Unexpected conserved non-coding DNA blocks in mammals

Daniel J. Gaffney and Peter D. Keightley

Ashworth Laboratories, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom

The significance of non-coding DNA is a longstanding riddle in the study of molecular evolution. Using a comparative genomics approach, Dermitzakis and colleagues have recently shown that at least some non-coding sequence, frequently ignored as meaningless, might bear the signature of natural selection. If functional, it could mark a turning point in the way we think about the evolution of the genome.

Few genomic features are more puzzling than the vast amounts of apparently functionless non-coding DNA that make up the greater proportion of human, mouse and many other eukaryotic genomes. However, although the view of non-coding sequence as genomic debris has been widespread, recent results by Dermitzakis and colleagues [1–3] offers a fascinating hint that a significant proportion can retain a function that, for the moment, remains a mystery.

For much of the past 50 years, the functional genome has been viewed as one that codes for protein and, until recently, most evolutionary studies of DNA sequences have focused almost entirely on this translated fraction, which we now think accounts for as little as 1–2% of both human and mouse DNA [4,5]. Many theories of the origin of non-coding DNA are founded on the perception that the bulk of such sequence is meaningless [6] and invoke random processes of accumulation of this ‘junk’, for example, the action of ‘selfish’ self-replicating elements [7]. Whole genome sequencing has, to some extent, borne these views out. Approximately 40% of mouse and human genomes are composed of the repetitive signatures that characterize past insertion of such retroelements [4,5]. Indeed, ~20% of the entire mouse genome appears to have originated via the activity of a single class of element, the long interspersed elements (LINEs) [5]. However, excluding repetitive DNA sequence still leaves enormous quantities of non-coding sequence that we know little about. One of the most intriguing suggestions arising from the comparison of human and mouse genomes is that protein-coding sequences only account for approximately a fifth of the total amount of each species’ genome that is subject to purifying selection [5]. The implication is that relatively large amounts of non-coding DNA are functional and it is clear, therefore, that the elucidation of potential functions (or otherwise) of non-coding DNA is a primary challenge in evolutionary genomics.
Development of comparative analyses

One powerful approach to address this challenge implements cross-species comparison of syntenic genomic regions with the aim of identifying potentially functional sequences. The underlying assumption, rooted in the neutral theory, is that conservation, above that expected given phylogenetic distance, implies selective constraint and, thus, function. Such ‘phylogenetic footprinting’ is not a new concept. Early comparative studies [8] were expanded by Duret et al. [9], who uncovered surprisingly strong conservation of flanking and untranslated regions (UTRs) of orthologous genes among widely diverged vertebrate groups. A similar approach on a larger scale by Koop et al. [10], comparing 100 kb of syntenic sequence surrounding T-cell receptor loci in mouse and human, revealed that conserved non-coding DNA could be found deep within intergenic regions in addition to sequence that is proximate to genes.

More recent human–mouse comparisons have revealed mosaics of constrained and randomly drifting sequence in intronic [11] and intergenic [12] DNA. The conserved blocks of non-coding sequence of which such mosaics consist have, in many cases, been found to correspond to exonic or regulatory regions [13,14]. Such pairwise comparisons are useful but can detect only a fraction of extant conservation and, as a result, multi-species phylogenies are becoming more commonly used. Analysis of sequence orthologous to a region of human chromosome 7 across 12 species by Thomas et al. [15] demonstrated the limitations of a single species pair in detecting the true pattern of non-coding conservation. High-density oligonucleotide arrays have also been employed recently in cross-species comparisons over entire chromosomes [16], revealing large quantities of conserved non-coding blocks that are not related to known exons. It is clear that large-scale comparative sequence analysis is rapidly becoming a useful tool to identify functionality within non-coding sequence [17,18].

Conserved non-genic sequences: functional non-coding DNA?

