What fraction of the human genome is functional?

Chris P. Ponting and Ross C. Hardison

Genome Res. published online August 29, 2011
Access the most recent version at doi:10.1101/gr.116814.110

Review

What fraction of the human genome is functional?

Chris P. Ponting[1,3] and Ross C. Hardison[2,3]

[1]MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom; [2]Pennsylvania State University, Department of Biochemistry and Molecular Biology, University Park, Pennsylvania 16802, USA

Many evolutionary studies over the past decade have estimated α_{sel}, the proportion of all nucleotides in the human genome that are subject to purifying selection because of their biological function. Most of these studies have estimated the nucleotide substitution rates from genome sequence alignments across many diverse mammals. Some α_{sel} estimates will be affected by the heterogeneity of substitution rates in neutral sequence across the genome. Most will also be inaccurate if change in the functional sequence repertoire occurs rapidly relative to the separation of lineages that are being compared. Evidence gathered from both evolutionary and experimental analyses now indicate that rates of "turnover" of functional, predominantly noncoding, sequence are, indeed, high. They are sufficiently high that an estimated 50% of mouse constrained noncoding sequence is predicted not to be shared with rat, a closely related rodent. The rapidity of turnover results in, at least, a twofold underestimate of α_{sel} by analyses that measure constraint across the eutherian phylogeny. Approaches that take account of turnover estimate that the steady-state value of α_{sel} lies between 10% and 15%. Experimental studies corroborate the predicted rates of loss and gain of noncoding functional sites. These studies show the limitations inherent in the use of deep sequence conservation for identifying functional sequence. Experimental investigations focusing on lineage-specific, noncoding, and functional sequence are now essential if we are to appreciate the complete functional repertoire of the human genome.

The proportion of all human genomic bases that convey biological function has proved a difficult quantity to predict computationally or to derive experimentally. Prior to the appearance of genome-scale experimental data sets, evolutionary approaches were developed to estimate the proportion, α_{sel}, of all human bases that have been evolutionarily constrained, that is, subject to purifying selection of deleterious alleles. These methods' predictions of α_{sel} are expected to slightly underestimate the true fraction of human functional DNA for two reasons. First, because there will be a small minority of sites whose functionality is not sequence specific, for example, DNA or protein "spacer sequences" whose length or conformation, but not sequence, is required to spatially separate functional elements; second, because there will be a (presumed) small amount of sequence that is functional, but is evolving rapidly, under positive rather than negative selection. These evolutionary approaches took advantage of alignments of newly sequenced genomes from other mammals such as mouse, rat, and dog (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004; Lindblad-Toh et al. 2005). These species' genomes provide suitable evolutionary yardsticks against which the human genome can be compared because of their large amounts of sequence (~40% and 50% for the human–mouse and human–dog comparisons) (Mouse Genome Sequencing Consortium 2002; Lindblad-Toh et al. 2005) that can be aligned with reasonable accuracy. Genome-wide alignments have allowed considerable progress in our understanding of mutational and selective processes and in the gain and loss of lineage-specific, particularly transposable element–derived, sequence.

More recently, cheaper DNA sequencing technologies have allowed in vitro assays to interrogate the entire human genome assembly, thereby functionally annotating DNA bases irrespective of their sequence conservation. These two vantage points—evolution and experimentation—have provided different perspectives on the true value of α_{sel}, which can only be reconciled if α_{sel} is large (exceeding 10%) and also, surprisingly, if the human genome's repertoire of constrained DNA has been continually changing along its evolutionary lineage. Here, we present an overview of these issues, considering first the evolutionary and then the experimental perspective, before commenting on how derived, as well as ancestral, functions will need to be determined if we are to fully appreciate human- or primate-specific biology and traits.

[3]Corresponding authors.
E-mail chris.ponting@dpag.ox.ac.uk.
E-mail rch8@psu.edu.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.116814.110. Freely available online through the *Genome Research* Open Access option.

Genome comparisons

Conservation has long been a touchstone for inferring the functionality of sequence. If sequence retains functionality over long evolutionary time spans, such as since the eutherian radiation ~100 Myr ago, then deleterious alleles will have been selectively purged within each eutherian lineage. In contrast, sequence that has always been free of functionality will accept mutations at an underlying neutral rate, and its sequence similarity will gradually erode over time. Comparing the conservation of a sequence against that of a putatively neutral sequence thus allows its degree of purifying selection (constraint) to be inferred. Despite such comparisons being the mainstay of evolutionary genomics for over a decade, they might present an incomplete, and thus misleading, picture (Pheasant and Mattick 2007). For although sequence that has retained constraint across an entire phylogeny might easily be identified, sequence that has gained or lost functionality on one or more lineages will be difficult to distinguish from among neutral sequence that has always been devoid of function (Fig. 1).

