Human Molecular Genetics, 2006, Vol. 15, No. 19 **2911–2922**
doi:10.1093/hmg/ddl232
Advance Access published on August 21, 2006

Conserved distances between vertebrate highly conserved elements

Hong Sun[1,†], Geir Skogerbø[1,†] and Runsheng Chen[1,2,*]

[1]Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics and [2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100101, P.R. China

Received June 1, 2006; Revised and Accepted August 11, 2006

High numbers of sequence element with very high (>95%) sequence conservation between the human and other vertebrate genomes have been reported and ascribed putative *cis*-regulatory functions. We have investigated the structural relationships between such elements in mammalian genomes and find that not only their sequences, but also the distances between them are significantly ($P < 2.2 \times 10^{-16}$) more conserved than corresponding distances between orthologous protein-coding genes or between exons within these genes. Regions of largely conserved distance between consecutive highly conserved elements (HCE) generally overlap previously identified HCE clusters, but may be far longer (up to 20 Mb) and possibly cover close to 25% of the human genome sequence. Similar conservation of distance is found between bird (chicken) and mammalian genomes and is also discernible in comparisons between fish and mammals. The data suggest either that a substantial amount of essential (functionally active) elements with lower sequence conservation occupy the space between the HCEs or that distance itself is an important factor in transcriptional regulation or chromatin modelling.

INTRODUCTION

Highly conserved sequences are present in the thousands in vertebrate genomes also outside protein-coding genes. More than 5000 ultraconserved elements (UCEs) longer than 100 bp and with absolute identity were found in human and rodent comparisons (1), and of these, 77% were located outside annotated exons. Long distance searches with slightly lower stringency detected 1400 (2) or 3500 (3) highly conserved elements (HCEs) between man and pufferfish. Recently, several classes of HCEs with origin in ancestral repeat sequences have also been identified in vertebrate genomes (4,5). Apart from some very few HCEs overlapping coding exons, vertebrate HCEs are not conserved in non-vertebrate animals (1), and within the insects (*Diptera*), the number of HCEs is also low (6).

In addition to the extreme sequence conservation, which surpasses that of most coding sequence and spans an evolutionary distance of at least 450 Myr (2), a number of other characteristics of vertebrate HCEs have also been described. Early observations of HCEs (7) occurred in the context of work on vertebrate *cis*-regulatory elements (cREs) of early/embryonic development genes, and in a number of cases regulatory modules controlling specific expression patterns of such genes have been found to be conserved from fish to man (8,9). Statistical analyses of genes associated with HCEs have also shown strong enrichment for certain functional categories, including genes involved in early embryo development and other transcription factors (1,3). Similarly, testing HCEs for enhancer activity in reporter gene systems revealed that a substantial fraction of the tested sequences were able to drive gene expression, not seldom in a fashion resembling that of their (assumed) associated genes (2).

No satisfactory explanation has been suggested to account for the extreme degree of sequence conservation of HCEs. Functions as *cis*-regulatory modules are not generally sufficient to explain the extreme conservation levels of HCEs (2,10). cREs are not necessarily strongly conserved and have been regarded as more 'evolvable' than coding sequence (11); furthermore, this evolvability has been called upon to explain morphological evolution as resulting from changes in regulatory rather than in protein-coding sequence (12,13). Transcription factor—DNA recognition modality has been described as degenerate in both directions, the specificity of

*To whom correspondence should be addressed at: Bioinfomatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Datun Road 15, Beijing 100101, P.R. China. Tel: +86 1064888546; Fax: +86 1064877837; Email: crs@sun5.ibp.ac.cn.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

transcriptional regulation being attributed as much to steric interactions between different transcription factors as to dependence of unique and invariable DNA sequences (14). What seems to be resolved, though, is that the conservation of these elements is due to selective pressure and not simply low local mutation rates (15). It has been frequently suggested that a strong sequence conservation could be due to the presence of two or more overlaid functional elements (1) (e.g. a regulatory element located within a coding exon); however, very few of such examples have been demonstrated (6,16,17), and none has come near explaining the existence of several thousand such elements in the vertebrate genomes. Suggestions have been put forward assuming that the existence of factors involved in controlling chromatin structure overlapping enhancer activity has been put forward (2,18), but no details as to how such a mechanism might work have been provided.

A commonly observed characteristic of HCEs is a strong tendency to occur in clusters along the chromosomes (1–3,19). The presence of such clusters might suggest that the order and orientation of the HCEs have been conserved among different species; however, no rigorous analysis of this aspect of the HCEs appears to have been carried out. The clusters are frequently found in relatively gene poor regions, straddling one or a few widely spaced genes (3,20), commonly of one of the categories mentioned above. These structures have generally been interpreted as clusters of cREs controlling one or several of the genes nearby or within the cluster (2,3,18). The clusters, however, can be quite wide, spanning several hundred kilobases, and in some cases even exceed a million base pairs (2,3).

Therefore, considerable distances may exist between the HCEs and the genes they are assumed to regulate (21). It has been suggested that the distance between an HCE and its target gene may be part of the regulatory mechanism itself (10); however, in a number of cases, genetic constructs in which the HCE has been brought into close proximity to the transcription start site have been shown to drive expression in a manner very similar to that of the endogenous promoter (2,22). Though subtle but important differences may exist between the expression patterns observed for an artificial construct and those of the endogenous gene, there is at present no strong evidence for a direct role of spacing in gene regulation, nor would such a role add much explanatory value to the level of sequence conservation of HCEs. For HCEs for which a likely HCE–gene association can be established, there seems to be a general conservation of HCE position and orientation relative to their putative target gene (23). The possibility therefore remains that the HCE clusters are imprints of other structures partly or fully overlapping elements with regulatory functions. We have therefore undertaken a structural analysis of the structural relationships among HCEs independently of whatever regulatory or other functions they may harbour.

