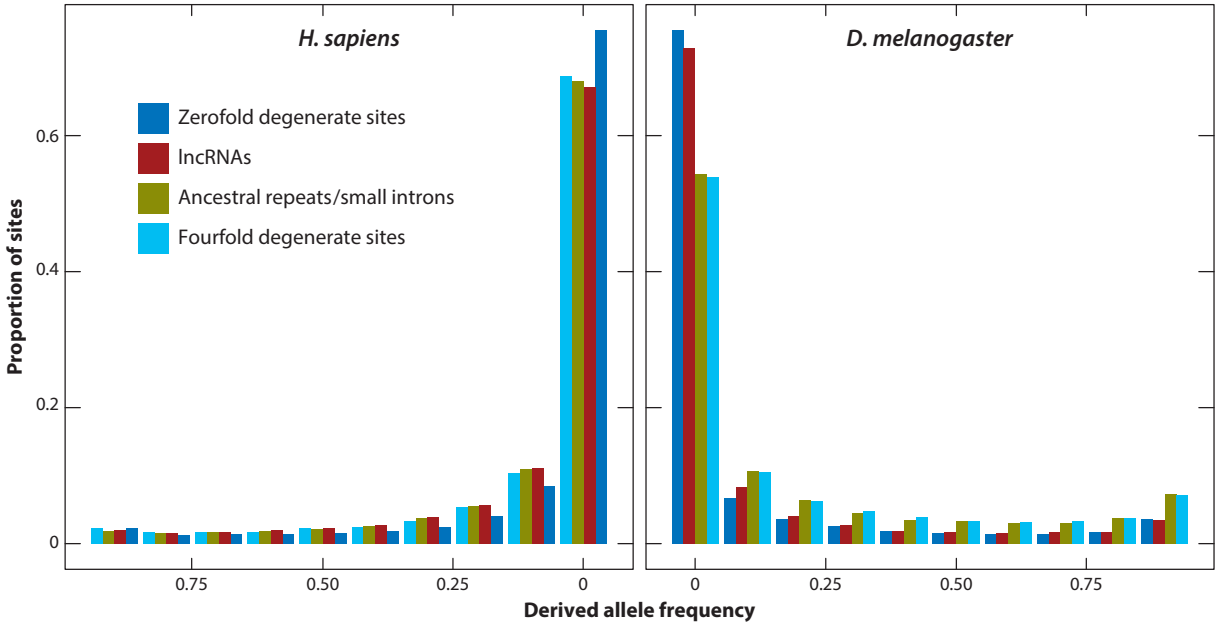

Figure 5

Comparison of derived allele frequency spectra in *Homo sapiens* and *Drosophila melanogaster* for zerofold (*dark blue*) and fourfold (*light blue*) degenerate sites, intergenic long noncoding RNAs (lncRNAs) (*red*), and neutrally evolving ancestral repeats (*H. sapiens*) or small introns (*D. melanogaster*) (*dark yellow*). Adapted from Haerty & Ponting (48).

(5, 88) and that the strength of selection depended on both the size of the binding site and the number of degenerate positions in the consensus motif.

It is striking that applying the DAF test to sequences and populations from other species yields results that contrast with those obtained from the human population. *Drosophila* conserved noncoding sequences, for example, appear to be under greater purifying selection than such sequences in humans (15). This is most evident for intergenic lncRNAs, which in fruit flies, but not in humans, exhibit a clear excess of low-frequency alleles (48) (**Figure 5**). These differences extend to other mammals because selective constraints on conserved noncoding sequences, including promoters, in rodents appear to be almost twice as strong as those in the human population (0.53 ± 0.01 and 0.29 ± 0.02, respectively) (60, 115). These differences are explained below as a direct effect of these species' contrasting effective population sizes.