Many evolutionary methods that infer α_{sel} assume the absence of purifying selection in a fraction of genomic sequences. Often these are "ancestral repeats" (ARs), which are aligned transposable element–derived sequences present in the last common ancestor of the species under consideration (Mouse Genome Sequencing Consortium 2002; Chiaromonte et al. 2003; Margulies

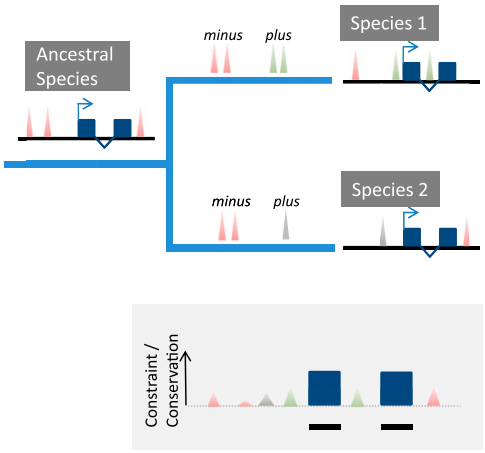

Figure 1. Conservation and turnover of functional sequence. Functional DNA, such as a spliced coding gene (blue) or regulatory elements (red triangles) present in the last common ancestor of two species, may become nonfunctional (*minus*) or be augmented by newly arisen regulatory sequence (*plus*) in a lineage-specific manner (green or gray triangles represent such derived functional sequence). Once the orthologous sequence from these two extant species is compared (*below*), then conservation is strong for retained ancestral functional sequence (here, coding exons, underlined) but is much weaker, and possibly undetectable, for lost ancestral (red) or lineage-specific (green, gray) sequence.

et al. 2003). If large proportions of ARs, instead, have been constrained, then α_{sel} will be greatly underestimated (Smith et al. 2004; Pheasant and Mattick 2007). However, any method that assumes AR neutrality should not immediately be discounted as being unreliable. This is because evolutionary models that exploit substitutions or insertions and deletions (indels) detect <1% of human–mouse ARs as having been subject to purifying selection (Lunter et al. 2006). Furthermore, AR evolutionary rates mirror closely those of pseudogenes and fourfold degenerate sites whose evolution is widely considered to be free of selection (Eory et al. 2010). Despite most ARs evolving neutrally, their substitution rates vary considerably across the mammalian genome (Hardison et al. 2003). In contrast, neutral rates are relatively uniform across short (<5-Mb) regions that show relatively homogeneous nucleotide content (Gaffney and Keightley 2005). Neutral rates can thus be accurately estimated for ARs and then used as a benchmark to infer the degree of constraint for their genomically neighboring sequence.

An initial estimate of α_{sel} from human–mouse alignments

The first, and most renowned, estimate of α_{sel} was described in the publication marking the sequencing of the mouse genome (Mouse Genome Sequencing Consortium 2002) but was more fully explained elsewhere (Chiaromonte et al. 2003). The human genome was divided into windows (size, W), and only those with at least T bases aligning to mouse were retained. The method produces a score that normalizes the sequence divergence in each window by that for local ARs. For neutral sequence, these scores are symmetrically distributed around zero. For genome-wide windows, however, there is a marked excess of windows with positive scores within which purifying selection on substitutions has occurred. The extent of this excess predicts the proportion of windows that were under purifying selection as being between 2.3% and 6.2%, depending on the choice of parameters (W = 50, 100 and T = 40–100). These results were summarized in the mouse genome sequence publication (Mouse Genome Sequencing Consortium 2002) as follows: "the proportion of small (50–100 bp) segments in the mammalian genome that is under (purifying) selection can be estimated to be ~5%."

This estimate was important in three respects. First, it established, using whole-genome alignments, that conserved sequence represents only a minor fraction of human and mouse genomes. Second, it predicts that the amount of constrained, and presumably functional, noncoding DNA is about four times greater than the amount of functional coding sequence (1.06%) (Church et al. 2009). Third, it legitimized the question of the value of α_{sel} and thus provided a precedent for subsequent studies. Nevertheless, this method is limited, first because it estimates the proportion of windows, rather than bases, that are under purifying selection. Thus it will tend to overlook constrained bases when they are distributed diffusely, and to over-count neutrally evolving bases lying within constrained windows. Importantly, the balance between such under- and overestimations in windowing approaches will vary according to evolutionary divergence, strength of selection, and the clustering of constrained bases, which together will lead to variation in α_{sel} estimates across a species phylogeny even when turnover of functional sequence is absent. Its second limitation is that it has most power to infer functionality for sequence that is constrained across both mouse and human lineages; the approach will mostly fail to capture sequence whose functionality is lineage-specific (Fig. 1; Pheasant and Mattick 2007).

The sequencing of the dog genome provided an opportunity to estimate α_{sel} for a second pair of mammalian genomes, namely, human and dog (Lindblad-Toh et al. 2005). This estimate was found to be similar to that for human and mouse genomes (α_{sel} ~ 5.3%) (Lindblad-Toh et al. 2005). Moreover, sequence that was aligned between human and dog, but not to mouse, contained little or no excess conservation (~0.1%). This implies that constrained sequence amounting to α_{sel} ~ 5% was present in the last common ancestral genome of these species and has been inherited by the three extant species with little or no loss of constrained sequence. The analysis, however, was unable to ascertain whether substantial amounts of species-specific constrained sequence have been acquired independently in each of the three lineages.

Estimates from alignments for multiple species

The Chiaromonte et al. (2003) method used pairwise alignments to estimate α_{sel} but is unable to identify where in the human genome most constrained sequence lies. When larger numbers of species' sequences are compared, it was expected that estimates of α_{sel} would become more reliable and that increasing proportions of functional elements would be detectable. The algorithms that were developed in subsequent years took advantage of multiply aligned sequences from many diverse mammalian species, first for single loci (e.g., Cooper et al. 2005) and then for whole genomes (e.g., Lindblad-Toh et al. 2011). These methods differed in several respects. They considered the evolutionary rates either of single base substitutions, some taking account of flanking bases or the pattern of mutations, or of indels; some compared these rates with those in presumed neutrally evolving ARs or fourfold degenerate sites in codons, and others used a random sample of aligned sequence as the neutral control, assuming that most aligned sequence has evolved entirely free of constraint (Table 1). Rather than considering DNA sequence only, one algorithm estimated constraint by calculating

the similarities among DNA structures inferred from hydroxyl radical cleavage patterns (Parker et al. 2009).

