NEWS AND COMMENTARY

Comparative genomics

Mining the crucial 1%

AL Hughes

Heredity (2004) **93**, 5, advance online publication, 5 May 2004; doi:10.1038/sj.hdy.6800461

The human genome is estimated to be about 99% identical at the nucleotide level to that of our closest living relative, the chimpanzee (Sibley and Ahlquist, 1984; Sakate *et al*, 2003). The search for the biological basis of uniquely human characteristics thus focuses on only about 1% of our genome. King and Wilson (1975) predicted that phenotypic differences between human and chimp are due mainly to differences in gene regulation, rather than differences in the sequences of the proteins encoded by the two species. This hypothesis remains to be tested, although there is evidence of expression differences between human and chimpanzee, particularly in brain tissue (Enard *et al*, 2002).

Even if it turns out to be true that the bulk of uniquely human traits arose as a consequence of changes in gene regulation, it is still worthwhile to ask whether changes in protein sequences have played a role in making us human. Clark *et al* (2003) have used newly available sequences from the chimpanzee genome to address this question. Reasoning that major phenotypic changes are likely to have arisen as a result of adaptive evolution, they searched a large set of homologous human and chimpanzee genes for a statistical signature of adaptive changes in amino-acid sequence.

Since most nonsynonymous mutations (those that cause amino-acid changes) are harmful to protein structure, they are quickly eliminated from population by conservative or 'purifying' natural selection (Kimura, 1977). On the other hand, synonymous mutations (those that do not cause amino-acid changes) are usually selectively neutral or nearly so. As a consequence, in most protein-coding genes, the number of synonymous nucleotide substitutions per synonymous site (d_S) exceeds the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) (Nei, 1987). In a small proportion of genes, however, d_N has been found to exceed d_S . This unusual pattern is evidence of positive Darwinian selection actively favoring changes at the amino-acid level (Hughes and Nei, 1988).

Clark *et al* (2003) examined a dataset of 7645 orthologous gene pairs from human and chimpanzee. (Two genes are said to be orthologous if they are descended from a common ancestor without gene duplication.) For each gene pair, a putative orthologous gene from the mouse was used as a reference to compare rates of evolution in the two primate genes. Clark and colleagues used two maximum likelihood-based approaches to test the null hypothesis that $d_N/d_S \leq 1$ against the alternative that $d_N/d_S > 1$. One of these approaches is believed to be capable of detecting positive selection acting on one domain of a protein while other regions are subject to purifying selection, which is the usual form taken by positive selection on protein-coding genes. Using this latter approach, Clark and co-workers reported evidence of positive selection (ie $d_N/d_S > 1$) that was significant at the 5% level in 667 orthologous gene pairs (8.7% of the total).

When genes showing evidence of positive selection were examined, a substantial number were found to be involved in sensory perception. These included numerous olfactory receptors, as well as genes involved in hearing. For example, one of the genes showing evidence of positive selection was the gene encoding alpha tectorin, a protein expressed in the tectorial membrane of the inner ear. Given the importance of language in human evolution, it is tempting to speculate that an increase in hearing acuity may have been a prerequisite for perceiving the subtle acoustic signals on which language is based.

Clark *et al*'s (2003) study can be faulted on statistical grounds. Although they conducted thousands of independent tests for positive selection, the authors applied no correction for multiple testing. It is expected that under these circumstances, 5% of tests will show a 'significant' result by chance alone; and the observed proportion of tests showing a significant result (8.7%) was not greatly different from that expected by chance. This possibility of 'false discovery' is an issue that often comes up in large genomic surveys, and

new statistical methods have been developed because traditional corrections for multiple comparisons can be over-conservative (Devlin *et al*, 2003). In addition, maximum likelihood methods of testing for positive selection have been shown in several studies to be nonconservative and overly sensitive to deviations from the model of sequence evolution that is assumed (see, for example, Suzuki and Nei, 2002). Nonetheless, it can be argued that the use of a nonconservative statistical approach is justifiable because Clark *et al* (2003) were engaged in data mining rather than hypothesis testing *per se*. Their study can be viewed as a genome-wide search for candidate loci for positive selection, and each candidate they found can be examined in more detail in the future.

In the most convincing examples of positive selection at the molecular level, knowledge of protein structure and function is incorporated in the analysis. As a result, the amino-acid changes that have occurred as a result of selection can be related directly to an adaptive phenotype (Hughes, 1999). In future studies, such detailed analysis can be applied to the candidate loci identified by Clark *et al* (2003). For example, detailed knowledge of the structure of alpha tectorin will make it possible to test the hypothesis that changes have targeted residues important for hearing acuity. Furthermore, it will be possible to combine data on gene expression with data on amino-acid sequence differences to provide a complete picture of the evolutionary changes that made us human. In terms of our understanding the basis for human uniqueness at the molecular level, the study of Clark *et al* (2003) is only the beginning, but it represents an important first step.

AL Hughes is at the Department of Biological Sciences at the University of South Carolina, Columbia, SC 29208 USA.

e-mail: austin@biol.sc.edu

Clark AG *et al* (2003). *Science* **302**: 1960–1963.
Devlin B *et al* (2003). *Heredity* **91**: 537–538 (News And Commentary).
Enard W *et al* (2002). *Science* **296**: 340–343.
Hughes AL (1999). *Adaptive Evolution of Genes and Genomes*. Oxford University Press: New York.
Hughes AL, Nei M (1988). *Nature* **335**: 167–170.
Kimura M (1977). *Nature* **267**: 275–276.
King MC, Wilson AC (1975). *Science* **188**: 107–116.
Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.
Sakate R *et al* (2003). *Genome Res* **13**: 1022–1026.
Sibley CG, Ahlquist J (1984). *J Mol Evol* **20**: 2–15.
Suzuki Y, Nei M (2002). *Mol Biol Evol* **19**: 1865–1869.