RESULTS

We initiated our study by a re-analysis of the data originally produced by Bejerano *et al*. (1), with a particular focus on the genomic distribution of the UCEs. Whereas the study of Bejerano *et al*. (1) concentrated on 481 UCEs of a size larger than 200 bp in length, we have looked at the entire set of 5425 UCEs longer than 100 bp detected in the human–mouse and human–rat comparisons. We further used BLAT (24) to locate all occurrences of the same UCEs with $\geq$80% identity in dog and chicken and with $\geq$35% identity in zebrafish, *Tetraodon nigroviridis* and *Fugu rubripes*. The number of UCE that can be mapped to genomes of the different species is variable, but amounts to around 5000 for the mammals, 3712 for chicken and 500–700 for the fish. Different subsets of the UCEs are common to different sets of species, and 210 UCEs are common to all eight species. With respect to genomic location, 46% of the UCEs are found in an intergenic region, 31% are intronic and 23% are located in or overlap an exon (Table 1).

Table 1. Genomic locations of UCEs in the human genome

Location	Number of UCEs	Percentage
Exonic	1242	23
Intronic	1694	31
Intergenic	2489	46
Total	5425	100

To investigate the structural relationships of HCEs among vertebrates, we compared the UCEs with a set of 12 001 protein-coding genes (abbreviated as CDSs) which occur in a 1:1:1 relationship in the human, mouse and dog genomes (henceforth called the test gene set) (25). In general, one might expect that the relatively short UCEs would be more influenced by chromosomal micro-rearrangements and thus show more differences in order and orientation than protein-coding genes. We do not find any evidence in this direction. The relative order of UCEs along the chromosomes is 98.7% identical in the human, mouse and dog genomes, which is very similar to that of CDSs in the same genomes (97.7%; Supplementary Material, Table S1) (26). Likewise, for both UCEs and CDSs with identical order on the mammalian chromosomes, the orientation is reversed in about 0.5% of the cases, showing that the generally very short UCEs have apparently not suffered more micro-rearrangements than the much longer protein-coding genes. A further indication that HCEs could have a high preference for more stable genomic environments than coding genes may be inferred from a recent study on mammalian chromosomal evolution that found more than 40% increase in coding gene density within the surrounding 1 Mb of a breakpoint compared with the genome-wide average gene density average (27). A similar analysis of human–mouse–rat breakpoint regions showed that a 1 Mb window around the breakpoint in the human genome had 75% less UCEs (0.025 UCEs/Mb) than sequence within synteny blocks (0.105 UCEs/Mb, $P < 0.001$; see Materials and Methods).

Conservation of distances between HCEs

The number of analysed UCE pairs was only one-third that of CDS pairs, and the average distance between two consecutive

Table 2. Distances between consecutive UCE and CDS pairs in the human genome

	Number of pairs	Min (bp)	Median (bp)	Mean (bp)	Max (bp)
All UCE pairs	4 824	101	37 172	317 084	16 165 061
All CDS pairs	12 001	45	74 736	204 568	52 040 302
All EXON pairs	80 324	34	2 010	14 314	6 723 840

The distances were measured from midpoint to midpoint.

UCEs (315 kb) was thus higher than the average distance between two consecutive CDSs (205 kb; Table 2). However, the median distance for UCE pairs (47 kb) is much shorter than that for CDS pairs (72 kb), underscoring the tendency of UCEs to occur in clusters. To establish whether the distances between UCEs show less change than distances between other genomic elements, we calculated a relative distance difference (RDD; see Fig. 1A for definition) between pairs of UCEs found in adjacent positions in the human, mouse and dog genomes and compared these with RDD values calculated for pairs of CDS from the test gene set. A relative distance measure could relate the absolute distance difference between genetic elements to the distance in any of two compared genomes, or as we have done, to the average distance between the two. As we in our background analyses as well as in on-going analyses (data not shown) included a number of comparisons with no common reference genome, we found it more practical to use the RDD value defined in the legend of Figure 1A. However, we also calculated the distance differences using the human genome as reference genome, reaching the same conclusions as with the RDD measure (data not shown).

There were clear differences in RDD distribution for the two data sets (Fig. 1A and B). The RDD values for UCE pairs were mostly concentrated in a narrow peak close to zero (median RDD_{UCE} of -0.026 and -0.054 for the human–mouse and the human–dog comparisons, respectively). RDDs for CDS pairs, in contrast, showed a broader distribution skewed towards more negative values (median RDD_{CDS} of -0.168 and -0.157 for the human–mouse and the human–dog comparisons, respectively). The distribution of RDD values for pairs of exons with the CDSs was also wider than that of the UCE pairs, but with a peak much closer to zero (median RDD_{EXON} of -0.099 and -0.058 for the human–mouse and human–dog comparisons, respectively). The more negative RDD values for CDS pairs in all likelihood reflects the size differences between the three genomes and might suggest that the factors responsible for these differences (25) predominantly act on intergenic (as opposed to intronic) sequence not occupied by HCEs.

Calculated as absolute values (|RDD|), the average distance difference for UCE pairs in the human–mouse comparisons was 0.13 (corresponding to an average 12.4% change relative to the distance in the human genome) which is about half that for CDS pairs (0.26 or 26.2%) and exons (0.36 or 34.9%; Supplementary Material, Table S5A). The figures for the human–dog comparison are similar, but slightly less than that for the human–mouse comparison for all three types of genetic elements (Fig. 2). Correspondingly, 67.1 and 72.4% of all human–mouse and human–dog UCE pairs, respectively, had |RDD| values below 0.15, compared to only 39.4 and 45.0% of the CDS pairs from the same comparisons (Fig. 3).

If RDDs between genomic elements are generally lower for shorter absolute distances, the apparent conservation of distance between UCE pairs could be a result of a larger portion of these than of CDS pairs having rather short distances. We therefore calculated |RDD| values for UCE and CDS pairs within different distance intervals (as measured in the human genome) (Fig. 4 and Supplementary Material, Table S2). There is a clear tendency towards lower |RDD| values for the shortest absolute distances; however, this tendency is far stronger for UCE pairs than for CDS pairs. In the 0–50 kb distance interval, the average $|RDD|_{CDS}$ values were approximately five times higher than the average $|RDD|_{UCE}$ values, compared with only about two times higher when the entire data sets were compared. Nonetheless, |RDD| values for CDS pairs were significantly higher than that for UCE pairs for all distances up to 250 kb ($P < 0.001$; Supplementary Material, Table S2), and the difference in distance conservation cannot be explained by differences in absolute distances. Because a considerable fraction of the UCEs overlaps with exons and introns of coding genes (Table 1), we also suspected that the relative stability of UCE pair distances might simply be due to the relative constant length of the genes that they overlap with; however, comparisons showed that in most cases there are no consistent differences between intergenic UCE pairs and pairs with one or two intragenic UCEs (Supplementary Material, Table S3). The exception was pairs where both UCEs were located in an intron, which had significantly ($P < 0.04$, Wilcoxon's test) more conserved distances than most other UCE pair combinations. We assume that this may be due to the fact that quite a number of such pairs (42%) span exons, which are likely to have more conserved lengths than most other genomic elements.