Since 2003, 16 publications applying these methods have estimated α_{sel} in the human genome as being between 2.6% and 12%, with an average of ~6.4% (Fig. 2; Table 1; for review, see Ponting et al. 2011). On one hand, it can be argued that these methods have yielded similar estimates of α_{sel} that never exceed 12%, which provides a strong indication that, indeed, most (~90%) of the mammalian genome is functionally inert. On the other hand, because estimates vary by around fivefold (2.3%–12%), it could be considered that there is, as yet, little consensus on the true value of α_{sel} for the human genome.

Each of these methods has its advantages and its deficits. In general, these methods have the greatest power to capture strongly constrained and long elements (with sensitivities for detecting coding sequence of between 65% and 85%) (Cooper et al. 2005; Siepel et al. 2005; Lunter et al. 2006; Davydov et al. 2010) and the least power to detect more weakly constrained or short elements (Pollard et al. 2010). Unlike others, one method (Chai) predicts a substantial proportion (~40%) of ARs to be constrained (Parker et al. 2009). This calls into question the proposed correspondence between the conservation of DNA structure and evolutionary constraint. Because neutral evolutionary rates vary considerably across mammalian genomes (Gaffney and Keightley 2005), methods that assume genome-wide uniformity of the neutral standard will produce inaccurate α_{sel} estimates. For such methods, constraint will be over- or underestimated for regions whose neutral rates are low or high, respectively, compared with the genome-wide average (Li and Miller 2003). The full extent of regions with low constraint is also likely to be underestimated by all sequence-based methods. This is because of the inevitability of a large minority (at least 15%) (Lunter et al. 2008) of aligned bases being incorrectly placed in mammalian whole-genome alignments (Margulies et al. 2007).

Four methods (Margulies et al. 2003; Cooper et al. 2005; Siepel et al. 2005; Asthana et al. 2007) considered multiple alignments of mammalian sequences representing 30 Mb (~1%) of the human genome that were generated as part of the Encyclopedia of DNA Elements (ENCODE) pilot project (The ENCODE Project Consortium 2007). Estimates for α_{sel} produced using this small portion of the mammalian genome are, inevitably, overestimates for the genome as a whole. This is because these ENCODE regions contain approximately twofold and 1.5-fold higher proportions of coding sequence and predicted constrained sequence, respectively (Asthana et al. 2007; Meader et al. 2010).

Unfortunately, these methods have not yet been separately applied to genome pairs sampled across independent mammalian lineages. If they had, might they have found α_{sel} to be constant, which would imply that the repertoire of constrained sequence has been relatively stable across mammalian evolution? Or, might they instead have identified α_{sel} to be more variable, perhaps with its values being higher for more closely related species and lower for more distantly related species, which would be indicative of turnover of functional sequence?

Exponential decay of shared constrained sequence with divergence

Rather than estimating α_{sel} shared across many mammalian genomes, Smith et al. (2004) sought evidence for such turnover using paired alignments from ~1.8 Mb of sequence from eight eutherian mammals. By using an evolutionary simulation that considered variations in mutation rates across these species and across a range of sequence scales, they identified noncoding sequence windows whose conservation exceeded a threshold value, and then subtracted the number of simulated windows that were conserved despite being free of selection. This provided the proportion of the 1.8 Mb noncoding sequence that appears constrained for a pair of species. By subsequently applying this method across pairwise aligned sequence from the eight species' multiple alignment, Smith et al. (2004) were surprised to find that predicted functional noncoding sequence varied greatly, being approximately sevenfold higher between more closely related species, such as mouse and rat, than between more distantly related species, such as mouse and human. Furthermore, they observed that the amount of constrained noncoding sequence predicted for a species pair declined roughly exponentially with these species' divergence.

More formally, a genomic proportion π of sequence does not turn over, while another proportion α^{O}_{sel} decays exponentially over divergence d. Here, $\pi \approx 1.1\%$, the proportion of the human genome that encodes protein. Thus, $(\alpha_{sel} - \pi) = (\alpha^{O}_{sel} - \pi)\exp(-Bd)$. Here B is a constant whose value reflects the rapidity of exponential decline: The divergence over which a 50% reduction occurs in the amount of functional noncoding sequence shared between two species is $d_{1/2} = \ln(2)/B$. Plotting the natural logarithm of $(\alpha_{sel} - \pi)$ against divergence d (Fig. 3) then provides, from the y-intercept, the value of $\alpha^{O}_{sel} - \pi$ at zero divergence. This total amount of functional sequence $\alpha^{O}_{sel} = \alpha_{sel}$ in the limit of no divergence provides an estimate for the fraction of the human genome that is constrained. Values from the Smith et al. (2004) study, which the investigators emphasize are only very approximate, predict $\alpha^{O}_{sel} = 11.0\%$ and a half-life for constrained noncoding sequence of $d_{1/2} = 0.14$ (Fig. 3).