Comparison of DAF test results between species is complicated by their population size histories and structures. The human population underwent multiple bottlenecks of various constrictions that reduced its effective population size to approximately 1,200 individuals and were then followed by demographic expansion (76). These changes strongly bias allele frequencies at neutral sites and, if not properly accounted for, lead to false inferences of selection. This is because there is an enrichment of rare variants at neutral sites in cases of population expansion (43) and an increased frequency of intermediate variants in cases of population contraction or bottlenecks (123, 127). Additionally, population range expansions associated with founder effects at the margin of the population spatial distribution can lead to a strong bias in allele frequency that mimics the action of positive selection (also called allele surfing) (65).

Methods that correct for the otherwise confounding demographic effects have been developed that infer the distribution of fitness effects from allele frequencies (127). Three methods, developed

by Eyre-Walker et al. (36), Boyko et al. (11), and Keightley & Eyre-Walker (59), estimate demographic parameters from user-provided putatively neutral sites to correct the frequency spectrum of sites of interest (11, 34, 36, 59, 127). Each of these methods is based on predictions from the nearly neutral theory of evolution (64, 92, 93) together with assumptions that adaptive polymorphisms should be very rare, because they are fixed rapidly, and that the majority of nonneutral mutations are deleterious. Consequently, based on the product of the effective population size (N_e) and the selection coefficient, mutations are classified as being effectively neutral ($|N_e s| < 1$), weakly deleterious ($1 < |N_e s| < 10$), deleterious ($10 < |N_e s| < 100$), or strongly deleterious ($|N_e s| > 100$). These algorithms have been applied to polymorphism data in humans, mice, and fruit flies to assess the variable strength of selection at both coding and putatively functional noncoding sites (**Figure 6**). In humans, 27–38% of mutations at nonsynonymous sites were deemed effectively neutral, 21–30% were weakly deleterious, and the remainder were strongly deleterious (11, 34). By comparison, the DAF spectrum of SNPs within intergenic lncRNAs was not different from putatively neutrally evolving neighboring sequences, providing no evidence for the action of purifying selection acting contemporarily on human lncRNAs. Nevertheless, the same analysis applied to intergenic lncRNAs from *D. melanogaster* predicted that approximately 30% of mutations within these loci are weakly deleterious (48) (**Figure 6**).

5. EFFECTIVE POPULATION SIZE AND RELAXATION OF SELECTION WITHIN THE PRIMATE LINEAGE

Many more substitutions at nonsynonymous sites are predicted to be deleterious in *Mus musculus castaneus* or *D. melanogaster* (59, 68) than in humans. Similarly, there appears to be a greater strength of purifying selection on codon usage, lncRNAs, UTRs, and conserved noncoding sequence in fruit flies than in humans (48, 49, 61, 70, 111). These observations are likely to reflect the low value of N_e for humans (1,200–15,000) (76) relative to *D. melanogaster* (1,450,000) (35) or *M. musculus castaneus* (290,000–580,000) (49). This is because N_e conditions the probability that a mutation under selection will be fixed. Any newly arising mutation can be neutral (selection coefficient $s = 0$), deleterious ($s < 0$), or advantageous ($s > 1$). Ohta (92, 93) and Kimura (64) showed that the probability θ that a mutation with $s \neq 0$ will be fixed is $\theta = N_e \mu s$, where μ is the mutation rate. If $|N_e s| > 1$, the mutation will be effectively selected, whereas if $|N_e s| < 1$, it will evolve effectively neutrally, under genetic drift. Consequently, the same mutation whose evolution is nearly neutral in a species of low N_e would have been effectively selected (positively or negatively) in a species with a much larger N_e.

This implies that the human population has accumulated weakly deleterious variants faster than species with larger N_e values. Such a relaxation of selection likely explains the greater-than-expected divergence in functional elements. A greater extent of positive selection in primates would also be consistent with this increased divergence in functional elements (99). Nevertheless, this is not a parsimonious explanation, because it requires positive selection to have acted relatively indiscriminately on very large numbers of elements that in other species are under the greatest constraint. As a consequence, functional elements are expected to turn over rapidly in species with low N_e values, such as primates. Functional elements in species with high N_e values, in contrast, are predicted to persist over longer evolutionary time and thus to have a greater chance of acquiring more fundamental functions.

