Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick[1]

ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia

While less than 1.5% of the mammalian genome encodes proteins, it is now evident that the vast majority is transcribed, mainly into non-protein-coding RNAs. This raises the question of what fraction of the genome is functional, i.e., composed of sequences that yield functional products, are required for the expression (regulation or processing) of these products, or are required for chromosome replication and maintenance. Many of the observed noncoding transcripts are differentially expressed, and, while most have not yet been studied, increasing numbers are being shown to be functional and/or trafficked to specific subcellular locations, as well as exhibit subtle evidence of selection. On the other hand, analyses of conservation patterns indicate that only ~5% (3%–8%) of the human genome is under purifying selection for functions common to mammals. However, these estimates rely on the assumption that reference sequences (usually ancient transposon-derived sequences) have evolved neutrally, which may not be the case, and if so would lead to an underestimate of the fraction of the genome under evolutionary constraint. These analyses also do not detect functional sequences that are evolving rapidly and/or have acquired lineage-specific functions. Indeed, many regulatory sequences and known functional noncoding RNAs, including many microRNAs, are not conserved over significant evolutionary distances, and recent evidence from the ENCODE project suggests that many functional elements show no detectable level of sequence constraint. Thus, it is likely that much more than 5% of the genome encodes functional information, and although the upper bound is unknown, it may be considerably higher than currently thought.

Only a tiny fraction of the human genome is currently recognized to encode functional products, mainly mRNAs (~2.2%) (Frith et al. 2005) plus a limited number of structural and regulatory RNAs, including microRNAs and other non-protein-coding RNAs (Mattick and Makunin 2006). Perplexingly, the currently estimated number of human protein-coding genes (~20,000) (International Human Genome Sequencing Consortium 2004; Goodstadt and Ponting 2006) is similar to those of the sea urchin (~23,000) (Sea Urchin Genome Sequencing Consortium 2006) and the nematode worm (~19,000) (Stein et al. 2003), and substantially less than that of the protist *Tetrahymena thermophila* (~27,000) (Eisen et al. 2006), despite enormous differences in their developmental complexity. Thus, it is unclear where the information that programs human development resides and how it is different from that of simpler organisms.

Part of the answer to this conundrum lies in the use of alternative splicing by complex organisms to expand the diversity of their proteomes (Xing and Lee 2006), although this requires a concomitant increase in regulatory information. In contrast to microorganisms, multicellular eukaryotes have extensive intronic and intergenic sequences whose extent broadly increases with developmental complexity (Taft et al. 2007). Thus it is possible that the non-protein-coding sequences in mammalian genomes contain large amounts of regulatory information used to program the complexities of mammalian development, including tetrapod body plan, placental development, and a highly developed brain, particularly in humans (Mattick 2007; Taft et al. 2007). This possibility is made all the more intriguing by the recent discovery that the vast majority of the mammalian genome is transcribed, apparently in a developmentally regulated manner (see below).

[1]Corresponding author.
E-mail j.mattick@imb.uq.edu.au; fax 61-7-3346-2111.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6406307.

However, while making some allowance for regulatory elements, and on the expectation that most genetic information is transacted by proteins, these extensive non-protein-coding sequences in humans and other mammals have been generally assumed to be nonfunctional, and mostly evolving without constraint, even though the fraction of noncoding sequences that are genetically inert is uncertain. Here we reassess the evidence concerning the amount of the human genome that is functional and under selection. We define functional sequences as those that (1) are required for replication and structural integrity of the chromosome, (2) encode functional products (RNAs and derived proteins), or (3) are required for the correct four-dimensional expression (regulation or processing) of these products during ontogeny and homeostasis. These include sequences that may act as required spacers, for example, between domains in proteins or RNAs, or in promoters, whose exact sequence may not be critical but that have a role in the functionality of the entity as a whole.

First, we review the amount and likely function of the transcriptional output of the genome. Second, bearing in mind that sequence conservation imputes function but is by definition a relative measure, we show that estimates of the extent of the genome that may be evolving "neutrally" (i.e., without obvious constraint, and by implication nonfunctional) are dependent on background assumptions of the nonfunctionality of certain classes of sequences, which may be questioned. Third, following from this, we suggest that the fraction of the genome under purifying selection may have been underestimated due to underestimation of the neutral rate of evolution. Finally, we show that experimentally validated gene regulatory sequences and functional noncoding RNAs are evolving at quite variable rates, often relatively quickly compared to sequences encoding proteins, presumably reflecting different structure–function constraints and different selection pressures. Since such sequences may not be included among those exhibiting detectable evolutionary con-

straint, and given the uncertainties in the measurement of the latter, it is possible that a considerable fraction of the human genome may be functional.

Transcriptional output of the genome

Recent cDNA and genome tiling array transcriptome analyses have revealed that at least 70% of the mammalian genome is transcribed, and possibly 60% of transcribed regions show evidence for transcription from both strands, in extremely complicated patterns of interlaced and overlapping transcripts, thousands of which are not polyadenylated (Katayama et al. 2005; Carninci 2007; Gerstein et al. 2007; Gingeras 2007). These observations have been reinforced by the recent detailed studies of the ENCODE regions of the human genome, which showed that 93% of bases in these regions appear in a primary transcript with at least two independent observations and 74% are detected by at least two different technologies (The ENCODE Project Consortium 2007). Hundreds of these intergenic, intronic, and antisense non-protein-coding transcripts show cell-specific or developmental regulation (Carninci et al. 2005; Cheng et al. 2005; Katayama et al. 2005; Ravasi et al. 2006) which may be extrapolated to thousands (Peters et al. 2007), and in the individual cases that have been examined in more detail, specific subcellular locations and functions (Prasanth et al. 2005; Willingham et al. 2005; Ginger et al. 2006; Ishii et al. 2006; for a recent review, see Mattick and Makunin 2006), all of which may indicate function. It is also now known that all snoRNAs and one-third to one-half of microRNAs in mammals are encoded within introns (Rodriguez et al. 2004; Baskerville and Bartel 2005; for review, see Mattick and Makunin 2005).

However, most of the tens of thousands of documented noncoding transcripts in mammals have not yet been studied, and it remains an open question whether they are functional or not. It has been suggested that many of these transcripts may be cell-type-specific transcriptional noise or by-products ("neutral transcription"), which may provide a reservoir for future evolution (Brosius 2005), or biochemically functional but selectively neutral transcripts with no significant advantage or disadvantage for the organism (The ENCODE Project Consortium 2007). On the other hand, recent evidence strongly implicates noncoding RNAs in the control of chromatin architecture and epigenetic memory (Andersen and Panning 2003; Bernstein and Allis 2005; Sanchez-Elsner et al. 2006; Schmitt and Paro 2006; Rinn et al. 2007), transcription (Janowski et al. 2005; Goodrich and Kugel 2006; Kim et al. 2006; Li et al. 2006; Martianov et al. 2007; Pagano et al. 2007), translation (Bartel 2004; Mattick and Makunin 2005), and possibly splicing (Mattick and Makunin 2006). Indeed, although most non-protein-coding RNAs (ncRNAs) with evidence for function are evolving quickly, they do retain more highly conserved patches within them (~600 long ncRNAs investigated in human and mouse) (Pang et al. 2006) and 3122 other long ncRNAs show subtle evidence of selection (Ponjavic et al. 2007).