Application of the neutral indel model

Although the Smith et al. (2004) study was important in indicating that functional noncoding sequence is rapidly turning over, it was unable to apply a neutral model to estimate the zero divergence fraction α^{O}_{sel} using genome-wide data. One such neutral model relies not on nucleotide substitutions but on indels to predict sequence under purifying selection (Lunter et al. 2006). In its simplest form, the model assumes that indels fall randomly in a pairwise alignment of neutrally evolving sequence, which immediately implies that the frequency distribution of between-indel distances follows a geometric distribution. Other advantages of the model over other approaches are that it accounts for much of the genome-wide variation in mutation rates and for the clustering of constrained bases, and it can detect lineage-specific functional sequence on a genome-wide scale. The subsequent demonstration of this prediction's validity provided an estimate that shared constraint is evident for fewer than 1% of human–mouse ARs. Applying this model to whole-genome alignments for human and mouse showed a considerable excess of long ungapped alignment blocks relative to the neutral expectation, which presumably reflects the preferential purging within them of deleterious indels during primate or rodent evolution. From this excess, human–mouse α_{sel} was predicted to lie between 2.56% and 3.25% (Table 1; Lunter et al. 2006).

Meader et al. (2010) applied this neutral indel model to genome-wide pairwise alignments for seven eutherian species and reported α_{sel} values that are approximately threefold higher for the more closely related species, such as mouse and rat, than for more distantly related species, such as human and mouse. The coding sequence repertoire is relatively stable across eutherian evolution (Ponting and Goodstadt 2009); so by subtracting the coding sequence portion ($\pi \approx 1.1\%$) from α_{sel}, the authors predicted that

Table 1. Publications describing estimates of α_{sel} and α^0_{sel}

Publication	Method (α_{sel} and α^0_{sel}, estimation)	α_{sel} (%) lower	α_{sel} (%) higher	Substitutions or indels or topography	Neutral model/standard	Whole or partial genome	Multiple, or pair of, genomes	Local or global neutral rate
Lunter et al. (2006)	NIM (α_{sel})	2.56	3.25	Indels	Randomly placed indels	Whole	Pair	Local
Thomas et al. (2003)	MCSs (α_{sel})	3.7	3.7	Substitutions	4D sites	Partial	Multiple	Local
The ENCODE Project Consortium (2007) (ENCODE)	Two of three methods (α_{sel})	4.9	4.9	Substitutions	4D sites/most aligned sites	Partial (ENCODE)	Multiple	Global
Lindblad-Toh et al. (2005)	Substitutions (α_{sel})	5.3	5.3	Substitutions	ARs	Whole	Pair	Local
Pollard et al. (2010)	Various (α_{sel})	5.3	5.3	Substitutions	Various	Partial (ENCODE)	Multiple	Global
Lindblad-Toh et al. (2011)	SiPhy (α_{sel})	5.4	5.4	Patterns	ARs	Whole	Multiple	Global
Eory et al. (2010)	Substitutions (α_{sel})	5.4	5.4	Substitutions	ARs	Whole	Pair	Local
Cooper et al. (2005)	GERP (α_{sel})	5.5	5.5	Substitutions	Most aligned sites	Partial	Multiple	Local
Chiaromonte et al. (2003)	Substitutions (α_{sel})	2.29	6.15	Substitutions	ARs	Whole	Pair	Local
Siepel et al. (2005)	phastCons (α_{sel})	3	8	Substitutions	none	Whole	Multiple	Global
Davydov et al. (2010)	GERP++ (α_{sel})	6	8	Substitutions	Most aligned sites	Whole	Multiple	Global
Smith et al. (2004)	Substitutions/Turnover (α_{sel} and α^0_{sel})	10	10	Substitutions	Simulation	Partial	Multiple pairs	Local
Meader et al. (2010)	NIM (α_{sel} and α^0_{sel})	6.5	10	Indels	Randomly placed indels	Whole	Multiple pairs	Local
Garber et al. (2009)	SiPhy (α_{sel})	5.8	10.2	Patterns	ARs	Partial (ENCODE)	Multiple	Global
Asthana et al. (2007)	SCONE (α_{sel})	5.5	11	Substitutions within trinucleotides	Most aligned sites	Partial (ENCODE)	Multiple	Global
Parker et al. (2009)	Topography (Chai; α_{sel})	12	12	Topography	Most aligned sites	Partial (ENCODE)	Multiple	Global

(4D) Four-fold degenerate; (ARs) ancestral repeats; (ENCODE) Encyclopedia of DNA Elements project; (GERP) Genomic Evolutionary Rate Profiling; (MCSs) multispecies conserved sequence; (NIM) neutral indel model; (SCONE) Sequence CONservation Evaluation.