As a considerable fraction of the CDS–CDS distances overlap to some extent with sequence between UCE pairs, the distribution of RDD values for all CDS pairs would not reflect what might be called the neutral distance change rate between genomic elements. We therefore calculated the distances for the 4221 CDS pairs that had no overlap with UCE–UCE distances. This gave a significantly higher average |RDD| value for both human–mouse (0.29) and human–dog (0.23) comparisons ($P < 1.0 \times 10^{-11}$; Wilcoxon's test) (Table 3), indicating that the neutral distance change rate might have an average |RDD| value which is above 0.2 at least. Similarly, calculating RDD values for only the intergenic part of the CDS–CDS distances (excluding CDS pairs with intervening non-test set genes) also gave much broader distribution profiles and considerably higher average

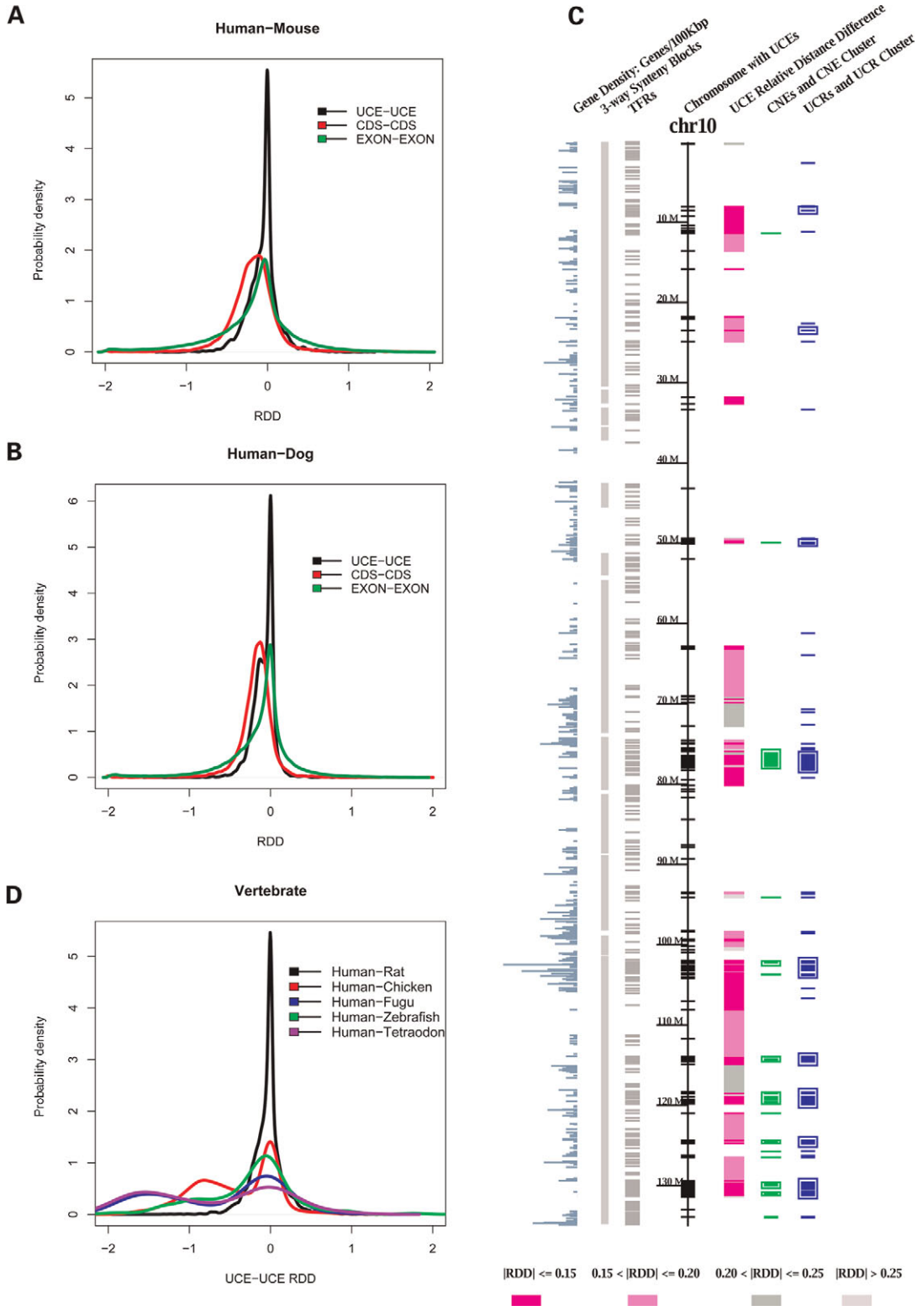

Figure 1. Conservation of distance between pairs of genetic elements in vertebrate genomes. (**A** and **B**) Distribution of RDD values calculated for pairs of UCEs, test set genes (CDS) and exons in the human and mouse (A) and human and dog (B) genomes. RDD was defined as RDD $= (d_q - d_h)/[(d_q + d_h)/2]$, d_q and d_h being the distance between the midpoints of two consecutive sequence element (CDSs, exons or UCEs) pairs in the query (non-human) and human genomes, respectively (see Materials and Methods for details). (**C**) Distance conservation between consecutive UCEs along human chromosome 10. The figure shows (from left to right) gene density (genes/100 kb), human–mouse–rat synteny blocks (28), TFRs (29), chromosome with UCEs indicated, UCE pairwise conserved distances (shades of red), CNEs (2) and CNE clusters (green rectangles) and UCRs (3) and UCR clusters (blue rectangles). For a full genomic view, see Supplementary Material, Figure S3. (**D**) Human–non-mammal UCE RDD distributions (human–rat included for comparison).

