NEWS AND COMMENTARY

Comparative genomics

Functional needles in a genomic haystack

L Bofkin and S Whelan

Heredity (2004) 92, 363–364. doi:10.1038/sj.hdy.6800429

The presence of vast amounts of nongenic, 'junk' DNA in most eukaryotic genomes has intrigued scientists for decades and, although most researchers agree that this DNA must be important, its function has remained stubbornly elusive. Comparative genomics has opened the gateway for powerful studies of junk DNA, typified by a recently published study by Dermitzakis *et al* (2003), who investigate a previously described class of highly conserved nongenic sequences (CNGs). They find that many CNGs appear more conserved between genomes than protein coding sequences (CODs) and functional nonprotein coding RNA molecules (ncRNAs). They also demonstrate that the patterns of evolutionary change in CNGs are different from CODs and ncRNAs, suggesting that they perform different functions. These evolutionary patterns provide evidence suggesting that CNGs are involved in protein binding.

Identifying new biologically interesting elements in a single genome is very much like finding needles in a haystack: one has to find the short series of letters that describe something useful among billions of As, Cs, Gs, and Ts. The fledgling field of comparative genomics is providing a magnet to find our needles, and as we sequence and align more genomes, the magnet is becoming larger and more powerful. Comparative genomics works by identifying sequences conserved between genomes (International Human Genome Sequencing Consortium, 2001; Dermitzakis *et al*, 2002; Dermitzakis *et al*, 2003; Margulies *et al*, 2003). Until recently, the majority of studies using this approach only concerned themselves with genes, the elements in the genome with the most phenotypically obvious function. The study of Dermitzakis *et al* (2003) suggests that CNGs may also perform a rich tapestry of functions; many of which may yield an improved understanding of genomic structure and regulation, and nuclear-genomic interactions under future investigation.

During evolution DNA sequences gradually accumulate changes through mutation; the greater the functional importance a piece of DNA has, the fewer changes it can accumulate as selection acts to preserve that function. It is estimated that around 5% of the human genome is restricted by selection in this way (Mouse Genome Sequencing Consortium, 2002), which greatly exceeds the 1.5% for gene sequence (International Human Genome Sequencing Consortium, 2001). Several studies have addressed issues with regard to the extent of non-genic conservation across a wide range of taxa, as distant as puffer-fish and human (Margulies *et al*, 2003). Dermitzakis *et al* (2003) also show the widespread presence of CNGs across mammalian taxa, but expand on this by demonstrating that CNGs have evolutionary characteristics that differ from both CODs and ncRNAs.

In a previous study, Dermitzakis *et al* (2002) identified numerous CNGs on human chromosome 21 from alignments between human and mouse using strict selection criteria: only sequences that produced an ungapped alignment in excess of 100 nucleotides and had greater than 70% sequence identity were classified as CNGs. The results from this study suggested that between 0.3 and 1% of the human genome consists of CNGs. In their recent paper, Dermitzakis *et al* (2003) sequence a selection of these CNGs from a number of placental and marsupial mammals to examine them in more detail, with a view to assigning putative functions.

To proceed in their comparative analyses, the authors compare the selected CNGs to CODs and ncRNAs from the species examined. Their first striking finding is that the CNGs identified in the human–mouse comparisons are generally more conserved than either the CODs or ncRNAs across the mammals, although one should be wary of inferring too much from this. The CNGs examined are clearly under very strong selective constraints, which is evident from widespread taxonomic distribu-

tion alone. Their relative importance to the other functional sequence examined is more difficult to assess. For example, CODs effectively contain nonfunctional sites due to the degeneracy of the genetic code and it is likely that if we removed degenerate third codon positions from the COD data, we may find that their functional sites are more conserved than CNGs.

The second interesting finding relates to the pattern of evolution in CNGs. The structure of the genetic code leads to a periodic pattern of evolution along the sequence in CODs: due to selective constraints the third codon position accumulates the most changes during the course of evolution and the second codon position the fewest. The CNGs do not exhibit this pattern of selective constraints, but were found instead to have stretches of comparatively uniform levels of constraint along their sequence, with alternating clusters of high and low nucleotide conservation. Furthermore, CNGs were not found to share the same mutational asymmetries associated with transcription from one strand of DNA, unlike both CODs and ncRNAs. This suggests that both strands of DNA are important in the function of CNGs, which is indicative of protein-binding regions, where specific nucleotides are optimally spaced on both strands of DNA. The authors also suggest that the mutational patterns shown by CNGs are comparable to other protein-binding DNA regions and that the alternating constrained–relaxed pattern of conservation in the CNGs may indicate that each may bind multiple proteins (see Dermitzakis *et al*, 2003; supplementary material).

This leaves the question of what are CNGs binding to that makes them so conserved through evolution? Dermitzakis *et al* suggest that the function of CNGs reflects the broad categories of protein-binding DNA and that they do not have a single function. Many CNGs are likely to be important regulatory elements conserved throughout mammals. This does not, however, explain all CNGs because many appear to fall in regions far away from the nearest gene, the so-called 'gene deserts'. While immediate proximity to a gene is not a prerequisite for transcriptional regulation, the number of protein-binding sites at such a distance suggests they may have other functions. Many CNGs may prove to be good candidates for regions associated with chromatin organization; the presence of multiple protein-binding sites would support this.

Either way, the identification of so many putative protein-binding CNGs is exciting and future studies of them may yield new DNA-binding motifs.

Overall, their work provides a good example of how comparative genomics is helping us find needles in very large haystacks. The exact function of these needles remains unclear, but hopefully further laboratory investigation has begun and may shed light on some fundamental aspects of genome biology.

Correspondence: L Bofkin and S Whelan are at the EMBL – European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Simon Whelan is also at the Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK.

e-mail: simon@ebi.ac.uk

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flege V, Bucher P, Jongeneel CV, Antonarakis (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**: 578–582.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE (2003). Evolutionary discrimination of mammalian conserved non-genic sequences. *Science* **302**: 1033–1035.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED (2003). Identification and characterisation of multi-species conserved sequences. *Genome Res.* **13**: 2507–2518.

Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.