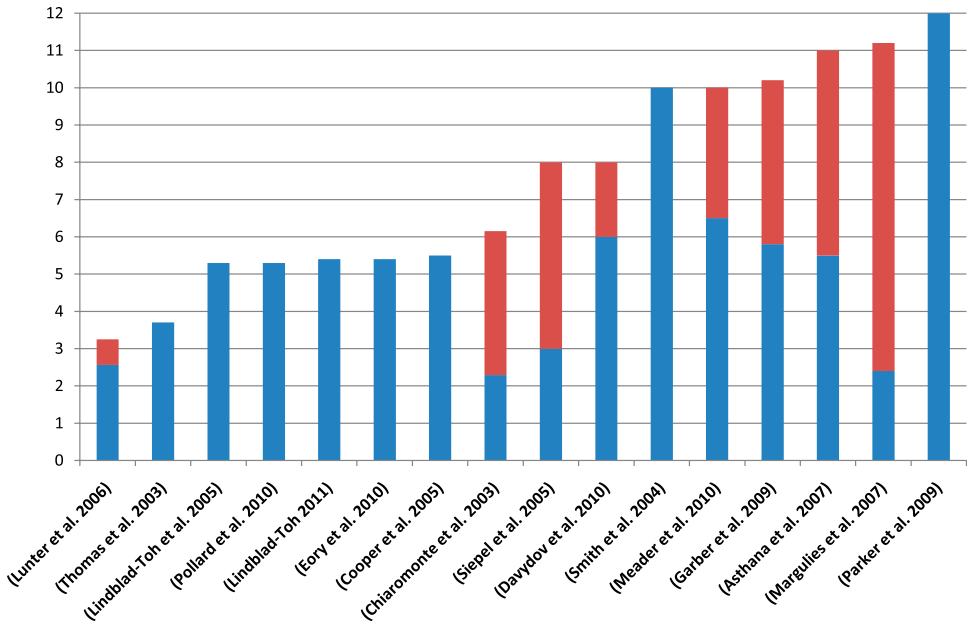

Figure 2. Estimates of α_{sel} from 16 studies ranked by increasing values. Lower and upper bound values are indicated in blue and red, respectively.

these closely related species share more than 5.5-fold more functional noncoding sequence than the more divergent species pairs.

The Meader et al. (2010) results support the Smith et al. (2004) finding that $\alpha_{sel} - \pi$ values decline exponentially with divergence (Fig. 3). Their study predicts that half of the functional noncoding sequence is lost in the time that it takes for two substitutions to occur in 10 bp of neutral sequence; that is, the half-life is about 0.2 in units of nucleotide divergence. Thus, the amount of constrained noncoding sequence shared between human and mouse (1.5% at a divergence, d, of ~0.6) is approximately half that for human and dog (3.0% at $d \sim 0.4$); this, in turn, is about half that for mouse and rat (6.1% at $d \sim 0.2$); and, finally, this amount is approximately half the presumed amount of constrained noncoding sequence present in extant genomes (13.8% at $d \sim 0$). Similarly, this implies that about one-quarter of human constrained noncoding sequence is not shared with rhesus macaque ($d \sim 0.075$) and ~4% is not shared with chimpanzee ($d \sim 0.012$). The neutral indel model also predicts a much higher figure for the constrained portion of the human genome than other approaches: $\alpha^{O}_{sel} = 14.9\%$ (Fig. 2). However, such estimates should be treated with caution as they have wide confidence limits and may be susceptible to non-uniformities in the indel rate that may not have been fully accounted for. Extrapolation furthermore relies on the accuracy of the single-rate exponential decay model to small divergences, for which currently no supporting data exist. For these reasons, Meader et al. (2010) more conservatively estimate α^{O}_{sel} to be 10%.

Experimental evidence for turnover and conservation of functional sequence

DNA sequences involved in regulating gene expression comprise one of the larger classes of functional noncoding sequences. While some gene regulatory sequences have been preserved over long phylogenetic distances, others are subject to turnover. The rapid rate by which these sequences are turned over appears to be compatible with the estimates derived from the neutral indel model.

Strong purifying selection on noncoding DNA sequences has been a productive approach to discovering gene regulatory sequences. An early example was the observation of strikingly similar DNA sequences within an intron of the human, mouse, and rabbit *IGK* genes encoding the immunoglobulin κ light chain (Emorine et al. 1983). Experimental assays showed that this constrained intronic sequence functions as an enhancer (Emorine et al. 1983; Picard and Schaffner 1984). Employing many more genomic sequences, and one of the rigorous methods discussed above for finding DNA sequences likely to be under selection (phastCons elements) (Siepel et al. 2005), it is clear that this intronic enhancer is subject to purifying selection (Fig. 4). High-throughput assays (Wold and Myers 2008) provide direct biochemical evidence that it is bound by the transcription factor complex NFKB in lymphoblastoid cells (Fig. 4; Kasowski et al. 2010; The ENCODE Project Consortium 2011). Many studies over the past 25 yr have successfully utilized signatures of purifying selection in noncoding DNA sequences for predicting regulatory regions (e.g., Aparicio et al. 1995; Gottgens et al. 2000; Flint et al. 2001; Woolfe et al. 2005; Pennacchio et al. 2006; Wang et al. 2006; Visel et al. 2008).

However, results of other lines of investigation emphasize evolutionary changes in regulatory regions. Some enhancers are lineage-specific, found, for example, only in primates (Bodine and Ley 1987) or in mice (Valverde-Garduno et al. 2004). Almost half of the functional transcription factor binding sites in human promoters are not functional in rodents (Dermitzakis and Clark 2002). Putative regulatory regions identified by high-throughput analyses of chromatin immunoprecipitated, factor-bound DNA in 1% of the human genome are rarely deeply conserved across vertebrates, and many do not show clear evidence of evolutionary constraint (The ENCODE Project Consortium 2007; King et al. 2007). Technical limitations in the ability to detect constraint or the accuracy of functional assignments are likely to account for only a small proportion of the large amount of apparently unconstrained but functional DNA sequences. Instead, this lack of constraint could