Genome-wide estimates of function from conservation

Initial comparison of the mouse and human genomes led to the conclusion that ~5% of small (50–100 bp) segments are under purifying selection for biological functions common to both species (more specifically, ~20% of all human–mouse aligned segments) (Waterston et al. 2002), a surprisingly high figure at the time as only ~1.2% of the human genome is protein-coding (Frith et al. 2005). It is important to note that Waterston et al. did not claim that this was the full extent of functional sequence in the genome as it does not include lineage-specific sequences (including transposon-derived sequences) that have diverged and/or been exapted during adaptive radiation or conserved specifically since the divergence of rodents and primates. Comparative analyses of mammals that are widely separated in evolution have insufficient power to detect lineage-specific elements or elements in species that are evolutionarily "too close," such as those elements that became functional in our ancestral primate lineage (Stone et al. 2005). The initial estimate of the conserved fraction of the genome was also dependent on various parameters including the window size used for the analysis (Stone et al. 2005), and ranged from 3% to 8%. The latter corresponds to 40% of all aligned sequences, even though these alignments only included 83% of RefSeq annotated genes (Waterston et al. 2002; Chiaromonte et al. 2003; Roskin et al. 2003).

Subsequent studies seeking to identify the particular segments under selection report similar results, including the most recent finding that 5% of bases are confidently predicted as being under evolutionary constraint in mammals by two out of three algorithms employed in the ENCODE project analysis (The ENCODE Project Consortium 2007). However, since conservation is relative, all of these methods require an estimate of the underlying neutral rate of evolution, generally taken to be the substitution rate measured from some class of sequence that is expected to be evolving free of constraint, with the implicit additional assumption that there are not many functional sequences that have evolved at a net rate that is statistically indistinguishable from the estimated neutral rate (Stone et al. 2005).

Classes of sequence used to estimate the neutral rate of substitutions have included lineage-specific nonexonic sequences (Cooper et al. 2003, 2004, 2005), synonymous sites in codons (fourfold degenerate sites or 4-D sites) (Cooper et al. 2003; Margulies et al. 2003), and alignable ancestral transposon-derived sequences (ancient repeats or ARs) (Waterston et al. 2002; Chiaromonte et al. 2003; Margulies et al. 2003; Roskin et al. 2003; Gaffney and Keightley 2006), none of which is unbiased (see below). Indeed the true rate of neutral sequence drift may never have actually been measured for lack of identifying functionally completely unconstrained sectors of DNA (Zuckerkandl 1992).

Lineage-specific nonexonic sequences present in two closely related species and absent from a third more distant species have been assumed to be neutrally evolving although they will include some fraction of functionally constrained sequence (Frazer et al. 2004). Moreover, extrapolation of the measured substitution frequencies to more distantly related species is problematic and results in varying estimates of the pan-mammalian neutral rate (~1.5-fold difference; Cooper et al. 2005).

Synonymous sites in codons, often thought to be fully redundant, can apparently encode subtle additional information. The genetic code has been shown to be almost optimal to encode such additional information, such as binding sequences, splicing signals, and RNA secondary structure (Bollenbach et al. 2007; Itzkovitz and Alon 2007). Synonymous sites can encode splicing regulatory information, and a high proportion of studied mutations produce a splicing defect (Pagani et al. 2005), which is another type of constraint, and may be a frequent cause of hereditary disease (Chamary et al. 2006; Xing and Lee 2006). They can also encode protein structural information (Kimchi-Sarfaty et al. 2007; Komar 2007). These conclusions are also supported by

genome-wide evolutionary studies. The rate of synonymous substitutions is 1.8-fold lower in alternative compared to constitutive exons between human and mouse (Xing and Lee 2005). There are 200 (and up to ~1600) regions of extreme selection on synonymous codons in 11,786 pairs of homologous human and mouse genes (Schattner and Diekhans 2006). Comparison between protein-coding and intergenic regions in human and chimp indicate that ~39% of synonymous sites are deleterious and subject to negative selection (Hellmann et al. 2003). Analysis of deep mammalian alignments within ENCODE regions may detect many more regions under weaker purifying selection with greater statistical power than possible with single pairwise analyses, but this has yet to be done. However, mounting evidence for functional selection and deleterious effects of mutations suggests that the assumption of neutrality of synonymous sites can no longer be maintained, and that it is possible the neutral rate cannot reliably be extracted from any sequence comparison (Chamary et al. 2006).

Uncertainty in the estimates of selection

The original estimate of 5% of the genome under selection for functions common to mammals is largely based on estimates of the neutral rate of evolution measured from ancient repeats. However, estimates based on ARs may be biased in two ways, although the extent of such bias is unknown: (1) the annotated and aligned ARs may comprise a slowly evolving subset of the distribution of all ARs, since the most rapidly evolving ones may have diverged to the extent of being unrecognizable or unalignable, and (2) some ARs are under, or have been subject to, purifying selection. If the fraction of ARs in either category is large, then the use of ARs as a neutral model will result in a significant underestimate of the true neutral rate and hence the fraction of the genome under selection. A third possibility is that some ARs are subject to positive selection pressures and are evolving faster than the neutral rate, leading to an overestimate of the fraction of the genome under purifying selection if significant numbers have not diverged beyond recognition. In this case, however, there will be underestimation of that fraction of the genome that encodes lineage-specific functions.

The evidence supporting the possibility of bias in the estimation of the neutral rate of evolution is as follows: First, it is evident that many ARs in mammalian genomes have diverged to the limit of detection, suggesting significant numbers are beyond recognition and cannot be identified (Smit and Riggs 1995; Smit 1999; Silva et al. 2003) (numbers are difficult to estimate, but the limit of detection is ~30% divergence from the consensus and is particularly problematic in mouse; Waterston et al. 2002). The ancestral mammalian genome is estimated at ~2.8 Gb and extant ancestral sequences in human ~2.2 Gb, but only ~152 Mb of ARs are alignable with both mouse and dog (although 200 Mb is alignable with mouse and 372 Mb with dog) (Lindblad-Toh et al. 2005), and these ARs can only be traced back ~120 Myr (Waterston et al. 2002). Comparisons of alignment algorithms in ENCODE regions using sequences from 28 vertebrates including 14 mammals show that less than half of identified ARs are alignable, ranging from 24% to 47% depending on the algorithm employed (Margulies et al. 2007). These analyses also concluded that the measured substitution rate in ARs varies more between alignment algorithms than it does regionally in aligned sequences by any one alignment algorithm and that "the 'true' neutral rate for any given region of the human genome is thus only estimable given some nontrivial technical uncertainty" (Margulies et al. 2007). Thus, the large amount of ancestral sequences, particularly those that are unaligned, almost certainly includes many other AR-derived sequences that are unrecognized due their divergence (see, e.g., Mikkelsen et al. 2007), which, if so, will introduce a significant error in the estimate of the neutral rate, as only the more conserved fraction is being measured.