The evolutionary analyses described above compare the decreased divergence or diversity of sequences of interest to others whose evolution is proposed to have been neutral. Whether these neutral proxy sequences have always escaped selection thus affects quantitatively, although rarely qualitatively, these analyses' results. Of all the commonly used neutral proxies (**Table 1**), 4D

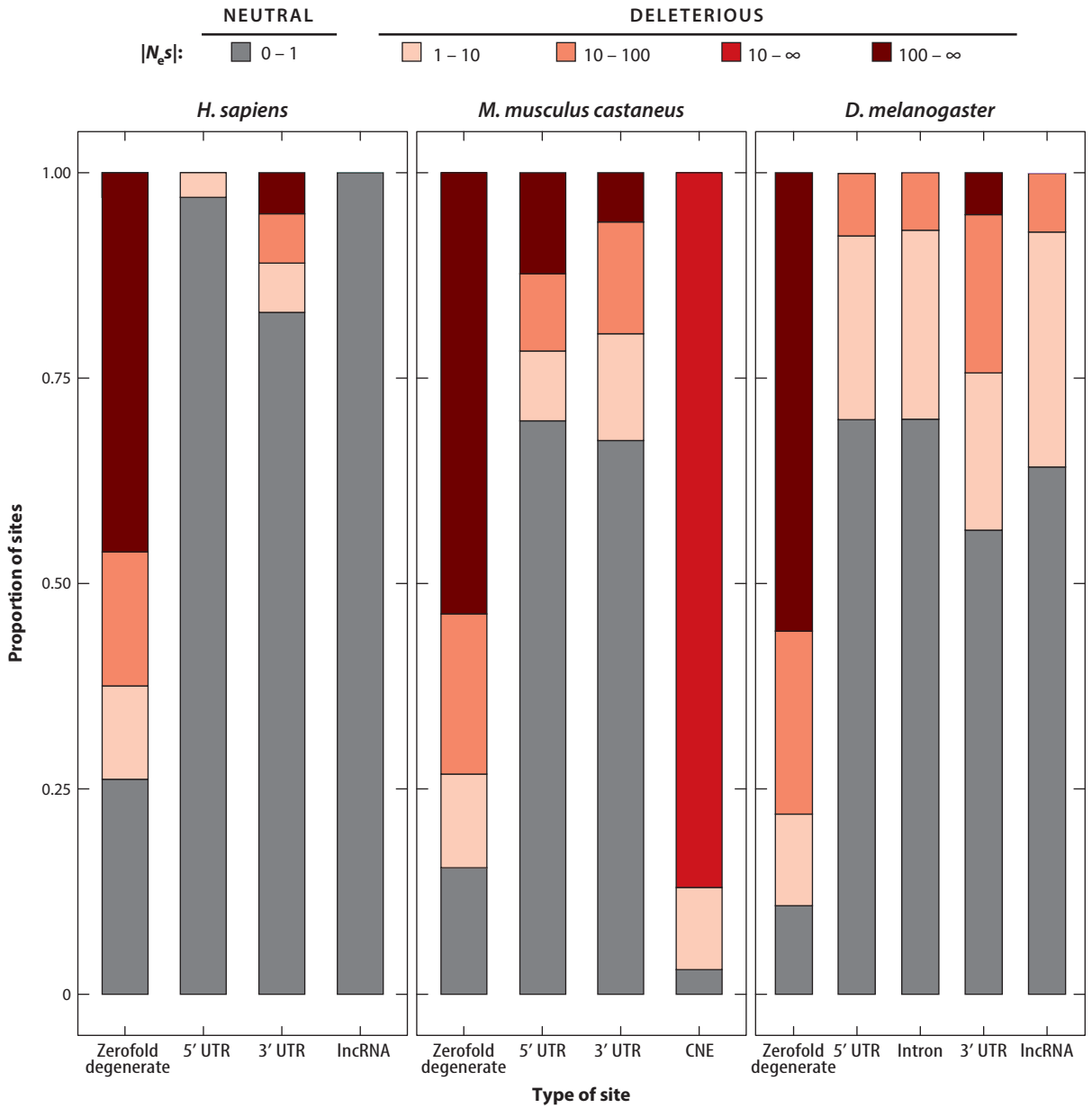

Figure 6

Comparison of the distribution of fitness effects of mutations occurring within 5′ and 3′ untranslated regions (UTRs), zerofold degenerate sites, introns, conserved noncoding elements (CNEs), and long noncoding RNAs (lncRNAs) in *Homo sapiens*, *Mus musculus castaneus*, and *Drosophila melanogaster*. For the CNEs in *M. musculus castaneus*, only three categories are present: $|N_e s| < 1$, $1 < |N_e s| < 10$, and $10 < |N_e s| < \infty$. Data from Haerty & Ponting (48), Eyre-Walker & Keightley (34), Halligan et al. (50), and Kousathanas et al. (69).

sites, particularly in high-N_e species, will be the least applicable. Ancestral repeats, defined as transposable element sequences that inserted prior to the last common ancestor of two species (often humans and mice), are proposed to be the most reliable neutral proxy because virtually all of them have been demonstrated to have evolved neutrally (79, 102). Only 0.2% and 1% of bases in rat–mouse and human–mouse ancestral repeat sequences, respectively, depart from neutrality (84).

Neutral sites are commonly collated across the whole genome, thereby forming a putatively neutral reference against which test sequence is compared. However, this overlooks important confounding issues. First, because test and neutral sites are most often not fully interdigitated, they are likely to not share the same genealogical history (4, 85, 133), and comparisons will not necessarily account for mutation rates, which vary greatly across the genome (31). Second, selection acting at linked sites, via either background selection (see sidebar, Background Selection) or hitchhiking, will distort polymorphism patterns at neutral proxy sites.