The work of Dermitzakis and colleagues [1] supports these previous findings and suggests that extensive conservation of non-coding DNA is not just a feature of a few mammalian species. Their initial study compared 33.5 Mb from the long arm of human chromosome 21 with syntenic sequence in mouse. The alignment of these regions revealed a high frequency of well-conserved, ungapped sequences located primarily in the Giemsa-dark, gene-poor region of chromosome 21 (Figures 1,2) a result that is supported by one previous study [16]. Initially, they removed those sequences containing existing, annotated exons and pseudogenes, which accounted for 1229 of the blocks sampled. However, the majority (2262) of the remaining conserved non-genic (CNGs) sequences appear not to match any of the ~230 known genes on chromosome 21. CNGs appear to be, on average, both smaller (~150 bp in length compared with the average chromosome 21 exon length of 270 bp) and more than twice as numerous as known exons. If functional, these sequences could include cis-regulatory elements, undiscovered protein-coding

Figure 1. The frequency of known exons (a) and conserved non-genic sequences (CNGs) (b) varies nonrandomly with respect to local GC content (c) across the long arm of human chromosome 21, starting at the centromere. Exons tend to be located in the GC rich distal end, whereas CNGs are predominantly located in the AT rich region proximal to the centromere. The average local GC content was calculated in non-overlapping 15-kb windows.

www.sciencedirect.com
exons, non-coding RNA genes (ncRNAs) and undescribed functional non-coding sequences. Alternatively, conservation could simply result from nonfunctional mutational ‘coldspots’.

Initially, the authors combined computational and experimental analyses to investigate the protein-coding potential of their data. Although similarity-based methods are often the most reliable methods of gene recognition, they are, by definition, limited by the availability of known protein-coding regions in related species. Instead, CNGs were compared with those exons predicted by the GrailEXP (http://compbio.ornl.gov/grailexp), Pro-Gen (http://www.anchorgen.com/pro_gen/pro_gen.html) packages and GenomeScan (http://genes.mit.edu/genomescan/). Such prediction algorithms generally operate on the principal that various compositional features are conserved across genes and can be used to predict probable coding structures in an unknown sequence. CNGs were also compared with existing human and mouse EST databases for similarities to existing cDNAs. These predictive approaches almost uniformly returned low numbers of candidate protein-coding genes. The authors also estimated the non-synonymous versus the synonymous (K_a/K_s) substitution ratio in the six potential reading frames of all CNGs; this is typically \(<0.3\) for most functional human–mouse orthologues [5]. Blocks within 100 kb of each other with low K_a/K_s ratios were identified as potential exon pairs. However, reverse transcription amplification (RT–PCR) of such ‘adjacent blocks’ indicated that only a small fraction produce mature mRNAs.

Combining the results of these analyses appeared to indicate that the coding potential of the majority of CNGs is small. Moreover, the distances between consecutive substitutions within CNGs appear randomly distributed compared with the highly distinctive ‘triplet’ pattern observed in coding sequence, which results from the elevation of substitution rates at codon third positions. Taken by itself, the apparent lack of protein-coding sequence in the data is unexpected. If we assume that most exons on human chromosome 21 are conserved between human and mouse, then it seems reasonable that most exons in this region, known and unknown, were part of the initial 3491 blocks that were detected. The removal of known exons and pseudogenes implies that only unknown exons, along with conserved non-coding sequence, could possibly have remained in the dataset. The failure to detect virtually any coding sequence in these remaining conserved blocks leaves two possible conclusions: (i) few unknown exons that are conserved between human and mouse exist within the sample region; or (ii) the variety of prediction methods employed have failed to detect unannotated exons that remain within the data.

**Defining a null hypothesis: the trouble with comparative genomics**

Although the results of Dermitzakis et al. would indicate that the CNGs they observe are not exonic, this does not consider the possibility that the conservation observed is merely a product of low divergence. The authors addressed this by attempting to amplify a subset of CNGs in rabbit [1] and a variety of more and less diverged mammalian outgroup species [2]. The rationale behind this is straightforward. Conservation in more widely diverged taxa provides further evidence that blocks are preserved as the result of selective optimisation of function, given that point mutations will have had more time to randomise neutral sequence between functional modules in more distant outgroups. Perhaps their most startling result shows that a fraction of CNGs is highly conserved in a wide variety of mammals. The level of conservation exceeds that observed in both known ncRNAs and coding sequence. Furthermore, comparison of the entire CNG dataset with the canine genome returned a high frequency of reciprocal best hits indicating that appreciable conservation across species might not be confined to the amplified subset. This
means that, if functional, at least some CNGs are sufficiently important to be retained, almost unchanged, across multiple, divergent evolutionary paths.

