Research Focus

Clues to function in gene deserts

James Taylor

Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802, USA

Recent work by Ivan Ovcharenko and colleagues has shed new light on the functional importance of gene deserts. They demonstrate that sequence conservation levels separate gene deserts into stable (more conserved) and variable classes. Both classes exhibit characteristics suggestive of function. The stable deserts in particular show features suggesting a role in the complex regulation of core vertebrate genes.

Introduction

The best understood functional elements, protein-coding regions, make up only ~1.5% of the human genome [1], whereas the functional importance of the remainder is still poorly understood. Furthermore, the distribution of protein coding regions is not random, yielding intergenic regions far longer and more frequent than would be expected owing to chance alone. It has been estimated that these gene deserts make up ~25% of the genome [2]. Several recent studies have begun to investigate the functional role of these peculiar features of the genomic landscape.

Questioning the utility of gene deserts

One intriguing hypothesis is that gene deserts contain distant sequences involved in the regulation of flanking genes. Using sequence conservation Nobrega *et al.* [3] identified enhancer elements in the region consisting of the human gene *DACH1* (a widely expressed gene involved in brain, limb and sensory organ development) and the two gene deserts that flank it. In the 870 Kb desert upstream of this gene they found five elements that could regulate its expression. These elements are spread between 225 Kb and 780 Kb upstream of *DACH1*, and are conserved in human, mouse and more-distant vertebrates, including the frog *Xenopus tropicalis* and the fish *Danio rerio*, *Tetraodon nigroviridis* and *Fugu rubripes*. Additionally the linear ordering of the five elements described and *DACH1* is the same in mammals and fish despite chromosomal rearrangements that have broken continuity with flanking genes, suggesting that not just the sequences, but also their relative positions on the chromosome, might be important.

However, some gene deserts are not necessary for survival. Although the discovery of new conserved features and classes of functional elements has prompted rethinking the existence of 'junk DNA', gene deserts still seem to be likely candidates for disposable sequence. Nobrega *et al.* [4] describe the deletion of two gene deserts in mice, a 1817 Kb region from chromosome 3 and a 983 Kb region from chromosome 19. The resulting mice were viable with no detectable phenotypic difference and only minor differences in gene expression compared to the wild type. Although the authors carefully warn that they might have been unable to detect changes caused by the desert deletions, the results seem to support the idea that some gene deserts might not have a functional role.

Sequence conservation reveals a dichotomy

Taken together, these results raise the possibility that there are different classes of gene deserts with distinct functional roles. Ovcharenko *et al.* [5] find that conservation clearly separates gene deserts into two classes: weakly conserved variable gene deserts and more conserved stable gene deserts. As a metric for conservation they use coverage by human/chicken evolutionarily conserved regions (ECRs), one of several criteria recently proposed for identifying short highly conserved regions in alignments [6–8]. Although the threshold for separating the two desert classes is somewhat arbitrary, the partitioning is supported by other conservation metrics [including human and pufferfish (*Fugu rubripes*) ECRs and phastCons [9] scores] and thus seems to be robust.

Neutral mutation rates vary substantially across the human genome [10]. This variation is a plausible explanation for the observed partitioning of gene deserts; the higher conservation seen in stable gene deserts than the variable gene deserts might result from low neutral rates. However, the authors find that neutral mutation rates are actually elevated in both the stable and variable gene deserts. This suggests that the separation is more than just a side effect of neutral rate variation, and conservation in stable deserts might be owing to selective pressure.

Signs of function in stable deserts

In addition to conservation, stable gene deserts exhibit characteristics that suggest gene regulatory function. Stable deserts flank several genes known to have distal regulatory elements. Stable deserts also contain three times the density of putative regulatory elements identified by regulatory potential scores [11] as found in variable deserts. Perhaps most remarkably, stable deserts show almost complete syntenic conservation in the chicken genome. Most of these deserts appear to be conserved as a complete syntenic unit, with the linear order of ECRs in the desert and flanking genes largely unbroken. This resilience to separation from flanking genes and

Corresponding author: Taylor, J. (james@bx.psu.edu).
Available online 19 April 2005

preservation of relative order strongly supports the idea that stable deserts harbor regulatory elements.