Second, the recent analysis of the opossum genome showed that 14% of all the most highly conserved noncoding elements (CNEs) and 16% of the eutherian-specific CNEs are derived from ARs (Mikkelsen et al. 2007). Thousands of fragments of ARs of all classes constitute at least 5.5% of the non-exonic mammalian conserved sequences and are often more highly conserved than those encoding proteins (Cooper et al. 2005; Siepel et al. 2005; Kamal et al. 2006; Lowe et al. 2007). Substitution rates are also significantly different between different classes of ARs, as well as between ARs of different age groups within a particular class (Waterston et al. 2002; Ganapathi et al. 2005; Gaffney and Keightley 2006; Pace and Feschotte 2007; Shankar et al. 2007), indicating that these sequences are evolving differently. Mammalian-wide interspersed repeats (MIRs), of which there are ~300,000 copies in the genome (2% of the genome) and date back ~130 Myr, have a lower than expected divergence from the mammalian MIR consensus, and the divergence is similar in both human and mouse even though neutrally evolving ARs should be twice as divergent in mouse than their human homologs, suggesting they are subject to selection (Silva et al. 2003). These elements have a 70-nt central region that is more highly conserved in the genome, and a 15- to 25-nt more highly conserved core within this, the most likely explanation being selection for function (Smit and Riggs 1995; Silva et al. 2003). *Alu* elements also have a core region conserved in mammals (Jelinek et al. 1980). While transposon-derived sequences (transposable elements or TEs) comprise 40%–60% of poorly conserved regions and have no identifiable ortholog, ~20% of conserved regions are composed of TEs that do have orthologs, suggesting selection of this subset. For example, MIR and L2 elements are twofold enriched in conserved regions, and >75% of murine MIR and L2 elements have human orthologs. Therefore, these elements must be ancestral repeats under negative selection, which suggests that the exaptation of MIR and L2-derived sequences may be common (Silva et al. 2003).

Evidence for functional exaptation of transposon-derived sequences

There are increasing numbers of transposon-derived sequences of all classes, both ancient and modern, including lineage-specific repeats, that have been shown to have undergone functional exaptation (Brosius 1999; Volff 2006) (also referred to as exaption, co-option, recruitment, or domestication; Silva et al. 2003). There is longstanding evidence that transposons and their derived sequences can significantly influence the information content and output of the genome (Baltimore 1985; Finnegan 1989; Oei et al. 2004). They have been shown to play important roles in early development (Peaston et al. 2004) and phenotypic variation (Whitelaw and Martin 2001). AR sequences can introduce new splice sites, protein domains, stop codons, and other sequences and can split genes, leading to the birth of new genes or alternative isoforms (Smit 1999; Lev-Maor et al. 2003; Yi et al. 2003; Dagan et al. 2004; Brandt et al. 2005a,b; Krull et al. 2005; Wheelan et al. 2005; Bejerano et al. 2006; Britten 2006; Cordaux and Batzer 2006; Cordaux et al. 2006; Zhang and Chasin 2006; Ni

et al. 2007), including noncoding RNAs (Kuryshev et al. 2001; Hasler and Strub 2006b).

AR sequences contain gene promoters (Ferrigno et al. 2001), which may be tissue-specific (Matlik et al. 2006; Romanish et al. 2007), transcription factor binding sites (Zhou et al. 2002), enhancers (Bejerano et al. 2006), silencers, polyadenylation signals, and other regulatory elements (Temin 1982; Hardman 1986), both sense and antisense (Matlik et al. 2006), which can become inserted into intergenic, intronic, protein-coding, and UTR regions (Landry et al. 2001; Smalheiser and Torvik 2006) of the genome and subsequently alter host gene expression and tissue specificity, and so the *potential* for exaptation of regulatory function is widespread around the genome (Smit 1999; Jordan et al. 2003; Shankar et al. 2004; Grover et al. 2005; Cordaux and Batzer 2006; Hasler and Strub 2006a; Polak and Domany 2006; Thornburg et al. 2006). This is not to say that the transposable elements themselves are under selection, but that sequences descended from them are (Silva et al. 2003; Lowe et al. 2007). There are RNAs derived from TEs that are developmentally modulated (Davidson and Posakony 1982), small RNAs from brain showing different strand biases (Berezikov et al. 2006a), and RNAs that undergo A-to-I editing (notably in *Alu*s) and may have important regulatory consequences (Athanasiadis et al. 2004; Blow et al. 2004; Kim et al. 2004; Levanon et al. 2004; Hasler and Strub 2006a).

Transposon-derived sequences may also underlie the creation of regulatory networks, an idea that dates back many years (Britten and Davidson 1969; Davidson and Britten 1979) and that has modern support (Zhou et al. 2002; Peaston et al. 2004; Cordaux et al. 2006; Johnson et al. 2006). Indeed, Barbara McClintock originally discovered transposable elements by studying "controlling elements" (McClintock 1956). Changes in the patterns of histone methylation in TEs in different mammalian cell types and lineages have been known for many years (Breznik et al. 1984; Nishioka 1988; Mietz and Kuff 1990; Chalitchagorn et al. 2004; Khodosevich et al. 2004; Martens et al. 2005), and they may contribute to epigenetic gene regulation (Lippman et al. 2004; Zuckerkandl and Cavalli 2007). TEs are a significant source of innovation of microRNAs (miRNAs)—at least 47 out of 545 human miRNAs are annotated as TEs (our updated analysis of Smalheiser and Torvik 2005). This suggests another mechanism for generating novel regulatory networks; any TE-derived sequence that is processed into a miRNA may be complementary to, and be able to regulate the expression of, a large number of 3′ UTRs containing similar TE-derived sequences (Smalheiser and Torvik 2006). Thus, while transposons may be mostly parasitic and TE-derived sequences may appear to have remained inert, they have contributed to the evolution of mammalian genomes through many mechanisms that create and modify gene expression and regulatory networks.

Different rates of evolution of functional sequences

It is also clear that there are widely different rates of evolution of different types of functional sequences in mammals. Rapidly changing sequences may be interpreted as neutrally evolving and nonfunctional, as functionally important but having flexible structure–function relationships, or as functionally important and undergoing adaptive improvements by acquiring advantageous mutations (Zuckerkandl 1992). Innovation in protein-coding sequences, which are usually governed by quite strict analog structure–function constraints, appears to be rare, whereas ~20% of eutherian conserved non-protein-coding elements (CNEs) are recent innovations that postdate the divergence of eutheria and metatheria (Mikkelsen et al. 2007).

