What fraction of the human genome is functional?

Chris P. Ponting1,3 and Ross C. Hardison2,3

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

Many evolutionary studies over the past decade have estimated $\alpha_{sel}$, 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 $\alpha_{sel}$ 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 $\alpha_{sel}$ by analyses that measure constraint across the eutherian phylogeny. Approaches that take account of turnover estimate that the steady-state value of $\alpha_{sel}$ 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, $\alpha_{sel}$, of all human bases that have been evolutionarily constrained, that is, subject to purifying selection of deleterious alleles. These methods' predictions of $\alpha_{sel}$ 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 $\alpha_{sel}$, which can only be reconciled if $\alpha_{sel}$ 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.

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 $\alpha_{sel}$ 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...
The sequencing of the dog genome provided an opportunity to estimate $\alpha_{sel}$ 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 ($\alpha_{sel} \sim 5.3\%$) (Lindblad-Toh et al. 2005). Moreover, sequence that was aligned between human and dog, but not to mouse, contained little or no excess conservation ($\sim 0.1\%$). This implies that constrained sequence amounting to $\alpha_{sel} \sim 5\%$ 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 $\alpha_{sel}$ 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 $\alpha_{sel}$ 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 $T = 40–100)$. 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 $\sim 5\%$."

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 $\alpha_{sel}$ 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 $\alpha_{sel}$ 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).

**An initial estimate of $\alpha_{sel}$ from human–mouse alignments**

The first, and most renowned, estimate of $\alpha_{sel}$ 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 $\sim 5\%$."

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 $\alpha_{sel}$ 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 $\alpha_{sel}$ 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 $\alpha_{sel}$ 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 ($\alpha_{sel} \sim 5.3\%$) (Lindblad-Toh et al. 2005). Moreover, sequence that was aligned between human and dog, but not to mouse, contained little or no excess conservation ($\sim 0.1\%$). This implies that constrained sequence amounting to $\alpha_{sel} \sim 5\%$ 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.
the similarities among DNA structures inferred from hydroxyl radical cleavage patterns (Parker et al. 2009). Since 2003, 16 publications applying these methods have estimated $\alpha_{\text{sel}}$ in the human genome as being between 2.6% and 12%, with an average of $\sim 6.4\%$ (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 $\alpha_{\text{sel}}$ that never exceed 12%, which provides a strong indication that, indeed, most ($\sim 90\%$) 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 these 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 $\alpha_{\text{sel}}$ shared across many mammalian genomes, Smith et al. (2004) sought evidence for such turnover using paired alignments from $\sim 1.8$ 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 unconserved 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.

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 $\alpha'_{\text{sel}}$ 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 $\alpha_{\text{sel}}$ 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 $\alpha_{\text{sel}}$ 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 ($\pi = 1.1\%$) from $\alpha_{\text{sel}}$, the authors predicted that...
Table 1. Publications describing estimates of $\alpha_{\text{sel}}$ and $\alpha_{\text{0,sel}}$

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

(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.
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 $a_{sel} - \pi$ 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 $d \sim 0.0$). 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: $a_{sel}^{\alpha} = 14.9\%$ (Fig. 2).

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 $\kappa$ 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). About half of the functional transcription factor binding sites in human promoters are not functional in rodents (Dermietzakis 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
Regression line for the natural logarithm of \( \alpha_{\text{sel}} \) 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 \( \alpha_{\text{sel}} \) 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, \( \alpha_{\text{sel}} \) is defined relative to the size of the human genome, rather than to the sizes of different animal genomes.

Figure 3. The constrained noncoding fraction of the human genome \((\alpha_{\text{sel}} - \pi)\) declines exponentially with species divergence, \(d\). The regression line for the natural logarithm of \((\alpha_{\text{sel}} - \pi)\) 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 \( \alpha_{\text{sel}} \) 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, \( \alpha_{\text{sel}} \) is defined relative to the size of the human genome, rather than to the sizes of different animal genomes.

Figure 4. Conservation and apparent turnover of regulatory regions in the \( I G K \) gene encoding immunoglobulin \( \kappa \). 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 a_{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 a_{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 (a_{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 phenotype 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 Rand 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


What fraction of the human genome is functional?

Chris P. Ponting and Ross C. Hardison

Genome Res. 2011 21: 1769-1776 originally published online August 29, 2011
Access the most recent version at doi:10.1101/gr.116814.110