

Signatures of adaptive evolution within human non-coding sequence

Chris P. Ponting* and Gerton Lunter

MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK

Received July 5, 2006; Revised and Accepted July 14, 2006

The human genome is often portrayed as consisting of three sequence types, each distinguished by their mode of evolution. Purifying selection is estimated to act on 2.5–5.0% of the genome, whereas virtually all remaining sequence is considered to have evolved neutrally and to be devoid of functionality. The third mode of evolution, positive selection of advantageous changes, is considered rare. Such instances have been inferred only for a handful of sites, and these lie almost exclusively within protein-coding genes. Nevertheless, the majority of positively selected sequence is expected to lie within the wealth of functional ‘dark matter’ present outside of the coding sequence. Here, we review the evolutionary evidence for the majority of human-conserved DNA lying outside of the protein-coding sequence. We argue that within this non-coding fraction lies at least 1 Mb of functional sequence that has accumulated many beneficial nucleotide replacements. Illuminating the functions of this adaptive dark matter will lead to a better understanding of the sequence changes that have shaped the innovative biology of our species.

Although the amount of the human genome that harbours functional, yet non-coding, elements remains ill-determined, models of sequence evolution are unanimous in predicting at least as much functional non-coding sequence as protein-coding material in the genome (1–3). Beyond the repertoire of sequences known to be promoting or regulating transcription and translation, there appears to be a large set of functional sequence [‘dark matter’, (4)] whose importance is yet to be understood. Determining the evolution and function of dark matter is critical to resolving an on-going debate as to whether it specifies much of the morphological diversity of animals, the phenotypic diversity of humans and an individual’s susceptibility to disease (5).

Between 10 and 15% of patients with rare Mendelian phenotypes exhibit no changes to a gene’s coding sequence, despite incontrovertible evidence of its association with disease (6). In these cases, it must be assumed that the as-yet-unknown mutations lie in the functional non-coding portions of the human genome. Indeed, mutations in intronic elements (6,7), promoters (8) and untranslated regions (UTRs) (9) have, on occasion, been associated with disease. The identification of other disease-associated mutations in the non-coding sequence is hindered by the fact that we currently possess few insights into how to identify functional sequence outside of the coding exons by computational means.

Here, we review recently published evidence for substantial amounts of human functional sequence outside of the protein-coding sequence. For this, we need to consider the nature and extent of neutrally evolved sequence in the human genome because such sequence represents the exact complement of all functional regions. Thereafter, we discuss the amount of constrained (‘conserved’) genomic sequence. Finally, we entertain the possibility that a significant proportion of human sequence has evolved adaptively and thus has diverged by a greater extent than expected from neutral evolution.

NEUTRALLY EVOLVED SEQUENCE

How much human DNA, during evolution, has been purified of deleterious mutations (‘purifying selection’); how much has accepted mutations because of their benefit (‘positive selection’) and thus what remaining proportion of the human genome has accumulated mutations that have not been selected for or against (‘neutral evolution’)? Because of their abundance and their ease of estimation from aligned sequences, nucleotide substitutions have provided the principal mutational signature from which neutrality or selection has been inferred. However, later in this review, we discuss a recently developed approach that harnesses

*To whom correspondence should be addressed. Tel: +44 1865285855; Email: chris.ponting@anat.ox.ac.uk

nucleotide insertions and deletions, rather than substitutions, to distinguish between selected and neutrally evolved sequence.

Three distinct classes of nucleotides have often been considered as having evolved neutrally: pseudogenes, the remnants of defunct genes or reverse-transcribed messenger RNA (10,11); ancestral repeats, the debris of transposons present in the last common ancestor of, for example, human and mouse (1,2); and 4-fold degenerate (4D) sites, the third position of codons that encode one particular amino acid whichever base is present (12). Unfortunately, none of these three types of sites is, in fact, universally neutral. In rare cases, pseudogenes appear to have been subject to selection (13,14) and a minority of transposable elements have acquired innovative function (15–17). Furthermore, 4D sites in mammals and invertebrates have been shown to be subject to weak and strong selection, respectively (18,19). Nevertheless, lacking alternatives, these type of sites remain as widely used proxies for unselected sequence.