The possible functions of CNGs remain unclear. Conservatively, their numbers are estimated to be almost twice that of the predicted protein-coding sequences, and a large fraction of CNGs might be functional. However, although sequence conservation is often a signature of the operation of selective forces, we should proceed with caution before accepting conservation ipso facto as proof of function. The fact remains that one of the most difficult objectives in any attempt to locate functional blocks via comparative analysis is the definition of a robust null hypothesis. A comprehensive understanding of the pattern of mutational variability across the genome is fundamental to the success of comparative genomics. Our knowledge of such variability, however, is still incomplete although this could explain at least some observed conservation [19]. One approach that can, at least partially, elucidate the pattern of evolution we expect under neutrality is a comparison of rates of change within putatively functional sequences that have elements adjacent, such as ancestral repeats, which we can be relatively sure are evolving free of selection. We have used this approach in a recent study [20] to examine patterns of constraint in rodent non-coding sequence under the assumption that, excluding various potentially important regions such as donor and acceptor splice sites, intrinsic sequence is evolving neutrally. Our results support those of Dermitzakis et al. in that we also find comparatively large quantities of conserved non-coding sequence. However, it is important to note the conservation we observed is relative to the local, presumably neutral substitution rate, within introns and is thus, not directly comparable with that observed by Dermitzakis et al. Our analysis reveals a steady decrease in conservation with increasing distance upstream and downstream of a protein-coding sequence to approximately zero within 4 kb of most genes. By contrast, CNGs appear markedly disassociated with annotated exons: we infer a negative and highly significant correlation between the numbers of each within intervals along the chromosome (Pearson $r = -0.47; P < 0.001$). The large median distance between known coding sequence and CNGs suggests that there is unlikely to be much overlap between our datasets, and, indeed, recent results have indicated that conservation of CNGs is independent of their position relative to known genes [3].

The comparatively low GC content of CNGs (~38%) appears to reflect their genomic position, with the majority of CNGs located in the GC-poor proximal half of chromosome 21 (Figure 1). GC-poor regions are expected to be characterized by lower substitution rates than those in GC-rich sequence because of the lower frequency of hypermutable CpG dinucleotides. Although this could explain at least some of the unexpected conservation of CNGs between species, we estimate that both local GC content and CpG dinucleotide frequency explain little of the variation in human–mouse divergence. In addition, such causality would also require the conservation of base composition between species that for the human–mouse–rabbit comparison appears not to be the case.

If CNGs are not exonic, are they RNA genes?

Of the working hypotheses that were originally suggested, we can reasonably conclude that the hidden or unannotated exon theory has been the most convincingly eliminated. It seems unlikely that such a comparatively large number of otherwise unrelated sequences would fail to produce an appreciable frequency of predicted genes, spliced transcripts or matches with known cDNAs if they were exonic. The case for several ncRNAs being present within the CNG data is perhaps more convincing. BLAST comparison of CNGs with data from a transcriptional study of human chromosome 21 [21] indicated that approximately a fifth matched sequences that are known to be transcribed. On the basis of substitution pattern within blocks, ncRNA prediction software returned a comparable number of probable ncRNAs within the data. It would be interesting to investigate the degree of overlap between matched oligonucleotides and those predicted RNA gene models and to verify some of the predicted RNA gene models experimentally. Dermitzakis and colleagues also attempted a fine-scale analysis of substitutional patterns, testing for significant differences in clustering of variable sites and substitutional asymmetries between CNGs, coding sequence and ncRNAs. Although the distribution of variable sites appears to distinguish CNGs from protein-coding sequence, the results are uncertain in the case of ncRNAs, where the patterns of evolution appear to be highly variable. Furthermore, the estimates of substitution rate from the admittedly small number of known ncRNAs indicates that they are more highly conserved than coding sequences obtained by Dermitzakis et al. Although this could indicate a conservation bias in known ncRNAs it also reflects how little we know about the evolutionary significance of such genes.