The optimal choice of neutral proxy is sequence that has escaped selection and is interdigitated or that lies adjacent to and is compositionally equivalent to sites of interest (for example, nonsynonymous sites, MREs, and lncRNAs) (4, 133). These requirements are met by matching candidate neutral sequence that lies in close physical proximity, and is compositionally similar, to each test sequence. Nevertheless, matches are often not possible when, for example, ancestral repeats of the required composition are absent from the genomic vicinity of the test sequence. Consequently, a compromise procedure, adopted by Halligan et al. (50), Haerty & Ponting (48), and Arbiza et al. (5), is to employ proxy neutral sequence that is noncoding and not conserved in diverse species and that lies outside of annotations such as TFBSs or DNase I–hypersensitive sites. Such sequence constitutes the majority of the reference human genome but only a small proportion of the fruit fly genome.

SUMMARY POINTS

1. Sequences that have long retained functionality are characterized by an increased interspecies sequence conservation that derives from the past purging of mutations that had a deleterious impact on fitness.
2. Purifying selection is strongest on protein-coding sequences, more moderate on flanking noncoding functional sequences (5′ and 3′ untranslated regions), and very weak on intergenic transcribed or bound functional elements that tend to turn over rapidly in evolution.
3. The most powerful approach to assess the current biological relevance of these rapidly evolving elements is to analyze variants' site frequency spectra.
4. Because a species' effective population size directly affects the efficacy of natural selection on mutations, classes of elements identified as evolving under selective constraints in one species may be effectively neutral in another species of smaller effective population size.
5. The observation of a biochemical event at a locus is necessary but not sufficient to infer that it is functional. Instead, indication of biochemical activity, such as transcription or transcription factor binding, combined with the inference of selective constraint between or within species provides the most compelling evidence for the functionality of elements that can then be prioritized for subsequent experimental validation.