Substitution rates in these putatively neutral sequences may well be relatively constant in small (<100 kb) regions, but certainly vary dramatically across mammalian genomes (1,12,20,21). Why this is so remains unclear, although there are predicted contributions to this variation from the hypermutability of CpG dinucleotides (22,23), from recombination (24), from the repair of sequence transcribed in the germ-line (25,26) and from base composition not being at equilibrium (27). Whatever the cause, the effect of substitution rate variation across the human genome is that a single neutral rate for the whole mammalian genome does not exist and, thus, that such rates need to be estimated locally.

CONSTRAINED NON-CODING SEQUENCE

Two essentially complementary approaches predict that only a small proportion of the human genome has been subject to strong purifying selection. The first of these, from Chiaromonte and coworkers (1,2), indicates that ~5% of the mammalian genome has purified substitutions since the mouse and human common ancestor. Theirs is a relatively simple model and thus unlikely to yield more than a rough estimate (28,29). It is predicated upon two assumptions: that substitutions in ancestral repeats were accumulated without selection, and that neutral rates in ancestral repeats and neighbouring sequence are equivalent, despite base composition differences (30).

We recently introduced a second, complementary, approach, one which considers insertions and deletions ('indels') between human, dog and mouse sequences rather than substitutions (3). The method, which also accounts for neutral rate variation genome-wide, predicts that between 2.56 and 3.25% of the human genome has been selectively purged of indels and thus is functional. Moreover, it provides quantitative support for the assumption of Chiaromonte *et al.* that ancestral repeats predominantly evolve neutrally, predicting that only ~0.1% of all transposable elements are selectively constrained.

Given that 1.2% of the human genome encodes protein (31), each of these two approaches thus indicates a greater

amount (~1.3–4%) of functional sequence residing outside of the coding sequence than inside. Many of these selectively constrained non-coding sequences are even better conserved than protein-coding sequence, yet on the whole their functions remain mysterious (32–36). Nevertheless, functional clues may be elicited from their non-uniform distribution among introns. It is observed that longer introns in general, and in particular, introns of genes regulating transcription, morphogenesis or organogenesis and introns within nervous-system-expressed genes on average possess higher than expected densities of conserved sequence (37,38). This suggests that conserved intronic sequence might often regulate processes during transcription and development.

Conversely, highly conserved intronic regions are significantly under-represented among genes with roles in response to pathogenic insults (38). Consequently, because the amino acid sequences of proteins involved in transcription and development generally evolve slowly, and those involved in immunity and host defence evolve rapidly, it appears that selection has acted in a relatively uniform manner across genomic loci: divergent protein sequences are encoded by genes whose introns are subject to relaxed constraints, whereas genes of conserved protein sequences contain longer and more conserved introns. In particular, it is notable that genes whose expression is limited to the nervous system often possess highly conserved protein-coding, UTR and intronic sequences (38–40).

ADAPTIVE EVOLUTION WITHIN NON-CODING SEQUENCE

The higher abundance of non-coding over coding sequence within the constrained portion of the human genome indicates that the majority of functional sequence is non-coding. It thus appears possible that recent adaptive events too might have involved more non-coding than coding sequence. Despite this, most attention has been paid to detect positive selection in coding sequence. Partly, this is because protein-coding sequence is more easily identified and annotated and partly because synonymous sites can be exploited to provide an estimate of the local neutral rate against which substitution rates within proximal non-synonymous sites can be compared (41,42). These methods, for example, have been exploited to identify adaptive amino acid substitutions proposed to be linked to the development of speech or to the enlargement of the hominin brain (43–45).

In contrast, predicting positive selection in non-coding sequence is hindered by the difficulty of identifying functional sequence when it has rapidly evolved, by the lack of proximal presumed neutral sites and by variations in neutral rate (46). Nevertheless, non-coding substitutions close to human *LCT* or *CYP1A2* genes have been identified which appear to affect their expression levels. These genes encode lactase and cytochrome P450 1A2, and the identified alleles have been associated with acquired tolerances to lactose or toxins (47–49).

An approach for detecting adaptive evolution within modern populations is to identify genomic regions that show evidence of a recent selective sweep (50). Such regions of

diminished sequence variation and high linkage disequilibrium are indeed enriched in the vicinity of protein-coding genes, such as those involved in the immune response and sensory perception, which are most expected to have been the targets of positive selection (51). Although a powerful approach, it only identifies relatively large intervals within which the site subject to positive selection still remains to be identified.