It is important to note that, despite the high degree of similarity between some CNGs across many species, identity by itself does not confirm the orthology of CNGs among groups, although the probability of sequences of such high similarity being non-orthologous is small. We can be almost certain of the orthology of CNGs in the human–mouse comparisons, which were based on the criteria of synteny. The same can not be said of those comparisons with rabbit or more distant species. This might constitute a weakness of the approach of Dermitzakis et al. and contrasts with the targeted sequencing protocol that was implemented by Thomas and colleagues [15] where a region known to be orthologous to a segment of human chromosome 7 was amplified in multiple species. It is notable, however, that these authors’ conclusions do not depart greatly from those of Dermitzakis et al., in that they also observed a high degree of sequence conservation in presumed functional ‘deserts’.
Also the subset of 220 CNGs that was selected for PCR might be slightly biased towards higher conservation, presumably to increase the frequency of PCR success, compared with the original dataset of 2262. The selection of this subset is of crucial importance in determining whether high conservation is a feature of all CNGs. The high frequency of reciprocal best hits in the dog genome for the entire CNG dataset would support the conclusion that most blocks are conserved across species. Confirmation of further conserved blocks in other species is required, however, before we will know the true prevalence of CNGs that have evidence of functionality.

Concluding remarks
These caveats notwithstanding, Dermitzakis et al. have shown that evolutionary conservation of non-coding blocks on chromosome 21 extends across the entire mammalian lineage, a result that, in itself, is fascinating. However, although a comparative study can identify candidate functional regions, characterization must typically be sought via experimental assay. An approach employing these two steps to great effect in a recent study [22] identified long-range regulatory elements of several important cytokines. Other experimental research is now beginning to shed light on the possible functions that CNGs could perform.

Functional assays have provided evidence of regulatory motifs that operate at great distances from their associated genes [23]. In addition, substantial transcriptional activity has recently been reported outside of annotated regions along chromosome 21, although the proportions that are due to coding versus non-coding regions are difficult to ascertain [24]. Furthermore, the distribution of the binding sites of well-characterized human transcription factors on chromosome 21 suggests that many are located outside defined promoter regions [25] and there appears to be evidence of the regulatory function of a small number of CNGs on chromosome 21 located around the single-minded homolog 2 (SIM2) transcription factor [26]. Whether such examples are representative of cis-regulatory regions as a whole and whether the pattern of conservation of CNGs is indicative of such function remains uncertain. One interesting consequence of the GC-poor location of most CNGs is that the genes they are proximate to tend to have large introns and often require complex regulation. By contrast, housekeeping genes, which require little transcriptional regulation, tend to be situated in GC-rich (and thus CNG-poor) regions. Thus, there is again a suggestion that, if functional, some CNGs might have a role in long-distance expression control. In addition, conserved non-coding sequence might not only function in gene regulation but also might have a structural role, for example, as matrix scaffold attachment regions [27].

It is the ‘second-steps’ of experimental characterization, such as the studies described previously, which have perhaps the most intriguing possibilities for further work. If they are functional, the conserved regions being uncovered using the comparative method could finally begin to unravel the secrets of non-coding DNA.

Acknowledgements
We thank Mark Blaxter, Dan Halligan and the anonymous referees for valuable comments.

References
3 Dermitzakis, E. et al. (2004) Comparison of human chromosome 21 conserved non-genic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genomic environment. Genome Res. 14, 852–859
11 Jareborg, N. et al. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Res. 9, 815–824
18 Elnitski, L. et al. (2003) Distinguishing regulatory DNA from neutral sites. Genome Res. 13, 64–72
27 Glazko, G.V. et al. (2003) A significant fraction of conserved noncoding
Antibody class switching: uncoupling S region accessibility from transcription

Denise A. Kaminski and Janet Stavnezer

Department of Molecular Genetics and Microbiology, Program in Immunology and Virology, University of Massachusetts Medical School, Worcester, MA 01655, USA

Immunoglobulin class switch recombination (CSR) is a regulated process that changes antibody effector functions. Recently, Nambu et al. showed that histone acetylation is induced at switch (S) regions undergoing CSR; however, histone acetylation without accompanying S region transcription is insufficient to attract activation-induced cytidine deaminase (AID), which is required for CSR. They also show that AID can associate with mRNA polymerase II. These results support the model that germline transcripts are required to form single-stranded DNA, the AID substrate and further suggest that AID is recruited to S regions by the transcriptional machinery.