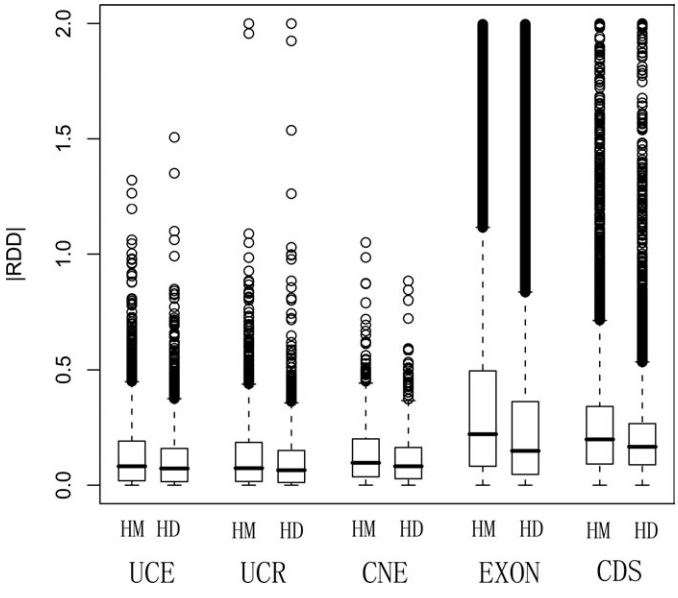

Figure 2. Absolute RDDs (|RDD|) between HCEs (UCEs, UCRs and CNEs), CDSs and exons in the human–mouse (HM) and human–dog (HD) comparisons. See Supplementary Material, Table S5A for further details.

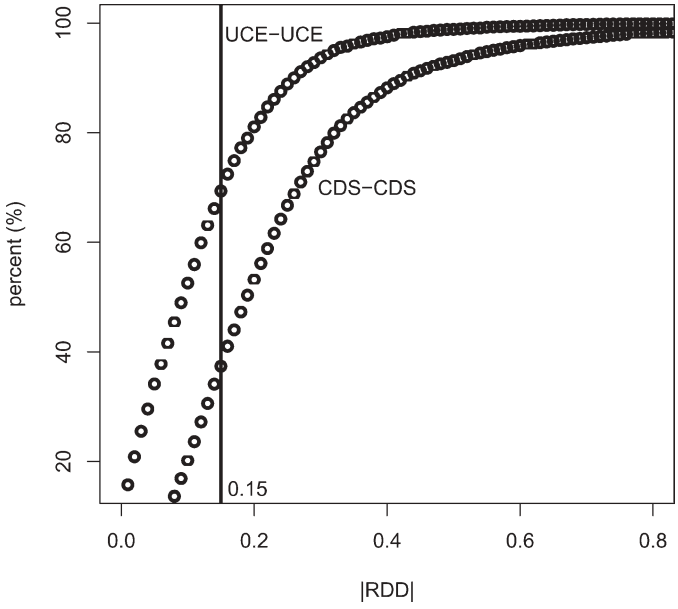

Figure 3. Cumulative plot (%) of UCE and CDS pairs with increasing |RDD| values.

distance differences (|RDD| = 0.48 and 0.37 for the human–mouse and human–dog comparison, respectively).

We also tested for HCE–HCE distance conservation using HCE data from two other studies (2,3) for HCE–HCE distance conservation. These two data sets were obtained using somewhat different length and conservation criteria, but should be comparable to the UCE data set. A direct comparison element by element shows that in the human genome only about 500 HCEs intersect with all three data sets, and of the 4183 non-exonic UCEs, two-thirds are not overlapped by HCEs from any of the two other data sets (Supplementary Material, Fig. S1). The smallest data set of about 1400 conserved non-coding elements (CNEs) (2) had the highest fraction of overlaps (~80%) with one or both of the other two sets, compared with about 50% for the set of ultraconserved regions (UCRs) (3). Despite the variable degree of overlap between the three data sets, the two additional data sets show approximately the same extent of HCE–HCE distance conservation among the mammals as found for the UCEs (Fig. 2), indicating that these three sets of HCEs essentially belong to the same population of sequence elements.

Blocks of conserved distance

To get a genomic perspective of the HCE distance conservation, we plotted UCE pairs with varying distance conservation along the human chromosomes (Fig. 1C; Supplementary Material, Fig. S2). Previous attempts to identify HCE clusters have been based on visual inspection alone (2) or in combination with nearest-neighbour analysis of consecutive UCR distances (3). The clusters identified by these two studies show a high degree of overlap (69/77%), but cover only a fraction of the most closely spaced HCEs. The full genome view of UCE pairwise distance conservation (Supplementary Material, Fig. S2) revealed a complex picture, of which a few details are seen along chromosome 10 (Fig. 1C). Blocks of consecutive UCE pairs with the most conserved distances (|RDD| < 0.20, indicated with red colours in Fig. 1C) intersect with 82 and 92% of CNE and UCR blocks, respectively, but differ from these both in extension and in continuity. Stretches of UCEs with strong distance conservation tend to include also far more widely spaced UCEs than previously identified HCE clusters and are commonly much longer than these, spanning several megabases of sequence. Altogether, UCE pairs with |RDD| < 0.20 cover 760 Mb or about one-fourth of the human genome. As expected, most block of conserved distance fall within established three-way synteny blocks (human–mouse–rat) (28); however, there appears to be no obvious relationship between blocks of conserved distance and synteny blocks. Gene density has been reported to be low in HCE-rich regions (1,3), and this also applies to regions with conserved distances (11.2 genes/Mb, Supplementary Material, Table S4), but not to HCE spans with higher |RDD| values (18.7 genes/Mb; compared with a genomic overall gene density of 15.5 genes/Mb).

One additional peculiarity is that regions overlapping previously identified denser HCE clusters frequently consisted of a mosaic of intermittent short distances with variable |RDD| values, often in conjunction with a longer span of more conserved distance (e.g. chr10, 100–110 Mb, in Fig. 1C). One explanation for this might be that the distance conservation observed between pairs of HCEs is actually the result of conserved distance to a more distant focal element (e.g. a target gene) or that what actually is conserved is the overall span of a larger chromosomal structure encompassing a number of HCEs. In both cases, a small insertion or deletion might a have a pronounced effect on the RDD between two closely spaced HCEs without having much effect on the length of the overall structure or the distance to a possible target gene.