INDELS AND HETEROGENEOUS SELECTION

Identifying the substrates of adaptive evolution would be more straightforward if the functional portion of the genome were already identified and separated from the sea of neutral sequence in which it lies scattered. The method of Chiaromonte and coworkers (1,2) that predicted ~5% of sequence to be conserved, and thus functional, cannot by itself exactly pinpoint the conserved bases, whereas other phylogenetic approaches are able to do so (17). Unfortunately, although conservation over long time spans does imply function (36), not all functional sequence is conserved. Consequently, these methods are less effective at pinpointing sequences that have evolved by positive selection than they are at identifying selectively purified sequence. What was required instead was a robust method to identify rapidly substituting, yet functional, sequence among all non-coding genomic regions.

To this end, we sought first to identify genomic regions that demonstrate another hallmark of purifying selection, the purging of inserted or deleted nucleotides, to obtain a set of likely functional sequence. Then, among these regions, we kept only those that, additionally, exhibit nucleotide substitutions at rates exceeding the expected neutral rates. We termed this confluence of purifying selection on indels and of positive selection on nucleotide substitutions as heterogeneous selection (3).

Specifically, we identified 54.4 Mb of sequence in which no indels had been fixed since the common ancestor of human and mouse (3). This set contained a predicted 1% of false positives that have evolved neutrally but had escaped any indel events purely by chance. If no regions of the human genome evolved by heterogeneous selection, then we might expect only 1% of this 54.4 Mb set to exhibit nucleotide substitution rates at, or over, their neutral rates. Nevertheless, we observed five times this amount of rapidly substituting sequence within the indel-purified 54.4 Mb (Fig. 1). This implies the strong admixture of sequences whose nucleotide substitutions either have been under relaxed constraints or have often been subject to positive selection. Much of this indel-purified and sequence-divergent DNA is within known functional material, such as protein coding exons, and it is substantially under-represented within transposable elements. This suggests that within this non-coding material should be the functional sequence that has been the target of positive selection upon nucleotide substitution.

AMOUNT OF POSITIVELY SELECTED SEQUENCE IN NON-CODING REGIONS

These results lead us to consider the possibility that ~1 Mb of sequence has evolved adaptively. This low value (~0.03% of

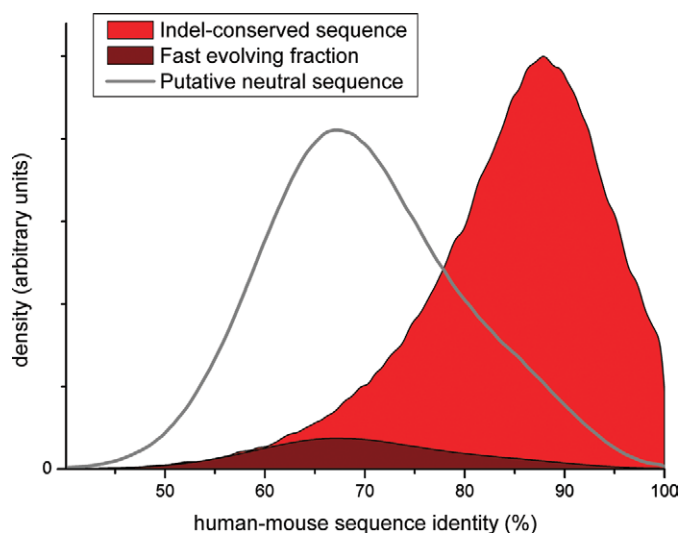


Figure 1. Evidence for elevated substitution rates within functional sequence. A lack of insertion and deletions (indels) indicates the past action of purifying selection and thus signals functional sequence. We used this signature to identify 54.4 Mb of human indel-free sequence in alignments with the mouse and dog genome. Our method was calibrated to include only 1% of false positives, neutrally evolving sequence which, by chance rather than selection, escaped any indel event over the evolutionary time span considered. If it is assumed that all indel-purified sequence is also purified of substitutions, the distribution of sequence identity within the 54.4 Mb set (red) would be a mixture of an unknown 'conserved' distribution and a 1% contribution from a 'neutral' distribution of sequence identity (grey outline). Instead, the observed distribution is consistent with an admixture of six times this amount of quasi-neutral evolution (maroon). Half of the 5% excess (about 1 Mb) has thus been acquiring substitutions above the mean neutral rate, despite being selectively purified of indels and thus presumably functional, suggestive of the possibility that positive selection caused the high rate of fixation of these substitutions. [Adapted from Lunter *et al.* (3), Fig. 7.]

the human genome) is consistent with the results of others who compared human polymorphism with human–chimpanzee divergence data (52).