Innovation and rapid evolution is also evident in thousands of gene regulatory sequences, which cover extended genomic regions and exhibit rapid turnover (Smith et al. 2004; Fisher et al. 2006; Frith et al. 2006; Taylor et al. 2006). This includes the remarkable functional conservation of regulatory sequences controlling *ret* gene expression in zebrafish and humans, although there is little recognizable primary sequence conservation (Fisher et al. 2006), and the independent exaptation of ARs as regulators of orthologous genes in human and rodents (Romanish et al. 2007). Taking turnover into account, it has been estimated that the extent of functional sequences in the human genome may be twice as great as that estimated from sequence conservation alone (Smith et al. 2004). Highly conserved epigenetic modifications can be used to identify tens of thousands of important regulatory elements, which cannot be identified by sequence conservation alone, half of which are lineage-specific (Roh et al. 2007). There are ~1000 regions of the human genome over 10 kb long that do not tolerate transposable element insertions, even though primary sequence is not highly conserved (Simons et al. 2006). Gene deserts are large regions covering >700 Mb of the human genome, which appear to harbor distant regulatory elements and are devoid of protein-coding genes and that contain rapidly evolving regions that apparently accept neutral substitutions at a higher rate than the bulk of the genome yet resist chromosomal rearrangements, suggesting they are subject to evolutionary constraints, which are not readily apparent in primary sequence, against harboring genes (Ovcharenko et al. 2005). There are other regions of the genome that show evolutionary constraint that is not evident at the primary sequence level, including shuffled *cis*-regulatory elements (Sanges et al. 2006), regions subject to heterogeneous selection, which are evolving rapidly in primary sequence but slowly with respect to indels (Lunter et al. 2006), the distances between ultra-conserved elements (Sun et al. 2006), and regions predicted to contain common RNA secondary structure (Washietl et al. 2005) or highly constrained RNA tertiary structures that may have weak constraints on primary sequence or cryptic patterns of non-Watson–Crick base pair conservation (Lescoute et al. 2005).

Different rates of evolution also occur both within and between different classes of functional gene products, both RNAs and proteins. While the majority of protein-coding sequences are highly constrained, some are much more flexible, or under positive selection (Bustamante et al. 2005). As Kimura (1968) originally pointed out, many substitutions in protein-coding sequences appear to be neutral or nearly neutral, but this does not mean that the segments in which they reside are nonfunctional, simply that they are relatively plastic. In addition, Zuckerkandl (1992) notes that Kimura's selectively neutral mutations are selectively equivalent and thus do not preclude them being functional. The first few hundred miRNAs to be discovered are highly conserved (Pang et al. 2006), but hundreds of more recently discovered miRNAs are not, being lineage- or even species-specific (Berezikov et al. 2006a,b; Piriyapongsa and Jordan 2007; Zhang et al. 2007) and expanding in the mammalian lineage (Hertel et al. 2006). There are also thousands of recently discovered small RNAs (piRNAs) expressed in testis that are not conserved between mouse and other species (Aravin et al. 2006; Girard et al. 2006; Lau et al. 2006).

As mentioned above, hundreds of longer ncRNAs, including the *Xist* and *Tsix* transcripts involved in X chromosome dosage