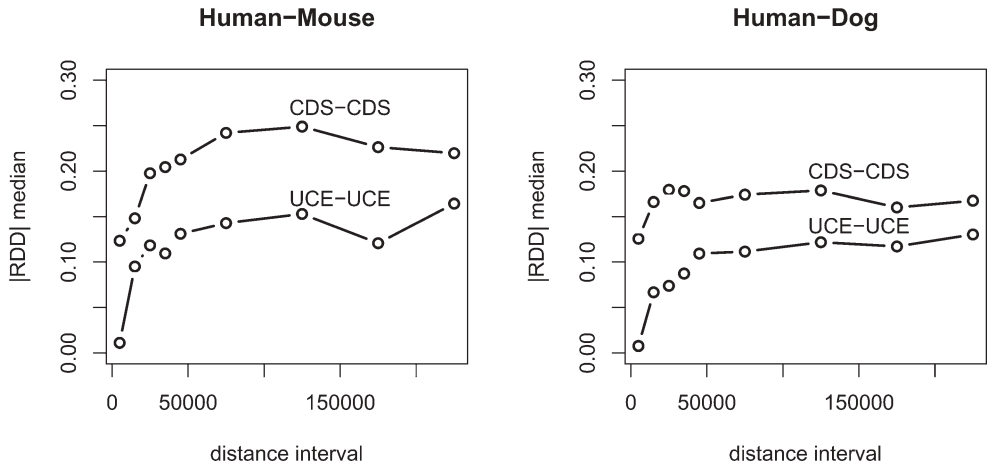

Figure 4. Absolute RDDs (|RDD|) as influenced by actual UCE and CDS pairwise distance. The figures show average |RDD| values for pairwise distances within 10 kb (0–50 kb range) and 50 kb (50–250 kb range) intervals for the HM and HD comparisons.

Table 3. RDDs for CDS pairs intersecting or not intersecting with UCE pairs

		Human−mouse				Human−dog			
		RDD		|RDD|		RDD		|RDD|	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
All CDS−CDS		−0.168	−0.177	0.199	0.256	−0.157	−0.176	0.167	0.210
CDS−CDS intersecting UCE pairs with |RDD| <0.15	A (1955 pairs)	−0.092	−0.106	0.125	0.166	−0.115	−0.123	0.122	0.144
CDS−CDS intersecting any UCE pair	B (7101 pairs)	−0.158	−0.164	0.182	0.234	−0.148	−0.163	0.157	0.191
CDS−CDS not intersecting any UCE pair	C (4921 pairs)	−0.189	−0.193	0.223	0.285	−0.170	−0.194	0.180	0.234
P(A,B)		2.2E−16		2.2E−16		2.2E−16		2.2E−16	
P(A,C)		2.2E−16		2.2E−16		2.2E−16		2.2E−16	
P(B,C)		1.0E−11		2.2E−16		2.2E−14		2.2E−16	

P(XY) denotes the significance level (two-sample Wilcoxon's test) for a comparison between X and Y.

HCE–gene associations

Possible associations between HCEs and nearby genes have been analysed by previous studies, which have all found an over-representation of gene functional categories involving nucleic acid binding, transcription regulation and early development (1–3). We therefore did not delve deeply into the matter of HCE–gene functional associations apart from comparing the nearest gene flanking or overlapping a UCE to the Uniprot database. This showed significant enrichment for genes in similar functional categories (i.e. transcriptional regulators and nucleotide and nucleic acid binding proteins) as previously reported ($P < 5.8 \times 10^{-5}$) (Supplementary Material, Table S6). A similar enrichment was also found for the 210 UCEs common to all eight vertebrates and for test set genes with |RDD| < 0.25. One deviation from the previous analyses was, however, that also genes coding for catalytic and ion-binding functions were significantly ($P < 7.5 \times 10^{-7}$) over-represented in the vicinity of UCEs, possibly suggesting that the relationship between UCEs and proximal genes may be more complex than previously supposed.

An additional question of interest is whether the conserved distances between HCEs can be extended to potential target genes or whether a potential HCE–gene relationship can be identified by the degree of distance conservation. An analysis of the genes manually associated with identified CNE clusters (2) showed that the distances between CNEs and these genes were generally almost as conserved as the HCE–HCE distances (data not shown). Further analysis of the general distance conservation between HCEs and surrounding genes is, however, complicated by the length of the HCE blocks with conserved distance, as these must necessarily cover numerous genes. If only UCE blocks with the most conserved distances are considered (i.e. |RDD| < 0.2), these alone cover altogether 798 annotated protein-coding genes, thus the number of potential target genes will be very high. The analysis is further complicated by the fact that comparison of distance conservation between HCEs and genes is generally only meaningful for a set of genes with only one orthologue in each of the compared genomes (i.e. the test gene set). An analysis of the distances between a UCE and the nearest test genes on either side shows that these distances are not significantly less conserved than the distances between UCEs in general (Wilcoxon's test, $P < 1.0 \times 10^{-5}$). We also asked whether the number of other non-test genes located between a HCE and the nearest test gene influences the distance conservation between the two, but found no significant increase