FUTURE ISSUES

1. With decreasing sequencing costs of large numbers of samples, population genomic approaches that assess the functionality of elements within the genome will spread. This will allow the in silico assessment of the biological relevance of a locus that draws upon both evolutionary and functional (for example, transcription or transcription factor binding) evidence.
2. Not all transcriptional or binding events will be biologically relevant. However, we do not have a clear understanding of the expected background levels of such biological noise when interpreting functional genomic data. The use of more quantitative methods to investigate the strength of transcription factor binding (77) will also greatly help to identify genomic regions in which binding is only transient and likely inconsequential.
3. Many tests for selection were initially developed for protein-coding sequence and only later adapted for non-protein-coding sequence. There is a considerable need for new tests that consider the deleterious effects of mutation and are tailored specifically to functional noncoding elements, because these are the primary substrate of rapid turnover along different evolutionary lineages.
4. We know almost nothing about the relative impacts that disruption of classes of functional noncoding sequence—including microRNA response elements, transcription factor binding sites, and long noncoding RNA loci—have on fitness and phenotypes. There is thus a pressing need for large-scale in vivo and cellular mutagenesis and phenotyping projects.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Jim Kent (University of California, Santa Cruz) for helpful discussions on the conservation of mouse/human genes. We are grateful to Chris Rands for his comments on the early versions of this review. W.H. and C.P.P. are supported by the Medical Research Council, and this work was also supported by a European Research Council Advanced Grant to C.P.P.

LITERATURE CITED

1. 1000 Genomes Proj. Consort. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–73
2. 1000 Genomes Proj. Consort. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
3. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, et al. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* 5:e234
4. Andolfatto P. 2008. Controlling type-I error of the McDonald–Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180:1767–71
5. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, et al. 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* 45:723–29

6. Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci. USA* 104:12410–15
7. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40:340–45
8. Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116:281–97
9. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* 42:806–10
10. Boffelli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* 5:456–65
11. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083
12. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, et al. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25:1915–27
13. Calin GA, Liu CG, Ferracin M, Hyslop T, Spizzo R, et al. 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* 12:215–29
14. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38:626–35
15. Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol. Biol. Evol.* 24:2222–34
16. Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, et al. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147:358–69
17. Chamary JV, Hurst LD. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. *Mol. Biol. Evol.* 21:1014–23
18. Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* 7:98–108
19. Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134:1289–303
20. Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. *Nat. Genet.* 38:1452–56
21. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–18
22. Clemente F, Vogl C. 2012. Evidence for complex selection on four-fold degenerate sites in *Drosophila melanogaster*. *J. Evol. Biol.* 25:2582–95
23. Cretekos CJ, Wang Y, Green ED, Martin JF, Rasweiler JJ IV, Behringer RR. 2008. Regulatory divergence modifies limb length between mammals. *Genes Dev.* 22:141–51
24. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22:1775–89
25. Dobzhansky T. 1964. Biology, molecular and organismic. *Am. Zool.* 4:443–52
26. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* 38:223–27
27. Ebisuya M, Yamamoto T, Nakajima M, Nishida E. 2008. Ripples from neighbouring transcription. *Nat. Cell Biol.* 10:1106–13
28. Eddy SR. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* 3:e10
29. Eißmann,Gutschner T, Hämmerle M, Günther S, Caudron-Herger M, et al. 2012. Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA Biol.* 9:1076–87
30. Elgar G, Vavouri T. 2008. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends Genet.* 24:344–52
31. Ellegren H, Smith NG, Webster MT. 2003. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* 13:562–68
32. ENCODE Proj. Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74

33. Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol. Biol. Evol.* 27:177–92
34. Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* 26:2097–108
35. Eyre-Walker A, Keightley PD, Smith NG, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* 19:2142–49
36. Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173:891–900
37. Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268
38. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* 312:276–79
39. Fisher WW, Li JJ, Hammonds AS, Brown JB, Pfeiffer BD, et al. 2012. DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proc. Natl. Acad. Sci. USA* 109:21330–35
40. Friedman RC, Farh KKH, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19:92–105
41. Fu W, Akey JM. 2013. Selection and adaptation in the human genome. *Annu. Rev. Genomics Hum. Genet.* 14:467–89
42. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–20
43. Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709
44. Galtier N, Duret L, Glémin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends Genet.* 25:1–5
45. Gazave E, Marqués-Bonet T, Fernando O, Charlesworth B, Navarro A. 2007. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol.* 8:R21
46. Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* 6:R67
47. Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ravi Ram K, et al. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics* 177:1321–35
48. Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol.* 14:R49
49. Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet.* 6:e1000825
50. Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. *Mol. Biol. Evol.* 28:2651–60
51. Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* 9:e1003569
52. Hedges SB. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* 3:838–49
53. Heger A, Ponting CP. 2007. Variable strength of translational selection among 12 *Drosophila* species. *Genetics* 177:1337–48
54. Hellmann I, Zöllner S, Enard W, Ebersberger I, Nickel B, Pääbo S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* 13:831–37
55. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, et al. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* 320:362–65
56. Hurst LD. 2002. The *Ka*/*Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486
57. Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–26
58. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, et al. 2007. Human genome ultraconserved elements are ultraselected. *Science* 317:915
59. Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177:2251–61

60. Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res.* 15:1373–78
61. Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3:e42
62. Kiezun A, Pulit SL, Francioli LC, van Dijk F, Swertz M, et al. 2013. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS Genet.* 9:e1003301
63. Kim SY, Pritchard JK. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet.* 3:1572–86
64. Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge Univ. Press
65. Klopfstein S, Currat M, Excoffier L. 2006. The fate of mutations surfing on the wave of a range expansion. *Mol. Biol. Evol.* 23:482–90
66. Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19:1752–59
67. Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. USA* 99:14878–83
68. Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics* 193:1197–208
69. Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. *Mol. Biol. Evol.* 28:1183–91
70. Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. *Hum. Mol. Genet.* 14:2221–29
71. Künstner A, Nabholz B, Ellegren H. 2011. Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection. *Genome Biol. Evol.* 3:1381–89
72. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, et al. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* 8:e1002841
73. Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, et al. 2012. Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150:457–69
74. Lawrie DS, Messer PW, Hershberg R, Petrov DA. 2013. Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genet.* 9:e1003527
75. Lewejohann L, Skryabin BV, Sachser N, Prehn C, Heiduschka P, et al. 2004. Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. *Behav. Brain Res.* 154:273–89
76. Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–96
77. Lickwar CR, Mueller F, Lieb JD. 2013. Genome-wide measurement of protein-DNA binding dynamics using competition ChIP. *Nat Protoc.* 8:1337–53
78. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478:476–82
79. Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* 2:e5
80. Main BJ, Smith AD, Jang H, Nuzhdin SV. 2013. Transcription start site evolution in *Drosophila*. *Mol. Biol. Evol.* 30:1966–74
81. Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12:1827–36
82. Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* 10:R124
83. McGaughey DM, Vinton RM, Huynh J, Al-Saif A, Beer MA, McCallion AS. 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at *phox2b*. *Genome Res.* 18:252–60
84. Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.* 20:1335–43
85. Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald–Kreitman test. *Proc. Natl. Acad. Sci. USA* 110:8615–20
86. modENCODE Consort., Roy S, Ernst J, Kharchenko PV, Kheradpour P, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330:1787–97