This small proportion pales in comparison with the estimated 20% contribution from positive selection to the divergence between the fruit fly *Drosophila melanogaster* and its sister species *D. simulans* in intronic and intergenic sequence (53). [This is in addition to the large contribution to fruit fly amino acid sequence divergence predicted to arise from adaptive evolution (54–57).] It is to be expected that adaptive evolution would impact most on species such as fruit flies, whose effective population sizes are considerably larger than those of mammals, simply because selection on mutations is more efficient (58). However, although likely to be less widespread, it would be curious if positive selection were not also to have acted upon mammalian introns and intergenic sequences as it has on fruit fly sequences.

For four reasons, the 1 Mb of adaptive human DNA may be a considerable under-estimate. First, the method necessarily only exploits orthologous regions that retain sufficient resemblance to allow their accurate alignment. Lineage-specific or orthologous segments whose sequences have diverged greatly as a consequence of positive selection are thus not able to be aligned and are not counted towards the genomic total. For example, it has been shown that sequences

unaligned between human and mouse often contain structural RNA elements (59). Secondly, sequences are often not included if they have recently gained function, owing to their sequence divergence being intermediate between those of neutral and constrained sequences. Thirdly, the method misses adaptive sequence within which selection has not acted heterogeneously, but instead has driven both beneficial indels and substitutions to fixation. Finally, it also overlooks positively selected sites, or short regions, that are scattered among a majority of constrained bases. Conversely, the 1 Mb total may wrongly include sequences that have not evolved adaptively, such as regions that either have lost constraint recently or have evolved by a combination of constraint and neutrality, because a minority of these regions will have accumulated high numbers of unselected substitutions purely by chance. The 1 Mb of sequence under positive selection thus should be considered to be an approximate first estimate.

More recently, we exploited this signature of selection upon indels to conduct a genome-wide scan for positive selection on small functional intronic elements. We find such elements to be especially abundant in the introns of genes that are expressed in the brain (Lunter and Ponting, submitted for publication). These results are consistent with a recent study of human sequences whose evolution has been rapid only in the few million years since our last common ancestor with chimpanzees. Haussler and coworkers found that genes involved in transcriptional regulation or in neurodevelopment are significantly associated with such human accelerated regions (HARs). One particularly striking example involves a 118 bp region (HAR1) that is expressed specifically in Cajal–Retzius neurons. Exceptionally, this region, which folds into a stable RNA structure, has undergone 18 base changes since the human–chimpanzee ancestor, of which 10 have been compensatory and thus consistent with the predicted secondary structure (K.S. Pollard, S.R. Salama and D. Haussler, personal communication). This example, perhaps the single most striking example of human-specific positive selection in non-coding regions to date, hints at a larger role of positive selection in non-coding sequence than hitherto appreciated.

POSITIVE SELECTION AND TURN-OVER OF FUNCTIONAL SEQUENCE

The division of genomic DNA into the well-known trichotomy of neutral, conserved and positively selected sites is, of course, an over-simplification. In particular, it does not consider sequence whose functionality has been intermittent over the long timescales separating mammalian species. The impermanence of functional sequence is most apparent within transcription factor binding sites (60). On the basis of limited experimental data, it is estimated that approximately one-third of these sites in human or rodents are not functional in the other species (61); a similar proportion is observed between two *Drosophila* species (62). Mammalian promoter and transcription start sites also appear to have been particularly prone to rapid evolution due to possible contributions from elevated mutation rates, reduced constraints, redundancy and positive selection (63,64).

If selection were often to be fleeting, rather than permanent, it would begin to explain the increasingly common identification of functional sequence that has not been conserved between diverse mammals. There are thousands of newly identified Piwi-interacting RNAs, for example, that are not conserved between mouse and more distant species (65). More generally, large numbers of non-coding sequences exhibit divergence levels, between mouse and either human or rat, that are similar to those of putatively neutral sequence (66–68).

If such sequences are indeed rapidly interchanging between neutrality and functionality, then our model organisms will not yield experimental findings on these sequences that are sufficiently relevant to human biology. Comparative genomics will remain central to the study of selection, but current evolutionary models and statistical techniques will need to be adapted to cope with transiently selected sequence. Moreover, the genomic data we currently have to hand will be too coarse-grained: we will need to determine the genome sequences of more nearly related species in order to investigate the more rapid fluctuations of selection relevant to the biology of our own species.