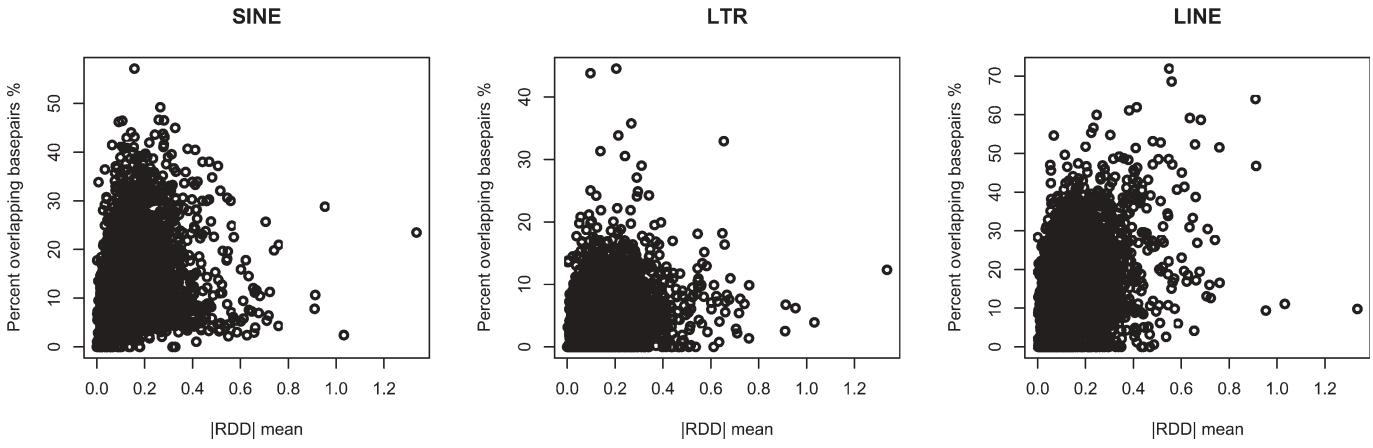

Figure 5. Transposon insertion and |RDD| values for UCE pairs. The figure shows the percentage of UCE–UCE distance base pairs that overlap with short interspersed elements (SINEs), LTRs and LINEs.

Table 4. Correlation coefficients (Spearman's rho statistics) between |RDD| values and transposon insertion frequency (number/Mb) or fraction (% of bp) for UCE pairs in the human genome

	Number/Mb		% of bp	
	Correlation coefficient	*P*-value	Correlation coefficient	*P*-value
All	0.679	2.2E−16	0.749	2.2E−16
SINEs	0.606	2.2E−16	0.622	2.2E−16
LTRs	0.609	2.2E−16	0.609	2.2E−16
LINEs	0.648	2.2E−16	0.668	2.2E−16

in RDD for up to five intervening genes (Supplementary Material, Table S7). Analysis of the distances to genes positioned outside the most conserved UCE block also indicated that the distance conservation commonly extends to the one or two nearest test genes, raising the number of potential targets even further. Therefore, although these data do not in any way exclude the possibility that the HCEs in the longest distance-conserved blocks are functionally related to only one or a few genes, distance conservation alone will probably not be sufficient to identify these genes (23).

RDD is related to transposon density and primary sequence conservation

It would appear likely that the RDD between a pair of HCEs in two genomes is related to the insertion rate of transposable elements and we therefore calculated the frequency of three different transposon families in the intervening distance between two UCEs. There is a general trend towards a lower fraction of transposon base pairs in sequence with low |RDD| values ($R = 0.75$, $P < 2.2 \times 10^{-16}$) (Fig. 5 and Table 4). When all UCE pairs are considered, there are small differences between the different transposon classes with respect to their contribution to |RDD| variation (Table 4); however, when only UCE pairs with intervening transposons are considered, their correlations are stronger for the longer transposon types [i.e. long terminal repeats (LTRs) and long interspersed elements (LINEs)] (Supplementary Material, Table S8). Twenty-seven percent of the UCE pairs contain no intervening transposons; these pairs are generally short (median length 466 bp) with very low RDD values (|RDD| = 0.017–0.024) (Supplementary Material, Table S9). We also tested the relationship between regions of conserved distance and the recently identified long transposon-free regions (TFRs), which have been associated with both protein-coding genes and UCEs (29). Of the 5425 UCEs, 1951 (36%) overlap with 1121 of the 9249 TFRs longer than 5 kb. However, TFRs are significantly enriched in sequence with both well-conserved (|RDD| < 0.15) and less well-conserved (|RDD| > 0.25) distance ($P < 0.001$) (Table 5; Supplementary Material), indicating a complex relationship between TFRs and distance conservation. Further analyses showed that any association between TFRs and distance-conserved sequence is almost exclusively limited to TFRs overlapping UCEs (21% of TFRs overlap one or more UCEs) and appears not to be a result of strong association with distance-conserved sequence *per se* (Fig. 6; see Supplementary Materials for details). TFRs would in any case only overlap 3.6% of the sequence with |RDD| < 0.15 (Table 5) and thus have little explanatory value for the vast majority of the distance conserved sequence. The lack of association between TFR and distance conservation is difficult to explain, but the pronounced enrichment of UCE base pairs in TFRs might be a hint that absolute absence of transposons is a characteristic of functional primary sequence (conserved or not) rather than of conservation of distance.

The relationship between distance conservation and primary sequence conservation will necessarily be quite complex because, on the one hand, only about 5% (∼150 Mb) of the human genome is conserved among mammals (25,30), whereas sequence with conserved distance may cover up to 25% of the human genome, and, on the other hand, the spans covered by a UCE–UCE pair are as complex as the genome itself, covering all aspects of both genic and intergenic sequence. Nevertheless, to obtain an estimate of the relationship between distance conservation and primary sequence conservation, we assigned PhastCons scores (31) to all base pairs spanned by UCE pairs. The data indicated a significant negative correlation between the average PhastCons scores and the |RDD| values ($R = -0.75$, $P < 2.2 \times 10^{-16}$) (Supplementary

Table 5. Degree of overlap between base pairs in TFRs and UCE–UCE pairs of varying distance conservation

UCE pairwise distance conservation	UCE–UCE distance (Mb)	TFR–UCE pair overlap (Mb)	% of UCE–UCE distance	% of TFR base pairs	*P*-value
|RDD| < 0.15	394	14.1	3.6	35.7	<0.001
|RDD| > 0.25	134	11.1	8.3	28.1	<0.001
All UCE pairs	1530	39.5	2.6	100.0	
Genome	2900	65.7			

The *P*-value is estimated from a simulation of assignment to the human genome of a set of random regions of similar size and number as the set of TFRs.

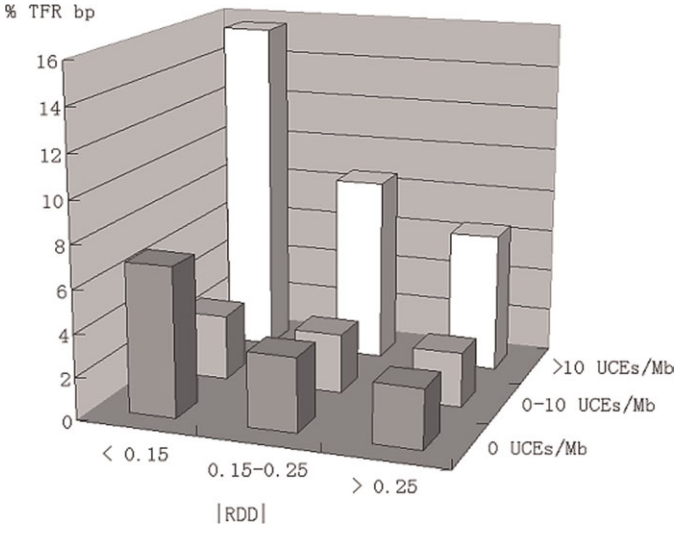

Figure 6. Relationship between TFR overlap, |RDD| values and UCE density for CDS–CDS pairs. The figure shows the average percentage of TFR (>5 kb) base pairs that overlap the distances between CDS–CDS with different |RDD| values and UCE densities.