87. Mouse Genome Seq. Consort. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
88. Mu XJ, Lu ZJ, Kong Y, Lam HYK, Gerstein MB. 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res.* 39:7058–76
89. Nakagawa S, Ip JY, Shioi G, Tripathi V, Zong X, et al. 2012. Malat1 is not an essential component of nuclear speckles in mice. *RNA* 18:1487–99
90. Nakagawa S, Naganuma T, Shioi G, Hirose T. 2011. Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J. Cell Biol.* 193:31–39
91. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, et al. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* 39:730–32
92. Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98
93. Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. *Theor. Popul. Biol.* 49:128–42
94. Osada N, Hirata M, Tanuma R, Kusuda J, Hida M, et al. 2005. Substitution rate and structural divergence of 5′UTR evolution: comparative analysis between human and cynomolgus monkey cDNAs. *Mol. Biol. Evol.* 22:1976–82
95. Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol.* 5:e14
96. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol. Biol. Evol.* 27:1226–34
97. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499–502
98. Piskol R, Stephan W. 2008. Analyzing the evolution of RNA secondary structures in vertebrate introns using Kimura's model of compensatory fitness interactions. *Mol. Biol. Evol.* 25:2483–92
99. Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–72
100. Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* 17:556–65
101. Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res.* 21:1769–76
102. Ponting CP, Lunter G. 2006. Signatures of adaptive evolution within human non-coding sequence. *Hum. Mol. Genet.* 15(Suppl. 2):R170–75
103. Ponting CP, Nellåker C, Meader S. 2011. Rapid turnover of functional sequence in human and other genomes. *Annu. Rev. Genomics Hum. Genet.* 12:275–99
104. Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* 136:629–41
105. Ramani AK, Chuluunbaatar T, Verster AJ, Na H, Vu V, et al. 2012. The majority of animal genes are required for wild-type fitness. *Cell* 148:792–802
106. Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* In press
107. Roy SW, Penny D, Neafsey DE. 2007. Evolutionary conservation of UTR intron boundaries in *Cryptococcus*. *Mol. Biol. Evol.* 24:1140–48
108. Schorderet P, Duboule D. 2011. Structural and functional differences in the long non-coding RNA *Hotair* in mouse and human. *PLoS Genet.* 7:e1002071
109. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034–50
110. Singh ND, Larracuente AM, Clark AG. 2008. Contrasting the efficacy of selection on the X and autosomes in *Drosophila*. *Mol. Biol. Evol.* 25:454–67
111. Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, et al. 2012. Analysis of variation at transcription factor binding sites in *Drosophila* and humans. *Genome Biol.* 13:R49

112. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* 123:1133–46
113. Stone EA, Cooper GM, Sidow A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* 6:143–64
114. Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14:103–5
115. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM. 2006. Heterotachy in mammalian promoter evolution. *PLoS Genet.* 2:e30
116. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69
117. Thomas LF, Sætrom P. 2012. Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microRNA-regulation. *PLoS Comput. Biol.* 8:e1002621
118. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489:75–82
119. Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, et al. 2009. Evolutionary processes acting on candidate *cis*-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet.* 5:e1000592
120. Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* 154:26–46
121. Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, et al. 2012. Personal and population genomics of human regulatory variation. *Genome Res.* 22:1689–97
122. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, et al. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat. Genet.* 40:158–60
123. Wakeley J, Aliacar N. 2001. Gene genealogies in a metapopulation. *Genetics* 159:893–905
124. Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337:1675–78
125. Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: Splicing related constraints on protein composition and evolution are common in intron-rich genomes. *Genome Biol.* 9:R29
126. Warnefors M, Pereira V, Eyre-Walker A. 2010. Transposable elements: insertion pattern and impact on gene expression evolution in hominids. *Mol. Biol. Evol.* 27:1955–62
127. Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA* 102:7882–87
128. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3:e7
129. Xu J, Zhang R, Shen Y, Liu G, Lu X, Wu CI. 2013. The evolution of evolvability of microRNA target sites in vertebrates. *Genome Res.* 23:1810–16
130. Yokoyama KD, Thorne JL, Wray GA. 2011. Coordinated genome-wide modifications within proximal promoter *cis*-regulatory elements during vertebrate evolution. *Genome Biol. Evol.* 3:66–74
131. Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, et al. 2012. Identification and properties of 1,119 lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol. Evol.* 4:427–42
132. Zhang B, Arun G, Mao YS, Lazar Z, Hung G, et al. 2012. The lncRNA *Malat1* is dispensable for mouse development but its transcription plays a *cis*-regulatory role in the adult. *Cell Rep.* 2:111–23
133. Zhen Y, Andolfatto P. 2012. Methods to detect selection on noncoding DNA. *Methods Mol. Biol.* 856:141–59