Given the strong association between stable deserts and their flanking genes, Ovcharenko *et al.* investigated the gene ontology [12] annotations of these genes. Genes flanking stable deserts are enriched for gene ontology (GO) categories related to regulation and transcription, as well as metabolism and development (Table 1). Important developmental genes can have complex expression patterns, and this enrichment points toward the possibility that stable gene deserts are involved in the regulation of these genes. The authors suggest an intriguing hypothesis for the origin of gene deserts; that they are regulatory units for key genes, and have drifted apart as vertebrate genomes have expanded, but cannot tolerate rearrangement. Further evidence for this hypothesis is provided by genes between what the authors call conjoined stable deserts; regions where two stable deserts are separated by no more than three genes and one megabase. The genes in these regions are also enriched for regulatory functions, and include many genes related to crucial biological processes and steps in vertebrate development.

Association between variable deserts and genes

Although lack of conservation in variable gene deserts makes them more difficult to characterize than the stable deserts, their importance should not be discounted. Examining the GO categories for genes flanking variable deserts, Ovcharenko *et al.* found that they also show strong enrichments (Table 1), suggesting that these deserts might have different but still significant functional associations.

Hillier *et al.* [13] recently undertook further investigation of the conservation structure in gene deserts as part of their annotation of human chromosomes 2 and 4.

Table 1. Gene ontology category enrichment for genes flanking deserts, relative to purely-by-chance case[a]

Gene ontology category	Enrichment
Stable gene deserts	
Regulation of metabolism	4.4
Transcription factor activity	4.2
Transcription coactivator activity	4.0
Regulation of biosynthesis	3.8
Transcription regulator activity	3.6
Transcription factor binding	3.2
DNA binding	2.8
Regulation of transcription	2.8
Transcription	2.7
Development	2.0
Variable gene deserts	
Glutamate receptor activity	7.8
Inotropic glutamate receptor activity	7.7
Amine receptor activity	6.2
Sulfotransferase activity	4.2
Cell adhesion	3.0
Transmission of nerve impulse	2.8
Neuromuscular physiological process	2.8
Synaptic transmission	2.7
Calcium ion binding	2.2
Organogenesis	1.9
Morphogenesis	1.7
Development	1.7
Cell communication	1.6

[a]Reproduced from [5] © 2005 Cold Spring Harbor Laboratory Press.

These chromosomes together contain several of the largest human gene deserts. The authors first investigated the *PCDH7* gene (a member of the protocadherin gene family, which code for membrane proteins involved in cell recognition and adhesion), which is flanked by the largest human gene desert (>5Mb) and another large desert (3.5Mb). Interestingly, although these are variable deserts according to the definition of Ovcharenko *et al.* [12], the synteny of this region, including flanking genes, appears to be conserved in mammals and birds. Hillier *et al.* [13] also show that *PCDH7* is not the only instance where a protocadherin is flanked by two large variable gene deserts. For example, *PCDH10* (which appears to have been separated from *PCDH7* before mammals and fish diverged) is surrounded by 5.1 and 4.0Mb deserts. The fact that this structure is so distantly conserved again suggests a functional role for variable gene deserts.

Hillier *et al.* [13] also investigated a 3.3Mb stable desert flanking the *ZFHX1B* (a transcription factor expressed early in development) gene on chromosome 2. This desert is unique in that it overlaps with the longest human/chicken 'non-coding conservation jungle' [14]. This desert shows an interesting pattern of conservation, with the entire interval well conserved between human, mouse and dog, but only ~2Mb immediately adjacent to *ZFHX1B* are conserved with chicken and fish. The authors then attempted to distinguish the regions conserved in mammals from the others based on differences in short nucleotide patterns in the human sequence. They found that short patterns in the human sequence, without the aid of comparative information, allow for a reasonably accurate discrimination between these intervals. The existence of distinguishable patterns in the primary sequence of these regions is another clue that they might contain functional elements.

Conclusion

Because they are depleted of protein-coding regions, gene deserts have long been thought by many to be devoid of function. Recently, it has been shown that for at least some gene deserts this is not true. Ovcharenko *et al.* took a significant step in unraveling the mystery of gene deserts by showing that these genomic regions fall in well-separated conservation classes. Investigating stable and variable gene deserts separately allows interesting functional characterization to emerge for each class. Ovcharenko *et al.* make a strong case for the functional importance of stable deserts, whose patterns of conservation and association with flanking genes suggest that they might host the complex regulatory machinery of important vertebrate genes. Although variable deserts are even more difficult to characterize, their functional role should not be discounted. This seems particularly true in light of work by Hillier *et al.* showing evidence of an ancient relationship between these deserts and certain gene families, despite a lack of conservation in primary sequence. Understanding the cause and significance of these findings offers a challenging avenue for future research.