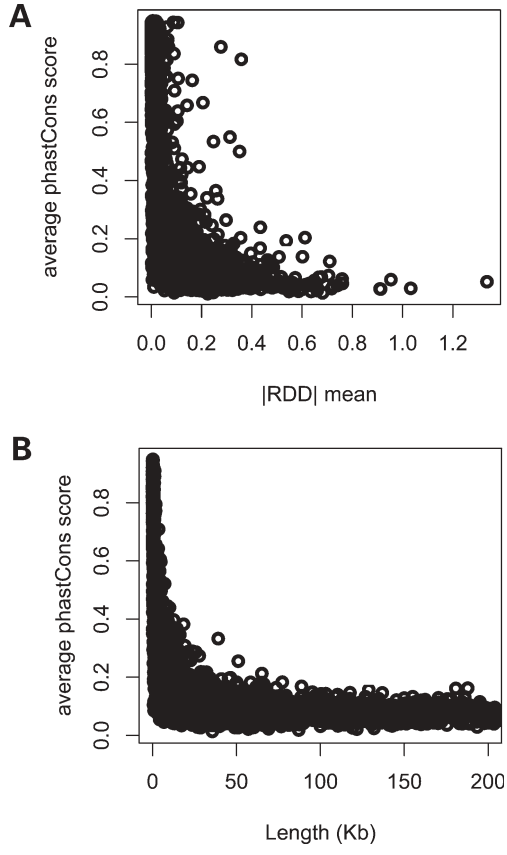

Figure 7. Correlation between average UCE–UCE PhastCons scores and (**A**) |RDD| values and (**B**) absolute UCE–UCE distances. |RDD|mean is the average |RDD| value from the HD and HM comparisons.

Material, Table S11) and even stronger correlation between PhastCons score and absolute UCE–UCE distance ($R = -0.89$, $P < 2.2 \times 10^{-16}$). We suspected that both correlations might be heavily influenced by the extreme sequence conservation of the UCEs themselves and repeated the calculation after removing PhastCons score for the UCE sequence. This reduced the correlation coefficients between PhastCons score and |RDD|, and between PhastCons score and UCE–UCE distance, to -0.51 and -0.24, respectively (Fig. 7; Supplementary Material, Table S11). Both correlations are statistically significant ($P < 2.2 \times 10^{-16}$), possibly indicating a residual influence of genic sequence (exons, conserved UTRs, etc.) residing between the shortest UCE pairs. Sequence with higher PhastCons scores is almost entirely limited to inter-UCE distances below 2 kb, and for UCE pairs with spans above this distance (i.e. 81.5% of all UCE pairs comprising 99.97% of UCE–UCE distances), PhastCons scores are all below 0.5, and there is only a weak correlation ($R = 0.40$, $P < 2.2 \times 10^{-16}$) between sequence and distance conservation. This accords with the observation that UCE pairs located in introns, which are generally shorter and tend to span exons, show somewhat higher PhastCons scores than both exonic and intergenic pairs ($P < 10^{-4}$) (Supplementary Material, Table S12). However, whereas some influence of conserved genic sequence is likely, this does not exclude the possibility that the weak but statistically significant correlation between conservation of distance and primary sequence found for longer UCE–UCE spans could indicate that functions residing in the primary sequence may have a small but important contribution to the observed conservation of distance.

Other vertebrates

Finally we asked whether the conservation of distance between pairs of HCEs would also be found between evolutionary more distant vertebrates. Extending the analysis to non-mammalian is hampered by lower numbers of common

UCE pairs as well as by the lack of a suitable set of 1:1 orthologous genes to serve as a standard. The average |RDD| for adjacent UCE pairs in the human and chicken genomes was 0.22, ~60% larger than |RDD| values among the mammals for the same UCE set (Supplementary Material, Table S13). For human–fish comparisons, RDDs could only be calculated for around 100 pairs between any two genomes, and the average |RDD| values (0.235–0.732) were also much higher than the within-mammal values (Supplementary Material, Table S14). A characteristic of the human–non-mammalian comparisons is that the RDD distributions show distinct two-peak profiles, with one peak close to zero and another peak (or 'shoulder') at a more negative value. Given the persistent nature of the two-peak distribution profiles (found in all mammal–non-mammal comparison for all sets of HCEs) (Fig. 1D), it is difficult to imagine that this kind of profile could arise by a random assortment of RDD values. Rather, it seems more likely that a subset of HCE pairs exists with RDD values close to zero and whose distances have been conserved between evolutionary very distant vertebrates. The majority of these low RDD values correspond to rather short distances (<40 kb), perhaps suggesting that distance conservation over very long evolutionary distances may be restricted to closely spaced HCEs; however, the low number of UCEs behind these data implies that this will need further investigation.

DISCUSSION

The data give few clues to their biological significance and can be interpreted in several ways, which are not necessarily mutually exclusive. One possible interpretation would be that conserved distance is actually an indicator of the presence of functional sequence with lower sequence conservation occupying the space between HCEs. This would be in accordance with a recent study that indicated that intron length is correlated to its content of significantly conserved sequence (and not, for instance, to transcriptional economy) (32). The lower gene density in sequence spans with conserved distance might likewise be a consequence of this sequence already harbouring functionally active elements. The existence of a weak but significant correlation between distance and sequence conservation might also be taken as evidence in this direction.