ACKNOWLEDGEMENTS

C.P.P. would like to thank Professor John Mattick (University of Queensland) for his generous hospitality during the writing of this review. We gratefully acknowledge the financial support of the UK Medical Research Council.

Conflict of Interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Lunter, G., Ponting, C.P. and Hein, J. (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.*, **2**, e5.
4. Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M. *et al.* (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, **302**, 842–846.
5. Mattick, J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet.*, **5**, 316–323.
6. Emison, E.S., McCallion, A.S., Kashuk, C.S., Bush, R.T., Grice, E., Lin, S., Portnoy, M.E., Cutler, D.J., Green, E.D. and Chakravarti, A. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, **434**, 857–863.
7. Naukkarinen, J., Gentile, M., Soro-Paavonen, A., Saarela, J., Koistinen, H.A., Pajukanta, P., Taskinen, M.R. and Peltonen, L. (2005) USF1 and dyslipidemias: converging evidence for a functional intronic variant. *Hum. Mol. Genet.*, **14**, 2595–2605.
8. Ye, S., Eriksson, P., Hamsten, A., Kurkinen, M., Humphries, S.E. and Henney, A.M. (1996) Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression. *J. Biol. Chem.*, **271**, 13055–13060.

9. Zito, F., Lowe, G.D., Rumley, A., McMahon, A.D. and Humphries, S.E. (2002) Association of the factor XII 46C>T polymorphism with risk of coronary heart disease (CHD) in the WOSCOPS study. *Atherosclerosis*, **165**, 153–158.
10. Li, W.H., Gojobori, T. and Nei, M. (1981) Pseudogenes as a paradigm of neutral evolution. *Nature*, **292**, 237–239.
11. Miyata, T. and Yasunaga, T. (1981) Rapidly evolving mouse alpha-globin-related pseudo gene and its evolutionary history. *Proc. Natl Acad. Sci. USA*, **78**, 450–453.
12. Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D. *et al.* (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, **13**, 13–26.
13. Mighell, A.J., Smith, N.R., Robinson, P.A. and Markham, A.F. (2000) Vertebrate pseudogenes. *FEBS Lett.*, **468**, 109–114.
14. Doxiadis, G.G., van der Wiel, M.K., Brok, H.P., de Groot, N.G., Otting, N., 't Hart, B.A., van Rood, J.J. and Bontrop, R.E. (2006) Reactivation by exon shuffling of a conserved HLA-DR3-like pseudogene segment in a New World primate species. *Proc. Natl Acad. Sci. USA*, **103**, 5864–5868.
15. Dunn, C.A., Medstrand, P. and Mager, D.L. (2003) An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc. Natl Acad. Sci. USA*, **100**, 12841–12846.
16. 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.
17. 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.
18. 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.
19. Duret, L. (2002) Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, **12**, 640–649.
20. Gaffney, D.J. and Keightley, P.D. (2005) The scale of mutational variation in the murid genome. *Genome Res.*, **15**, 1086–1094.
21. Ellegren, H., Smith, N.G. and Webster, M.T. (2003) Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.*, **13**, 562–568.
22. Cooper, D.N. and Youssoufian, H. (1988) The CpG dinucleotide and human genetic disease. *Hum. Genet.*, **78**, 151–155.
23. Fryxell, K.J. and Moon, W.J. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.*, **22**, 650–658.
24. Meunier, J. and Duret, L. (2004) Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.*, **21**, 984–990.
25. Green, P., Ewing, B., Miller, W., Thomas, P.J. and Green, E.D. (2003) Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.*, **33**, 514–517.
26. Majewski, J. (2003) Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am. J. Hum. Genet.*, **73**, 688–692.
27. Smith, N.G., Webster, M.T. and Ellegren, H. (2002) Deterministic mutation rate variation in the human genome. *Genome Res.*, **12**, 1350–1356.
28. Smith, N.G., Brandstrom, M. and Ellegren, H. (2004) Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics*, **84**, 806–813.
29. Clark, A.G. (2006) Genomics of the evolutionary process. *Trends Ecol. Evol.*, **21**, 316–321.
30. Kondrashov, F.A., Ogurtsov, A.Y. and Kondrashov, A.S. (2006) Selection in favor of nucleotides G and C diversifies evolution rates and levels of polymorphism at mammalian synonymous sites. *J. Theor. Biol.*, **240**, 616–626.
31. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
32. Jareborg, N., Birney, E. and Durbin, R. (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.*, **9**, 815–824.
33. Dermitzakis, E.T., Reymond, A. and Antonarakis, S.E. (2005) Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, **6**, 151–157.