References

1 Lander, E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science* 291, 1304–1351
3 Nobrega, M.A. *et al*. (2003) Scanning human gene deserts for long-range enhancers. *Science* 302, 413
4 Nobrega, M.A. *et al*. (2004) Megabase deletions of gene deserts result in viable mice. *Nature* 431, 988–993
5 Ovcharenko, I. *et al*. (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Res.* 15, 137–145
6 Ovcharenko, I. *et al*. (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.* 32, W280–W286
7 Margulies, E.H. *et al*. (2003) Identification and Characterization of Multi-Species Conserved Sequences. *Genome Res.* 13, 2507–2518
8 Bejerano, G. *et al*. (2004) Ultraconserved elements in the human genome. *Science* 304, 1321–1325
9 Siepel, A. and Haussler, D. (2004) Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* 11, 413–428
10 Hardison, R.C. *et al*. (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* 13, 13–26
11 Kolbe, D. *et al*. (2004) Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res.* 14, 700–707
12 Ashburner, M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
13 Hillier, L.W. *et al.* The DNA sequence of chromosomes 2 and 4. *Nature* (in press)
14 Hillier, L.W. *et al*. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716

0167-7799/$ - see front matter
doi:10.1016/j.tibtech.2005.04.003

Small molecule microarrays: from proteins to mammalian cells – are we there yet?

Gabriela Chiosis[1] and Jeffrey L. Brodsky[2]

[1]Department of Medicine and Program in Molecular Pharmacology and Chemistry, Memorial Sloan-Kettering Cancer Center, New York, NY 10021, USA
[2]Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

A recent publication by Stockwell and colleagues documents a leap forward toward the continued development of small molecule microarray (SMM) technology. By creating microarrays of small molecules impregnated in a biodegradable polymer, the authors have, for the first time, shown that SMMs can be used in a cell-based format. This technological improvement opens the door for using SMMs to perform high-throughput screens in mammalian cells.

Introduction

Microarray technology is modernizing biomedical research by allowing the simultaneous analysis of tens of thousands of samples and by examining low nanomolar to picomolar amounts of materials. A recent study estimated an annual compound growth rate for the microarray market of 63%, from US$232 million to US$2.6 billion, between 1999 and 2004. An ever-expanding sector of the microarray field is small molecule microarrays (SMMs), in which small molecules are immobilized on a surface to probe protein targets. The first reported use of SMMs dates back to 1999 and derived from work performed in the laboratory of Stuart Schreiber at Harvard University [1]. Since then, significant technological advances have led to improvements in SMMs [2]. For example, libraries compatible with SMM immobilization strategies were created [1,3–5] and superior methods of immobilization and target presentation were generated [6–9]. In the first SMMs, small molecules (SMs) were covalently immobilized on the surface of the array (Figure 1a), regardless of whether the molecules were synthesized by split-pool or parallel methods [1,3–5], or else they were generated *in situ* [10]. These microarrays were used to identify modulators of distinct proteins in diverse biological pathways, including Ure2p, which represses a glucose-sensing pathway in yeast [11], and Hap3p, a component of the yeast Hap2/3/4/5p transcription factor complex, which has a role in nutrient signaling and aerobic respiration [12]. A novel calmodulin ligand [4] and ligands of human IgGs [5] were also identified.

Another advance was the development of a technology that circumvented the covalent attachment of molecules to arrays. Diamond and Gosalia instead 'printed' a 352 member combinatorial library in nanoliter glycerol droplets onto microarrays [13]. Because only 1 nanomole of each compound was required in each of these micro-'reactors', 100 replicate slides with 1 mM of each compound could be prepared and arrayed in quadruplicate. Next, caspases 2, 4 and 6 and their respective fluorogenic substrates were aerosolized over the array, which allowed a parallel analysis of the inhibitory effects of each small molecule on enzyme activity (Figure 1b). As a result of this strategy, a caspase inhibitor was discovered that acted on

Corresponding author: Chiosis, G. (chiosisg@mskcc.org).
Available online 19 April 2005