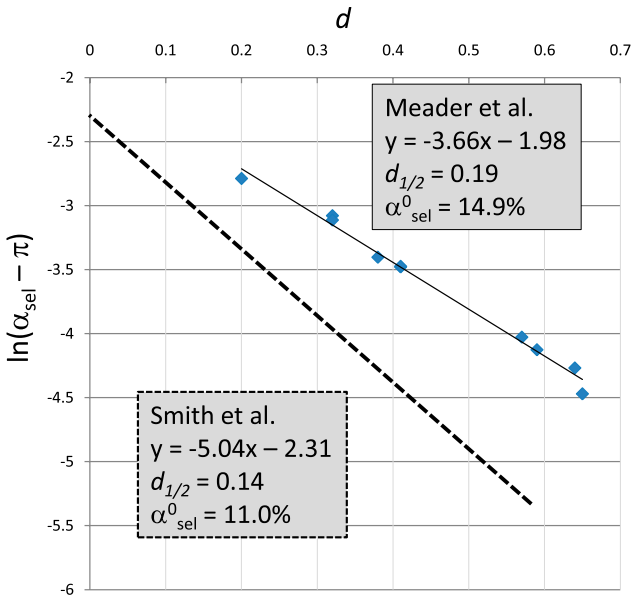

Figure 3. The constrained noncoding fraction of the human genome ($\alpha_{sel} - \pi$) declines exponentially with species divergence, *d*. The regression line for the natural logarithm of ($\alpha_{sel} - \pi$) against *d* for the Smith et al. (2004) study is shown by the broken line. Data points for the Meader et al. (2010) study are shown (blue diamonds), together with their regression line (solid line). The equations for these lines are presented, together with the inferred values of α^0_{sel} and $d_{1/2}$. Meader et al. (2010) data were taken to be the midpoint between lower and upper bound estimates. Divergence values in the Smith et al. (2004) and Meader et al. (2010) studies were estimated from full alignments and from synonymous sites, respectively. As elsewhere in this review, α_{sel} is defined relative to the size of the human genome, rather than to the sizes of different animal genomes.

reflect spreading of a biochemical mark (e.g., histone modifications in chromatin) beyond an initiating element that could itself be constrained. Elements with lineage-specific functions would also show a lack of constraint. Another possibility is that the experimental assays detect a pool of biochemically functional elements (e.g., transcription factor binding sites) that confer no advantage for the most part. However, over time some of these could be mutated to acquire a biological role. Such a reservoir of elements could serve as a source for new elements with a biological function (The ENCODE Project Consortium 2007). These and other studies point to widespread turnover in binding sites within regulatory regions and even for entire *cis*-regulatory modules. The breadth of evolutionary profiles of regulatory regions, from deep phylogenetic preservation to lineage-specific occupancy, suggested by these studies of single loci or selected subsets of genomic intervals is strongly confirmed by genome-wide investigations (Cheng et al. 2009). For the *IGK* example, the well-known enhancer is a strongly conserved binding site for NFKB, but a second NFKB-bound site in the same gene shows no signature for evolutionary constraint (Fig. 4).

Furthermore, when occupancy is directly examined by chromatin immunoprecipitation, conservation of transcription factor binding between different mammals (for example, human vs. mouse or dog) is observed at only 10%–22% of the bound sites in liver (Odom et al. 2007; Schmidt et al. 2010). An even smaller fraction, ~3% of the CEBPA-occupied segments, is bound in all three mammals. Most binding events are thus species-specific, which is indicative of much turnover, that is, loss and gain of binding along each lineage.

The considerable amount of turnover observed in these studies of transcription factor binding in different mammals is consistent with predictions of the rate of turnover of functional, noncoding sequences. The equation in Figure 3 predicts that ~11% of functional noncoding sequences in human would be also found in mouse, at a divergence *d* of about 0.6. This is very similar to the frequency of conserved binding by liver transcription factors between human and mouse (Schmidt et al. 2010). Likewise, the alignability of putative regulatory regions (identified by chromatin immunoprecipitation in a manner agnostic to conservation) follows an exponential decay similar to that presented in Figure 3 (Miller et al. 2007). Hence the turnover model derived from estimating the fraction of DNA segments under selection detected as a function of evolutionary distance fits well with multiple lines of experimental results. Most regulatory regions are undergoing turnover, albeit at a rate slower than the bulk of the genome (Miller et al. 2007). While some have interpreted the turnover of transcription factor–occupied regions as indicating that a majority of these are evolving neutrally (Schmidt et al. 2010), we find that their turnover rate is consistent with the rate generally observed in *functional* noncoding sequences.

So how does one interpret the evolutionary patterns revealed by alignments of DNA sequences or correspondence in binding sites between species? The most deeply preserved regulatory regions are frequently involved in control of genes that encode developmental regulators (Woolfe et al. 2005; Pennacchio et al. 2006; Visel et al. 2008). These comprise only a few percent of putative regulatory regions (King et al. 2007) and only rarely cover enhancers for some tissues, such as heart (Blow et al. 2010). Likewise, transcription factor binding is conserved between species at only a small fraction of factor-bound DNA segments. Conservation of occupancy by liver transcription factors in multiple species is strongly associated with key genes active in that tissue (Schmidt et al. 2010). Thus strong constraint on protein-bound DNA sequences and conservation of protein occupancy are features of particularly important DNA sequences. The strong constraint suggests that these are not subject to the same rate of exponential decay observed for the bulk of the functional noncoding sequences. Only rarely will they be lost and, then, usually when deletion alleles are advantageous (Sagai et al. 2004; Chan et al. 2010). Exactly what makes them so important will be the subject of future work. Perhaps they play a central role in