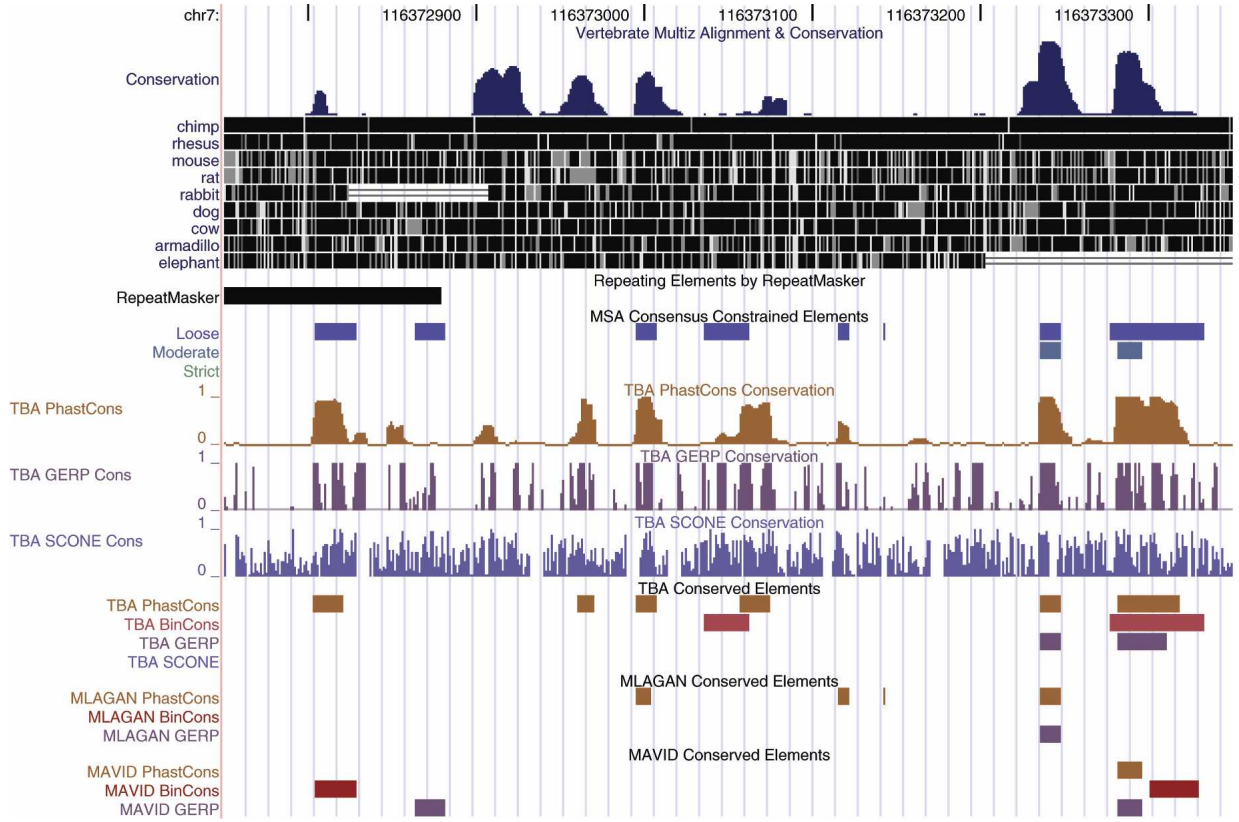

Figure 1. Conservation in the ENCODE *CFTR* locus. The diagram shows a 600-bp region in an intron of the *ST7* gene (hg17 chr7:116372751–116373350). The *top* panel ("Vertebrate Multiz Alignment & Conservation") shows phastCons conservation scores based on 17-way alignments (Siepel et al. 2005). In black *below* this are alignments of human with chimp, rhesus, mouse, rat, rabbit, dog, cow, armadillo, and elephant. "Repeating Elements by RepeatMasker" shows an ancient repeat annotated as a MIR, which is 27% divergent from the MIR consensus, near the limit of detection. "MSA Consensus Constrained Elements" shows eight regions predicted to be conserved by at least one algorithm ("Loose" set), two regions predicted to be conserved by at least two algorithms in at least two alignments ("Moderate" set), and no regions predicted to be conserved by all algorithms in all alignments ("Strict" set). "TBA phastCons Conservation," "TBA GERP Conservation," and "TBA SCONE Conservation" show conservation scores over the TBA alignment from phastCons, GERP, and SCONE algorithms, respectively. "TBA Conserved Elements," "MLAGAN Conserved Elements," and "MAVID Conserved Elements" show elements predicted conserved based on the scores from the phastCons, BinCons, GERP, and SCONE algorithms across alignments from TBA, MLAGAN, and MAVID, respectively (Margulies et al. 2007) (image from http://genome.ucsc.edu/). The figure illustrates several difficulties in identifying selective constraints from regions that are not highly conserved: (1) conserved blocks are predicted within ARs assumed to evolve neutrally; (2) conservation scores vary depending on the species aligned (phastCons scores in the *top* panel are different from scores in TBA phastCons scores); (3) patterns of identified conservation vary between algorithms over the same alignment (compare the pattern of TBA scores from phastCons, GERP, and SCONE); and (4) conserved element predictions based on these scores vary between different algorithms on the same alignment as well as between the same algorithm over different alignments (compare phastCons, BinCons, and GERP elements over TBA, MLAGAN, and MAVID alignments).

compensation, are evolving quickly (Nesterova et al. 2001; Pang et al. 2006). A recent study of 3122 mouse long ncRNAs with weak evidence for purifying selection on their primary sequences nonetheless showed clear evidence for selection when their promoters, indel distribution, and conserved splice sites were considered (Ponjavic et al. 2007). There is also evidence of recent positive selection of ncRNAs in human, such as the *HAR1* transcript expressed in particular regions of the brain (Pollard et al. 2006). Although functionally validated RNAs do not presently add up to a large fraction of the genome, they do (1) illustrate the point that low conservation of the primary sequence does not necessarily equate to or demonstrate lack or loss of function (Zuckerkandl 1992; Smith et al. 2004; Xing and Lee 2005; Pang et al. 2006) and (2) point to the possibility that many functional transcripts, particularly regulatory ncRNAs, may not be highly conserved over significant evolutionary distances, presumably because of more relaxed structure–function constraints and/or positive selection for regulatory variants associated with phenotypic radiation and adaptive evolution.

Consistent with this, the recent analysis of the ENCODE regions concluded that "many functional elements are seemingly unconstrained across mammalian evolution" (The ENCODE Project Consortium 2007). This has been interpreted to indicate that there may be many sequences that are "biologically active but provide no specific benefit to the organism" (The ENCODE Project Consortium 2007). However, this apparent contradiction can be readily resolved if the actual neutral rate of evolution is higher than current estimations. These observations are also consistent with the possibility that many of these apparently weakly constrained sequences encode lineage-specific functional elements and/or functionally similar but nonorthologous elements that have been subject to rapid drift. The problems with detecting which sequences, and in determining the extent of sequences, in the genome that may be under evolutionary con-

straint, particularly in regions that are not highly conserved, is exemplified by Figure 1, which shows a close-up view of a region within an intron of the *ST7* gene in the ENCODE *CFTR* region and illustrates several difficulties in identifying selective constraints from regions that are not highly conserved.

A common objection to the possibility that mammalian genomes may contain large amounts of functional sequence under weak selection is the prediction that only strongly advantageous or disadvantageous alleles are subject to selection in mammals due to their small effective population sizes, and thus alleles that have a small functional impact evolve neutrally. This objection is apparently contradicted by the "unexpected strength of natural selection" in synonymous sites discussed in Chamary et al. (2006). In addition, Zuckerkandl (1992) points out that functionality in the more rapidly evolving noncoding regions of the genome cannot be negated on the basis of other observations that support both neutralist and alternative interpretations.

How much of the genome might be functional?

The assumption that recognizable ARs are nonfunctional and are representative leads to the conservative estimate that 3%–8% of genomic regions are under purifying selection in mammals. However, it is clear that all estimates of the extent of neutrally evolving segments of the human genome, and reciprocally of those under selection and imputed to be functional, are entirely dependent both qualitatively and quantitatively on the assumption of the neutral evolution of extant ARs, which may or may not be correct, but which is at least subject to doubt. Evidence continues to mount that AR-derived sequences can modify genetic output, and that both individual ARs and classes of ARs are evolving non-neutrally. There may also be significant underrepresentation of faster-evolving unrecognized or unaligned ARs, with the consequence that the extent of purifying selection in mammals, and hence the proportion of functional sequences, may be significantly underestimated. Moreover, there are significant discrepancies and difficulties in estimating the presumed neutral rate (Margulies et al. 2007), all of which are dependent on the underlying assumptions and parameters and which may be interpreted differently. Unfortunately, however, the available data in large part do not permit distinction between, nor assessment of the extent of, sequences that may be inert and evolving without constraint versus those that are functional and evolving at different rates under different structure–function constraints and different selection pressures, with different evolutionary histories, especially those involved in gene regulation. It therefore remains an open question whether the majority of the genome is evolving neutrally and whether it may be functional or not. A recent study has shown that a substantial fraction of purifying selection in human noncoding sequences occurs outside of previously identified conserved noncoding sequences and is diffusely distributed across the genome. This finding suggests that there are many human noncoding variants that may impact gene expression and phenotypic traits, most of which will have escaped detection with current approaches to genome analysis (Asthana et al. 2007).

It seems clear that 5% is a minimum estimate of the fraction of the human genome that is functional, and that the true extent is likely to be significantly greater. If the upper figure of 11.8% under common purifying selection in mammals from ENCODE (Margulies et al. 2007) is realistic across the genome as a whole, and if turnover and positive selection approximately doubles this figure (Smith et al. 2004), then the functional portion of the genome may exceed 20%. It is also now clear that the majority of the mammalian genome is expressed and that many mammalian genes are accompanied by extensive regulatory regions. Thus, although admittedly on the basis of as yet limited evidence, it is quite plausible that many, if not the majority, of the expressed transcripts are functional and that a major component of genomic information is rapidly evolving regulatory DNA and RNA. Consequently, it is possible that much if not most of the human genome may be functional. This possibility cannot be ruled out on the available evidence, either from conservation analysis or from genetic studies (Mattick and Makunin 2006), but does challenge current conceptions of the extent of functionality of the human genome and the nature of the genetic programming of humans and other complex organisms.

Acknowledgments

We thank Cas Simons, Igor Makunin, Evgeny Glazov, and Chris Ponting for their advice and comments on the manuscript. We also thank the reviewers and the editor for constructive criticisms and helpful suggestions. We acknowledge the financial support of the Australian Research Council, the Queensland State Government, and the University of Queensland.

References

Andersen, A.A. and Panning, B. 2003. Epigenetic gene regulation by noncoding RNAs. *Curr. Opin. Cell Biol.* **15:** 281–289.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442:** 203–207.