34. Frazer, K.A., Sheehan, J.B., Stokowski, R.P., Chen, X., Hosseini, R., Cheng, J.F., Fodor, S.P., Cox, D.R. and Patil, N. (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res.*, **11**, 1651–1659.
35. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
36. Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S.E., Dermitzakis, E.T. *et al.* (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.*, **38**, 223–227.
37. Sironi, M., Menozzi, G., Comi, G.P., Bresolin, N., Cagliani, R. and Pozzoli, U. (2005) Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. *Trends Genet.*, **21**, 484–488.
38. Sironi, M., Menozzi, G., Comi, G.P., Cagliani, R., Bresolin, N. and Pozzoli, U. (2005) Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum. Mol. Genet.*, **14**, 2533–2546.
39. Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.*, **17**, 68–74.
40. Winter, E.E., Goodstadt, L. and Ponting, C.P. (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res.*, **14**, 54–61.
41. McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, **351**, 652–654.
42. Hurst, L.D. (2002) The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.*, **18**, 486–487.
43. Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S., Wiebe, V., Kitano, T., Monaco, A.P. and Paabo, S. (2002) Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature*, **418**, 869–872.
44. Evans, P.D., Anderson, J.R., Vallender, E.J., Gilbert, S.L., Malcom, C.M., Dorus, S. and Lahn, B.T. (2004) Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. *Hum. Mol. Genet.*, **13**, 489–494.
45. Evans, P.D., Gilbert, S.L., Mekel-Bobrov, N., Vallender, E.J., Anderson, J.R., Vaez-Azizi, L.M., Tishkoff, S.A., Hudson, R.R. and Lahn, B.T. (2005) Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science*, **309**, 1717–1720.
46. Hudson, R.R., Kreitman, M. and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.
47. Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. and Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.*, **30**, 233–237.
48. Olds, L.C. and Sibley, E. (2003) Lactase persistence DNA variant enhances lactase promoter activity *in vitro*: functional role as a *cis* regulatory element. *Hum. Mol. Genet.*, **12**, 2333–2340.
49. Wooding, S.P., Watkins, W.S., Bamshad, M.J., Dunn, D.M., Weiss, R.B. and Jorde, L.B. (2002) DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. *Am. J. Hum. Genet.*, **71**, 528–542.
50. Bamshad, M. and Wooding, S.P. (2003) Signatures of natural selection in the human genome. *Nat. Rev. Genet.*, **4**, 99–111.
51. Smith, A.V., Thomas, D.J., Munro, H.M. and Abecasis, G.R. (2005) Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.*, **15**, 1519–1534.
52. Keightley, P.D., Lercher, M.J. and Eyre-Walker, A. (2005) Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.*, **3**, e42.
53. Andolfatto, P. (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, **437**, 1149–1152.
54. Bierne, N. and Eyre-Walker, A. (2004) The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.*, **21**, 1350–1360.
55. Sawyer, S.A., Kulathinal, R.J., Bustamante, C.D. and Hartl, D.L. (2003) Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J. Mol. Evol.*, **57** (Suppl. 1), S154–S164.
56. Smith, N.G. and Eyre-Walker, A. (2002) Adaptive protein evolution in *Drosophila*. *Nature*, **415**, 1022–1024.
57. Welch, J.J. (2006) Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics*, **173**, 821–837.
58. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK.

59. Torarinsson, E., Sawera, M., Havgaard, J.H., Fredholm, M. and Gorodkin, J. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.*, **16**, 885–889.
60. 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.
61. Dermitzakis, E.T. and Clark, A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.
62. Costas, J., Casares, F. and Vieira, J. (2003) Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene*, **310**, 215–220.
63. 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.
64. Frith, M.C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Sandelin, A. (2006) Evolutionary turnover of mammalian transcription start sites. *Genome Res.*, **16**, 713–722.
65. 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.
66. Wang, J., Zhang, J., Zheng, H., Li, J., Liu, D., Li, H., Samudrala, R., Yu, J. and Wong, G.K. (2004) Mouse transcriptome: neutral evolution of ‘non-coding’ complementary DNAs. *Nature*, **431**, 1; page following 757, online.
67. Hyashizaki, Y. (2004) Mouse transcriptome: Neutral evolution of ‘non-coding’ complementary DNAs (reply). *Nature*, **431**, online.
68. 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.