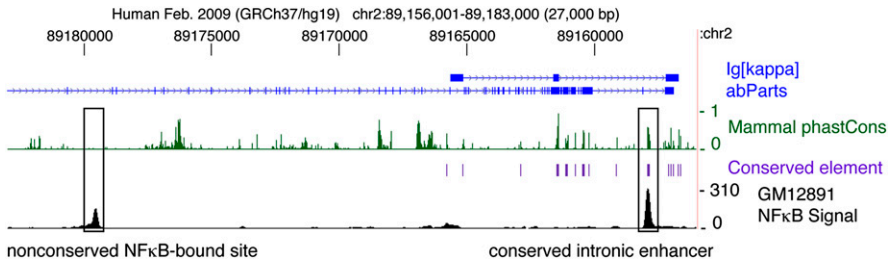

Figure 4. Conservation and apparent turnover of regulatory regions in the *IGK* gene encoding immunoglobulin κ. The intronic enhancer discovered by interspecies conservation of noncoding DNA (Emorine et al. 1983) is toward the *right* end; it lies in a region likely to be under evolutionary constraint, as shown by the mammalian phastCons and conserved element tracks (Siepel et al. 2005). This site and a second, nonconserved site are both bound by NFKB (RELA subunit) in the lymphoblastoid cell line GM12891, shown as the density of mapped ChIP-seq reads on the last track. The ChIP-seq data are from Kasowski et al. (2010) and the ENCODE project (The ENCODE Project Consortium 2011).

regulatory mechanisms, such as providing strong connections with the transcriptional apparatus. Perhaps they are involved in multiple modes of regulation or in multiple tissues. Clearly this special subset of regulatory regions is worthy of intensive investigation. At the other extreme, protein-bound DNA segments found only in one or in very closely related species are perhaps less likely to confer biological function, although experimentally proving such an absence of function presents a considerable challenge (Nobrega et al. 2004). Those that have lineage-specific functions appear to lie close to genes that are enriched for lineage-specific activity, such as immune response genes (King et al. 2007).

However, the majority of regulatory regions are neither narrowly lineage-specific nor conserved over a broad phylogenetic span (The ENCODE Project Consortium 2007; King et al. 2007). Enhancers active in heart are in this less strongly constrained category (Blow et al. 2010). Putative regulatory regions conserved in only a subset of mammals may be enriched for regulation of certain categories of genes (King et al. 2007), although this issue should be examined again with larger data sets. The change, loss, and gain of these regulatory regions, fitting the exponential decay illustrated in Figure 3, may reflect greater degrees of freedom in carrying out their function. For instance, one possible role is modulating the activity of the core regulatory regions (perhaps the strongly constrained regions). Such modulation may be accomplished by a wider range of binding patterns than is found in the core functional regions, and hence these modulatory regions could show more evolutionary turnover.

Gain of functional noncoding sequence

Given the lack of evidence to the contrary, we must assume that all mammalian genomes have contained approximately similar amounts of functional sequence. The exponential decay of shared constrained noncoding sequence (Fig. 3) implies that as quickly as functional sequence is lost at one location, it is gained at another. Such pairs of compensatory events involving ~8-bp transcription factor binding sites are unlikely to occur together within short sequences but become very frequent at the scales (~1 Mb) over which DNA-bound transcription factors exert their effects (Durrett and Schmidt 2008). Consequently, there is likely to be a high degree of functional redundancy among such closely linked sites that together buffer against the complete loss of regulatory functionality.

Functional noncoding sequence may be gained from advantageous mutations within preexisting nonfunctional, neutrally evolving sequence. It may also have been acquired from the insertion of sequence, such as that derived from transposable elements, duplicated from another genomic context. About one-quarter of transcription factor binding sites or promoters appear to have been introduced into genomes via transposable elements (Jordan et al. 2003; Wang et al. 2007; Kunarso et al. 2010), indicating that the transcriptional regulation evolves rapidly and in a lineage-specific manner (Bourque et al. 2008).

Nevertheless, a prediction from the neutral indel model is that the evolution of virtually all ancestral transposable elements (i.e., ARs), present in pairs of divergent eutherian genomes, has been predominantly neutral (Lunter et al. 2006; Meader et al. 2010). Transposable element-derived sequence that has retained functionality over tens of millions of years is rare, occupying only ~1 Mb of the human genome (Lowe et al. 2007). It thus appears that the evolutionary lifespan of transposable element–derived functional sequence is relatively short, of the order indicated by Figure 3.

Conclusions

Evolutionary models and experimental findings now indicate that a surprisingly large portion of the human genome (approximately α^{O}_{sel} = 10%–15%) (Meader et al. 2010) might be functional. Although this is a larger proportion than indicated by initial estimates (Mouse Genome Sequencing Consortium 2002), it is lower than a suggestion that α^{O}_{sel} exceeds 20% (Pheasant and Mattick 2007). Five questions, however, should be addressed in forthcoming years. First, we do not yet know whether the total amount of constrained sequence in the human genome (α^{O}_{sel} multiplied by genome size) differs substantially from amounts for other mammals or for birds. The amount does far exceed numbers of inferred functional nucleotides in fish, fruitflies, or nematode worms (Meader et al. 2010). Second, as additional nonmammalian clades are populated with sequenced genomes, we should be able to assess whether rates of turnover of functional noncoding sequence are equivalent across the animal phylogeny. Third, experimental investigations should determine more comprehensively the proportions of human noncoding sequences derived from either a transposable element or ancestral sequence that have acquired functionality in the primate lineage. Fourth, we will need to investigate to which phenotypic traits, molecular functions, and cellular processes does either de novo functional, or erstwhile functional, sequence contribute most? Finally, what is the complete set of human genomic regulatory regions that can be identified experimentally? To date, experiments have focused primarily on immortalized cell lines, thus leaving most primary cells and developmental stages unstudied. Direct experimental identification of all regulatory sequences would greatly reduce our current reliance on evolutionary approaches and their inherent assumptions.