Asthana, S., Noble, W.S., Kryukov, G., Grant, C.E., Sunyaev, S., and Stamatoyannopoulos, J.A. 2007. Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci.* **104:** 12410–12415.

Athanasiadis, A., Rich, A., and Maas, S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2:** e391. doi: 10.1371/journal.pbio.0020391.

Baltimore, D. 1985. Retroviruses and retrotransposons: The role of reverse transcription in shaping the eukaryotic genome. *Cell* **40:** 481–482.

Bartel, D. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Baskerville, S. and Bartel, D.P. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241–247.

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

Berezikov, E., Thuemmler, F., van Laake, L., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R.H.A. 2006a. Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38:** 1375–1377.

Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S., et al. 2006b. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.* **16:** 1289–1298.

Bernstein, E. and Allis, C.D. 2005. RNA meets chromatin. *Genes & Dev.* **19:** 1635–1655.

Blow, M., Futreal, P.A., Wooster, R., and Stratton, M.R. 2004. A survey of RNA editing in human brain. *Genome Res.* **14:** 2379–2387.

Bollenbach, T., Vetsigian, K., and Kishony, R. 2007. Evolution and multilevel optimization of the genetic code. *Genome Res.* **17:** 401–404.

Brandt, J., Schrauth, S., Veith, A.M., Froschauer, A., Haneke, T., Schultheis, C., Gessler, M., Leimeister, C., and Volff, J.N. 2005a. Transposable elements as a source of genetic innovation: Expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* **345:** 101–111.

Brandt, J., Veith, A.M., and Volff, J.N. 2005b. A family of neofunctionalized Ty3/gypsy retrotransposon genes in mammalian genomes. *Cytogenet. Genome Res.* **110:** 307–317.
Breznik, T., Traina-Dorge, V., Gama-Sosa, M., Gehrke, C.W., Ehrlich, M., Medina, D., Butel, J.S., and Cohen, J.C. 1984. Mouse mammary tumor virus DNA methylation: Tissue-specific variation. *Virology* **136:** 69–77.
Britten, R. 2006. Transposable elements have contributed to thousands of human proteins. *Proc. Natl. Acad. Sci.* **103:** 1798–1803.
Britten, R.J. and Davidson, E.H. 1969. Gene regulation for higher cells: A theory. *Science* **165:** 349–357.
Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238:** 115–134.
Brosius, J. 2005. Waste not, want not—Transcript excess in multicellular eukaryotes. *Trends Genet.* **21:** 287–288.
Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.
Carninci, P. 2007. Constructing the landscape of the mammalian transcriptome. *J. Exp. Biol.* **210:** 1497–1506.
Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.
Chalitchagorn, K., Shuangshoti, S., Hourpai, N., Kongruttanachok, N., Tangkijvanich, P., Thong-ngam, D., Voravud, N., Sriuranpong, V., and Mutirangura, A. 2004. Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene* **23:** 8841–8846.
Chamary, J.V., Parmley, J.L., and Hurst, L.D. 2006. Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7:** 98–108.
Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308:** 1149–1154.
Chiaromonte, F., Weber, R.J., Roskin, K.M., Diekhans, M., Kent, W.J., and 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.
Cooper, G.M., Brudno, M., Green, E.D., Batzoglou, S., Sidow, A., and NISC Comparative Sequencing Program. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13:** 813–820.
Cooper, G.M., Brudno, M., Stone, E.A., Dubchak, I., Batzoglou, S., and Sidow, A. 2004. Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.* **14:** 539–548.
Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15:** 901–913.
Cordaux, R. and Batzer, M.A. 2006. Teaching an old dog new tricks: SINEs of canine genomic diversity. *Proc. Natl. Acad. Sci.* **103:** 1157–1158.
Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci.* **103:** 8101–8106.
Dagan, T., Sorek, R., Sharon, E., Ast, G., and Graur, D. 2004. AluGene: A database of *Alu* elements incorporated within protein-coding genes. *Nucleic Acids Res.* **32:** D489–D492. doi: 10.1093/nar/gkh132.
Davidson, E.H. and Britten, R.J. 1979. Regulation of gene expression: Possible role of repetitive sequences. *Science* **204:** 1052–1059.
Davidson, E.H. and Posakony, J.W. 1982. Repetitive sequence transcripts in development. *Nature* **297:** 633–635.
Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **4:** e286. doi: 10.1371/journal.pbio.0040286.
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.
Ferrigno, O., Virolle, T., Djabari, Z., Ortonne, J.P., White, R.J., and Aberdam, D. 2001. Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* **28:** 77–81.
Finnegan, D.J. 1989. Eukaryotic transposable elements and genome evolution. *Trends Genet.* **5:** 103–107.
Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312:** 276–279.
Frazer, K.A., Tao, H., Osoegawa, K., de Jong, P.J., Chen, X., Doherty, M.F., and Cox, D.R. 2004. Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14:** 367–372.
Frith, M.C., Pheasant, M., and Mattick, J.S. 2005. The amazing complexity of the human transcriptome. *Eur. J. Hum. Genet.* **13:** 894–897.
Frith, M.C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayshizaki, Y., and Sandelin, A. 2006. Evolutionary turnover of mammalian transcription start sites. *Genome Res.* **16:** 713–722.
Gaffney, D.J. and Keightley, P.D. 2006. Genomic selective constraints in murid noncoding DNA. *PLoS Genet.* **2:** e204. doi: 10.1371/journal.pgen.0020204.
Ganapathi, M., Srivastava, P., Das Sutar, S.K., Kumar, K., Dasgupta, D., Pal Singh, G., Brahmachari, V., and Brahmachari, S.K. 2005. Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. *BMC Bioinformatics* **6:** 126. doi: 10.1186/1471-2105-6-126.
Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S., and Snyder, M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res.* **17:** 669–681.
Ginger, M.R., Shore, A.N., Contreras, A., Rijnkels, M., Miller, J., Gonzalez-Rimbau, M.F., and Rosen, J.M. 2006. A noncoding RNA is a potential marker of cell fate during mammary gland development. *Proc. Natl. Acad. Sci.* **103:** 5781–5786.
Gingeras, T.R. 2007. Origin of phenotypes: Genes and transcripts. *Genome Res.* **17:** 682–690.
Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199–202.
Goodrich, J.A. and Kugel, J.F. 2006. Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* **7:** 612–616.
Goodstadt, L. and Ponting, C.P. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comp. Biol.* **2:** e133. doi: 10.1371/journal.pcbi.0020133.
Grover, D., Kannan, K., Brahmachari, S.K., and Mukerji, M. 2005. ALU-ring elements in the primate genomes. *Genetica* **124:** 273–289.
Hardman, N. 1986. Structure and function of repetitive DNA in eukaryotes. *Biochem. J.* **234:** 1–11.
Hasler, J. and Strub, K. 2006a. *Alu* elements as regulators of gene expression. *Nucleic Acids Res.* **34:** 5491–5497. doi: 10.1093/nar/gkl706.
Hasler, J. and Strub, K. 2006b. *Alu* RNP and *Alu* RNA regulate translation initiation in vitro. *Nucleic Acids Res.* **34:** 2374–2385. doi: 10.1093/nar/gkl246.
Hellmann, I., Zollner, S., Enard, W., Ebersberger, I., Nickel, B., and Paabo, S. 2003. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res.* **13:** 831–837.
Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I.L., and Stadler, P.F. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7:** 25. doi: 10.1186/1471-2164-7-25.
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.
Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M., et al. 2006. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *J. Hum. Genet.* **51:** 1087–1099.
Itzkovitz, S. and Alon, U. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Res.* **17:** 405–412.
Janowski, B.A., Huffman, K.E., Schwartz, J.C., Ram, R., Hardy, D., Shames, D.S., Minna, J.D., and Corey, D.R. 2005. Inhibiting gene expression at transcription start sites in chromosomal DNA with antigene RNAs. *Nat. Chem. Biol.* **1:** 216–222.
Jelinek, W.R., Toomey, T.P., Leinwand, L., Duncan, C.H., Biro, P.A., Choudary, P.V., Weissman, S.M., Rubin, C.M., Houck, C.M., Deininger, P.L., et al. 1980. Ubiquitous, interspersed repeated sequences in mammalian genomes. *Proc. Natl. Acad. Sci.* **77:** 1398–1402.
Johnson, R., Gamblin, R.J., Ooi, L., Bruce, A.W., Donaldson, I.J., Westhead, D.R., Wood, I.C., Jackson, R.M., and Buckley, N.J. 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.* **34:** 3862–3877. doi: 10.1093/nar/gkl525.
Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19:** 68–72.
Kamal, M., Xie, X., and Lander, E.S. 2006. A large family of ancient

repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci.* **103:** 2740–2745.
Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al. 2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564–1566.
Khodosevich, K., Lebedev, Y., and Sverdlov, E.D. 2004. Large-scale determination of the methylation status of retrotransposons in different tissues using a methylation tags approach. *Nucleic Acids Res.* **32:** e31. doi: 10.1093/nar/gnh035.
Kim, D.D., Kim, T.T., Walsh, T., Kobayashi, Y., Matise, T.C., Buyske, S., and Gabriel, A. 2004. Widespread RNA editing of embedded *Alu* elements in the human transcriptome. *Genome Res.* **14:** 1719–1725.
Kim, D.H., Villeneuve, L.M., Morris, K.V., and Rossi, J.J. 2006. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nat. Struct. Mol. Biol.* **13:** 793–797.
Kimchi-Sarfaty, C., Oh, J.M., Kim, I., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. 2007. A "silent" polymorphism in the *MDR1* gene changes substrate specificity. *Science* **315:** 525–528.
Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217:** 624–626.
Komar, A.A. 2007. SNPs, silent but not invisible. *Science* **315:** 466–467.
Krull, M., Brosius, J., and Schmitz, J. 2005. *Alu*-SINE exonization: En route to protein-coding function. *Mol. Biol. Evol.* **22:** 1702–1711.
Kuryshev, V.Y., Skryabin, B.V., Kremerskothen, J., Jurka, J., and Brosius, J. 2001. Birth of a gene: Locus of neuronal BC200 snmRNA in three prosimians and human BC200 pseudogenes as archives of change in the Anthropoidea lineage. *J. Mol. Biol.* **309:** 1049–1066.
Landry, J.R., Medstrand, P., and Mager, D.L. 2001. Repetitive elements in the 5′ untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* **76:** 110–116.
Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363–367.
Lescoute, A., Leontis, N.B., Massire, C., and Westhof, E. 2005. Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Res.* **33:** 2395–2409. doi: 10.1093/nar/gki535.
Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22:** 1001–1005.
Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. 2003. The birth of an alternatively spliced exon: 3′ Splice-site selection in *Alu* exons. *Science* **300:** 1288–1291.
Li, L.C., Okino, S.T., Zhao, H., Pookot, D., Place, R.F., Urakami, S., Enokida, H., and Dahiya, R. 2006. Small dsRNAs induce transcriptional activation in human cells. *Proc. Natl. Acad. Sci.* **103:** 17337–17342.
Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas 3rd, E.J., Zody, M.C., et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438:** 803–819.
Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471–476.
Lowe, C.B., Bejerano, G., and Haussler, D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci.* **104:** 8005–8010.
Lunter, G., Ponting, C.P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comp. Biol.* **2:** e5. doi: 10.1371/journal.pcbi.0020005.
Margulies, E.H., Blanchette, M., Haussler, D., Green, E.D., and NISC Comparative Sequencing Program. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* **13:** 2507–2518.
Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., 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.
Martens, J.H., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. 2005. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *EMBO J.* **24:** 800–812.
Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., and Akoulitchev, A. 2007. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature* **445:** 666–670.
Matlik, K., Redik, K., and Speek, M. 2006. L1 antisense promoter drives tissue-specific transcription of human genes. *J. Biomed. Biotechnol.* **2006:** 71753.
Mattick, J.S. 2007. A new paradigm for developmental biology. *J. Exp. Biol.* **210:** 1526–1547.
Mattick, J.S. and Makunin, I.V. 2005. Small regulatory RNAs in mammals. *Hum. Mol. Genet.* **14:** R121–R132.
Mattick, J.S. and Makunin, I.V. 2006. Non-coding RNA. *Hum. Mol. Genet.* **15:** R17–R29.
McClintock, B. 1956. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21:** 197–216.
Mietz, J.A. and Kuff, E.L. 1990. Tissue and strain-specific patterns of endogenous proviral hypomethylation analyzed by two-dimensional gel electrophoresis. *Proc. Natl. Acad. Sci.* **87:** 2269–2273.
Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.
Nesterova, T.B., Slobodyanyuk, S.Y., Elisaphenko, E.A., Shevchenko, A.I., Johnston, C., Pavlova, M.E., Rogozin, I.B., Kolesnikov, N.N., Brockdorff, N., and Zakian, S.M. 2001. Characterization of the genomic *Xist* locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res.* **11:** 833–849.
Ni, J.Z., Grate, L., Donohue, J.P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T.A., Blume, J.E., and Ares Jr., M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes & Dev.* **21:** 708–718.
Nishioka, Y. 1988. Tissue specific methylation of human Y chromosomal DNA sequences. *Tissue Cell* **20:** 875–880.
Oei, S.L., Babich, V.S., Kazakov, V.I., Usmanova, N.M., Kropotov, A.V., and Tomilin, N.V. 2004. Clusters of regulatory signals for RNA polymerase II transcription associated with *Alu* family repeats and CpG islands in human promoters. *Genomics* **83:** 873–882.
Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. 2005. Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15:** 137–145.
Pace II, J.K. and Feschotte, C. 2007. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res.* **17:** 422–432.
Pagani, F., Raponi, M., and Baralle, F.E. 2005. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci.* **102:** 6368–6372.
Pagano, A., Castelnuovo, M., Tortelli, F., Ferrari, R., Dieci, G., and Cancedda, R. 2007. New small nuclear RNA gene-like transcriptional units as sources of regulatory transcripts. *PLoS Genet.* **3:** e1. doi: 10.1371/journal.pgen.0030001.
Pang, K.C., Frith, M.C., and Mattick, J.S. 2006. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends Genet.* **22:** 1–5.
Peaston, A.E., Evsikov, A.V., Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev. Cell* **7:** 597–606.
Peters, B.A., St. Croix, B., Sjöblom, T., Cummins, J.M., Silliman, N., Ptak, J., Saha, S., Kinzler, K.W., Hatzis, C., and Velculescu, V.E. 2007. Large-scale identification of novel transcripts in the human genome. *Genome Res.* **17:** 287–292.
Piriyapongsa, J. and Jordan, I.K. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* **2:** e203. doi: 10.1371/journal.pone.0000203.
Polak, P. and Domany, E. 2006. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7:** 133. doi: 10.1186/1471-2164-7-133.
Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.
Ponjavic, J., Ponting, C.P., and Lunter, G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.* **17:** 556–565.
Prasanth, K.V., Prasanth, S.G., Xuan, Z., Hearn, S., Freier, S.M., Bennett, C.F., Zhang, M.Q., and Spector, D.L. 2005. Regulating gene expression through RNA nuclear retention. *Cell* **123:** 249–263.
Ravasi, T., Suzuki, H., Pang, K.C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M.C., Gongora, M.M., et al. 2006. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **16:** 11–19.
Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. 2007.

Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311–1323.
Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L., and Bradley, A. 2004. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **14:** 1902–1910.
Roh, T.Y., Wei, G., Farrell, C.M., and Zhao, K. 2007. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.* **17:** 74–81.
Romanish, M.T., Lock, W.M., de Lagemaat, L.N., Dunn, C.A., and Mager, D.L. 2007. Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet.* **3:** e10. doi: 10.1371/journal.pgen.0030010.
Roskin, K.M., Diekhans, M., and Haussler, D. 2003. Scoring two-species local alignments to try to statistically separate neutrally evolving from selected DNA segments. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology* (eds. K.M. Roskin et al.), pp. 257–266. ACM Press, New York.
Sanchez-Elsner, T., Gou, D., Kremmer, E., and Sauer, F. 2006. Noncoding RNAs of trithorax response elements recruit *Drosophila* Ash1 to Ultrabithorax. *Science* **311:** 1118–1123.
Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F., and Stupka, E. 2006. Shuffling of *cis*-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.* **7:** R56. doi: 10.1186/gb-2006-7-7-r56.
Schattner, P. and Diekhans, M. 2006. Regions of extreme synonymous codon selection in mammalian genes. *Nucleic Acids Res.* **34:** 1700–1710. doi: 10.1093/nar/gkl095.
Schmitt, S. and Paro, R. 2006. RNA at the steering wheel. *Genome Biol.* **7:** 218. doi: 10.1186/gb-2006-7-5-218.
Sea Urchin Genome Sequencing Consortium. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314:** 941–952.
Shankar, R., Grover, D., Brahmachari, S.K., and Mukerji, M. 2004. Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. *BMC Evol. Biol.* **4:** 37. doi: 10.1186/1471-2148-4-37.
Shankar, R., Chaurasia, A., Ghosh, B., Chekmenev, D., Cheremushkin, E., Kel, A., and Mukerji, M. 2007. Non-random genomic divergence in repetitive sequences of human and chimpanzee in genes of different functional categories. *Mol. Genet. Genomics* **277:** 441–455.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Silva, J.C., Shabalina, S.A., Harris, D.G., Spouge, J.L., and Kondrashov, A.S. 2003. Conserved fragments of transposable elements in intergenic regions: Evidence for widespread recruitment of MIR- and L2-derived sequences within the mouse and human genomes. *Genet. Res.* **82:** 1–18.
Simons, C., Pheasant, M., Makunin, I.V., and Mattick, J.S. 2006. Transposon-free regions in mammalian genomes. *Genome Res.* **16:** 164–172.
Smalheiser, N.R. and Torvik, V.I. 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet.* **21:** 322–326.
Smalheiser, N.R. and Torvik, V.I. 2006. Alu elements within human mRNAs are probable microRNA targets. *Trends Genet.* **22:** 532–536.
Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.
Smit, A.F. and Riggs, A.D. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23:** 98–102.
Smith, N.G., Brandstrom, M., and Ellegren, H. 2004. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* **84:** 806–813.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1:** e45. doi: 10.1371/journal.pbio.0000045.
Stone, E.A., Cooper, G.M., and Sidow, A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **6:** 143–164.
Sun, H., Skogerbo, G., and Chen, R. 2006. Conserved distances between vertebrate highly conserved elements. *Hum. Mol. Genet.* **15:** 2911–2922.
Taft, R.J., Pheasant, M., and Mattick, J.S. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29:** 288–299.
Taylor, M.S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Semple, C.A. 2006. Heterotachy in mammalian promoter evolution. *PLoS Genet.* **2:** e30. doi: 10.1371/journal.pgen.0020030.
Temin, H.M. 1982. Function of the retrovirus long terminal repeat. *Cell* **28:** 3–5.
Thornburg, B.G., Gotea, V., and Makalowski, W. 2006. Transposable elements as a significant source of transcription regulating signals. *Gene* **365:** 104–110.
Volff, J.N. 2006. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28:** 913–922.
Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A., and Stadler, P.F. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23:** 1383–1390.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. 2005. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. *Genome Res.* **15:** 1073–1078.
Whitelaw, E. and Martin, D.I. 2001. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat. Genet.* **27:** 361–365.
Willingham, A.T., Orth, A.P., Batalov, S., Peters, E.C., Wen, B.G., Aza-Blanc, P., Hogenesch, J.B., and Schultz, P.G. 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309:** 1570–1573.
Xing, Y. and Lee, C. 2005. Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc. Natl. Acad. Sci.* **102:** 13526–13531.
Xing, Y. and Lee, C. 2006. Alternative splicing and RNA selection pressure—Evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.* **7:** 499–509.
Yi, P., Zhang, W., Zhai, Z., Miao, L., Wang, Y., and Wu, M. 2003. Bcl-rambo beta, a special splicing variant with an insertion of an Alu-like cassette, promotes etoposide- and taxol-induced cell death. *FEBS Lett.* **534:** 61–68.
Zhang, X.H. and Chasin, L.A. 2006. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc. Natl. Acad. Sci.* **103:** 13427–13432.
Zhang, R., Peng, Y., Wang, W., and Su, B. 2007. Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res.* **17:** 612–617.
Zhou, Y.H., Zheng, J.B., Gu, X., Saunders, G.F., and Yung, W.K. 2002. Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Res.* **12:** 1716–1722.
Zuckerkandl, E. 1992. Revisiting junk DNA. *J. Mol. Evol.* **34:** 259–271.
Zuckerkandl, E. and Cavalli, G. 2007. Combinatorial epigenetics, "junk DNA," and the evolution of complex organisms. *Gene* **390:** 232–242.

Received February 17, 2007; accepted in revised form July 12, 2007.