Activation of antibody-producing B lymphocytes (B cells) by antigen and co-stimulatory signals results in class switch recombination (CSR). CSR enables B cells to change the antigen constant (C) region, enhancing the ability of the antibody to eliminate pathogens, while maintaining the same antigen-binding variable region (Figure 1a). Following activation of mouse B cells, the C\(_{\mu}\) gene encoding the heavy chain C region of the initial IgM class can be replaced by any of the downstream C genes, resulting in expression of IgG3, IgG1, IgG2b, IgG2a, IgE or IgA (Figure 1b). CSR occurs by deletional recombination between repetitive G-rich switch (S) region sequences located upstream of each C gene except C\(_{\varepsilon}\). The intervening DNA is then detected as an extrachromosomal switch circle [1]. CSR requires the B-cell-specific enzyme, activation-induced cytidine deaminase [\(\text{AID} (\text{see Glossary})\)] [2,3]. Owing to its homology with an mRNA-editing cytidine deaminase, AID was first proposed to initiate CSR indirectly by editing an mRNA to create a novel endonuclease. However, accumulating evidence is consistent with AID acting directly on the DNA of the Ig S region, converting dC to dU residues, thereby initiating DNA repair processes that result in the DNA breaks required for CSR [4–8]. How AID is targeted to S regions has remained unclear and is the focus of a recent report by Nambu et al. [9].

S region accessibility during CSR

It has been proposed that S region susceptibility to recombination is caused by the obligatory transcription of S regions in their germline (GL) configuration [10–13]. This transcription initiates upstream (5') of the target S regions and continues through the C region exons (Figure 1b). Germline transcription is induced specifically by the same stimuli that induce CSR to each specific S region, namely B-cell activators, such as, lipopolysaccharide (LPS) together with cytokines [10]. The germline transcripts (GLTs) must also be spliced, however, suggesting that the function of GL transcription is not simply to induce accessibility of S region chromatin [10,12,13]. This hypothesis has been recently supported and extended by results from Nambu et al. [9], which suggest that AID association with S region DNA requires GL transcription and not simply chromatin accessibility.

To examine the accessibility of chromatin associated with specific S regions during CSR, Nambu et al. [9] used chromatin immunoprecipitation (ChiP) to assay histone acetylation, a modification known to occur in chromatin associated with actively transcribed genes [14]. Mouse splenic B cells were treated with LPS and interleukin 4 (IL-4), which induces GLTs initiating upstream of the \(\gamma_1\) and \(\varepsilon\) switch regions, and subsequently CSR to IgG1 and to IgA.

Glossary

AID: activation-induced cytidine deaminase; a B-cell-specific enzyme that is required for class switch recombination (CSR) and for somatic hypermutation of antibody variable region genes.

S region: 2–10 kb DNA segments located upstream of heavy chain constant (C) genes where switch recombination occurs. S regions consist of tandem repeats of 20–80-bp consensus sequences and contain frequent GAGCT, GGCGT and GGGCT elements.

Germline (GL) transcription: RNA synthesis that initiates upstream of switch (S) regions and is necessary for CSR. Although the GL transcripts are polyadenylated and spliced, no protein products have been detected.

R-loop: a non-coding nucleic acid structure consisting of RNA that forms a duplex with one strand of DNA, leaving the non-coding DNA strand single stranded.

UNG: the uracil DNA glycosylase required for normal levels of CSR. This might be due to its ability to excise dU bases that result from AID activity in genomic DNA.

Histone acetylation: modification of histone proteins by the addition of acetyl groups; this modification occurs most commonly on lysine residues of the N-terminal tails of histones H3 and H4. It is associated with actively transcribed genes.