Acknowledgments

C.P.P. thanks Gerton Lunter and Stephen Meader for helpful discussions over many years and their constructive criticism of early drafts of this manuscript. We thank Kerstin Lindblad-Toh and Manolis Kellis for sharing a manuscript prior to publication and Chris Rands for useful conversations. C.P.P. is funded by the UK Medical Research Council. R.C.H. is supported by grants from the US National Institutes of Health, R01DK065806 and RC2HG005573.

References

Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S. 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. *Proc Natl Acad Sci* **92:** 1684–1688.

Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. 2007. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol* **3:** e254. doi: 10.1371/journal.pcbi.0030254.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42:** 806–810.

Bodine DM, Ley TJ. 1987. An enhancer element lies 3′ to the human A λ globin gene. *EMBO J* **6:** 2997–3004.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, et al. 2008. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* **18:** 1752–1762.

Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327:** 302–305.

Cheng Y, Wu W, Kumar SA, Yu D, Deng W, Tripic T, King DC, Chen KB, Zhang Y, Drautz D, et al. 2009. Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res* **19:** 2172–2184.

Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D. 2003. The share of human genomic DNA under selection estimated from human–mouse genomic alignments. *Cold Spring Harb Symp Quant Biol* **68:** 245–254.
Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, et al. 2009. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7:** e1000112. doi: 10.1371/journal.pbio.1000112.
Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15:** 901–913.
Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6:** e1001025. doi: 10.1371/journal.pcbi.1001025.
Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19:** 1114–1121.
Durrett R, Schmidt D. 2008. Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. *Genetics* **180:** 1501–1509.
Emorine L, Kuehl M, Weir L, Leder P, Max EE. 1983. A conserved sequence in the immunoglobulin Jκ-Cκ intron: possible enhancer element. *Nature* **304:** 447–449.
The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046. doi: 10.1371/journal.pbio.1001046.
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:** 177–192.
Flint J, Tufarelli C, Peden J, Clark K, Daniels RJ, Hardison R, Miller W, Philipsen S, Tan-Un KC, McMorrow T, et al. 2001. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster. *Hum Mol Genet* **10:** 371–382.
Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. *Genome Res* **15:** 1086–1094.
Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25:** i54–i62.
Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ, Bahn S, Mistry S, Grafham D, McMurray A, Vaudin M, et al. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat Biotechnol* **18:** 181–186.
Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* **13:** 13–26.
Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19:** 68–72.
Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in transcription factor binding among humans. *Science* **328:** 232–235.
King DC, Taylor J, Zhang Y, Cheng Y, Lawson HA, Martin J, Chiaromonte F, Miller W, Hardison RC. 2007. Finding *cis*-regulatory elements using comparative genomics: some lessons from ENCODE data. *Genome Res* **17:** 775–786.
Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42:** 631–634.
Li J, Miller W. 2003. Significance of interspecies matches when evolutionary rate varies. *J Comput Biol* **10:** 537–554.
Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ 3rd, Zody MC, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.
Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. Evolutionary constraint in the human genome based on 29 eutherian mammals. *Nature* (in press).
Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci* **104:** 8005–8010.
Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2:** e5. doi: 10.1371/journal.pcbi.0020005.
Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res* **18:** 298–309.
Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* **13:** 2507–2518.
Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M, et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* **17:** 760–774.
Meader S, Ponting CP, Lunter G. 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res* **20:** 1335–1343.
Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17:** 1797–1808.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Nobrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM. 2004. Megabase deletions of gene deserts result in viable mice. *Nature* **431:** 988–993.
Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39:** 730–732.
Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324:** 389–392.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.
Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. *Genome Res* **17:** 1245–1253.
Picard D, Schaffner W. 1984. A lymphocyte-specific enhancer in the mouse immunoglobulin κ gene. *Nature* **307:** 80–82.
Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121.
Ponting CP, Goodstadt L. 2009. Separating derived from ancestral features of mouse and human genomes. *Biochem Soc Trans* **37:** 734–739.
Ponting CP, Nellaker C, Meader S. 2011. Rapid turnover of functional sequence in human and other genomes. *Annu Rev Genomics Hum Genet* (in press). doi: 10.1146/annurev-genom-090810-183115.
Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.
Sagai T, Masuya H, Tamura M, Shimizu K, Yada Y, Wakana S, Gondo Y, Noda T, Shiroishi T. 2004. Phylogenetic conservation of a limb-specific, *cis*-acting regulator of Sonic hedgehog (Shh). *Mamm Genome* **15:** 23–34.
Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328:** 1036–1040.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050.
Smith NG, Brandstrom M, Ellegren H. 2004. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84:** 806–813.
Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Valverde-Garduno V, Guyot B, Anguita E, Hamlett I, Porcher C, Vyas P. 2004. Differences in the chromatin structure and *cis*-element organization of the human and mouse GATA1 loci: implications for *cis*-element identification. *Blood* **104:** 3106–3116.
Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40:** 158–160.
Wang H, Zhang Y, Cheng Y, Zhou Y, King DC, Taylor J, Chiaromonte F, Kasturi J, Petrykowska H, Gibb B, et al. 2006. Experimental validation of predicted mammalian erythroid *cis*-regulatory modules. *Genome Res* **16:** 1480–1492.
Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci* **104:** 18613–18618.
Wold B, Myers RM. 2008. Sequence census methods for functional genomics. *Nat Methods* **5:** 19–21.
Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3:** e7. doi: 10.1371/journal.pbio.0030007.