The alternative is that conserved distance is indeed indicative of precise spacing of functional elements relative to each other (or to a common focus, e.g. a regulated gene) being an important aspect of gene or chromatin structure. In either case, there is also the problem of defining what actually 'conserved distance' is, as this aspect of comparative genomics has not been much studied. A conservative approach would be to regard only distances with |RDD| < 0.15 as genuinely conserved, as these fall safely below the average gene–gene distance |RDD| of ~0.25. This might mean that real conservation of distance acts mostly on closely spaced HCE-embedded cREs or between cREs and nearby target genes. In turn, this type of distance conservation might not amount to much more than maintaining cREs within operating ranges commonly found between such elements and their target genes (11). The fact that a large fraction of TFRs are both relatively short (on average <15 kb) and low in conserved primary sequence could also indicate a spacer function (29). An interpretation along these lines is also supported by the (still very limited) data on fish–mammal distance conservation, which appears to be almost exclusively limited to shorter HCE–HCE distances. It also makes good sense in the light of what is known about *cis*-regulation in the well-studied *Drosophila* in which cREs are generally located within a few kilobases of their target genes, but it accords less well with the recent findings of *cis*-regulatory activity apparently conserved (21) and acting (9,33) over spans of up to 1 Mb of sequence or more.

However, the idea that HCEs are merely well-conserved cREs is in itself problematic. One reason is that *cis*-regulatory function alone is not sufficient to explain the levels and extension of sequence conservation of HCEs (100% for up to several hundred base pairs), because as a rule cREs are rather short and commonly far more malleable to sequence changes than, for instance, protein-coding sequence (11,12,14). Sequence conservation does not even appear to be necessary for preservation of function, as human cREs are able to reproduce expressional patterns of corresponding zebrafish cREs, despite all recognizable sequence similarity between the two having vanished (34). Nor has the distance between cREs and their target genes been shown to be of crucial importance to transcriptional regulation, and very distant cREs are able to reproduce (at least partially) the expression patterns of the endogenous target gene even when placed within short distance of the basal promoter of a reporter gene (9). Considerable variation in transcription start positions between orthologous genes in the human and mouse genomes also argues against spacing between cREs and core promoters being of crucial importance for transcriptional control (35). Though the sensitivity of cRE-reporter gene constructs may still be limited and fail to detect more subtle variation in gene expression patterns, there is at present no convincing argument implicating cRE order and distance as crucial factors in their function as transcriptional regulatory elements, particularly not on the distance scale of tens of megabases. Therefore, though there is accumulating evidence in favour of *cis*-regulatory activity embedded in HCEs, their overall sequence and structural conservation is not easily explained by the current understanding of cRE function.

The less conservative interpretation is that, as indicated by the data, the neutral distance change rate between genomic elements is considerably higher (i.e. |RDD| ~ 0.25–0.45) and that |RDD| values up to 0.20–0.25 may be indicative of conserved distance and long contiguous HCE-containing structures. This, however, expands the size range for such structures to more than 20 Mb for the longest structures, which is 10–20 times the longest span between HCEs and their putative target genes indicated by recent reports (21,23). The possibility of the HCEs being components of conserved structures on this scale further challenges the idea of a simple *cis*-regulatory relationship between HCEs and a small set of target genes. This is not to say that large-scale ordered structures may not be important in transcriptional regulation. The conserved order of homeobox genes across most of *Metazoa* is perhaps the most prominent example, though the actual link between the conserved gene order and, in these complexes, their spatio-temporal expression still remains obscure (36).

Apparent regular spacing of eukaryotic genes regulated by common transcription factors (37) may have been an artefact of spacing on the chip (38), but the existence of large domains of co-expressed genes in eukaryotic genomes appears nonetheless to be a fact, and long distance coordination of gene expression is best explained in terms of effects of chromatin structure (39). Although there is no apparent regularity in HCE spacing, the extreme distances over which order and distances of HCEs are conserved might indicate a role in chromatin modelling for these large HCE structures.

In conclusion, our analysis indicates that in addition to an extreme level of sequence conservation, HCEs also display strong conservation of mutual distances among vertebrate species. This conservation of distance appears to extend to all sequenced vertebrate genomes and, at least among mammalian genomes, may cover up to several tens of megabase of sequence. Overlapping, multiple functions have been suggested by several studies (2,19) to account for the extreme sequence conservation of HCEs; however, there is no consensus as to what these multiple functions might be. A tempting model is the one that would combine features of long-range chromatin modelling with more local *cis*-regulatory and/or other functions. It is conceivable that chromosomal regions might contain specific 'anchor points' with generally moderate sequence similarity, which are required for attaining one or more chromatin structures necessary for correct spatio-temporal expression of genes or clusters of genes. These 'anchor points' may in general not be sufficiently conserved to be recognized in alignments between different species, but might become visible whenever they overlap with one or more other functional elements (e.g. protein-coding sequence, cREs or other). The lack of similarity among HCEs (and thus 'anchor points') within a genome is difficult to account for, but the fact that function may be conserved *in lieu* of discernible sequence conservation suggest that there may exist aspects of conserved primary sequence information that are not picked up by available bioinformatics tools (34). These long-ranging chromatin modelling structures may have originated with (and/or have been a pre-requisite for) vertebrate development, during when a number of the anchor point sequences with multiple overlapping functions became frozen in evolutionary time. Later vertebrate evolution has modelled further on these structures, added some HCEs, discarded others, but kept sufficiently many to leave a mark that is visible in all studied vertebrates.

MATERIALS AND METHODS

Data

The sets of 5425 UCEs (1) and 1373 conserved non-coding elements (2) were obtained directly from the respective authors. The UCR data set was obtained from http://mordor.cgb.ki.se/cgi-bin/SCRbrowse/c (3). The TFR (>5 kb) data set was obtained from http://jsm-group.imb.uq.edu.au/tfr/ (29). The complete genome sequences for the eight species were downloaded from the UCSC website (http://hgdownload.cse.ucsc.edu/goldenPath/): Human (NCBI Build 34, hg16); Mouse (NCBI Build 30, mm3); Rat (rn3, Baylor HGSC v. 3.1); Dog (canFam1, Broad Institute v. 1.0); Chicken (February 2004, galGal2); Fugu (August 2002, fr1); Zebrasfish (June 2004, danRer2); Tetraodon (February 2004, tetNig1). We got the human–mouse–dog orthologous test gene set from the website http://www.broad.mit.edu/mammals/dog/ (25), the 391 three-way human–mouse–rat synteny blocks from Bourque *et al*. (28) and four-way human–mouse–rat–chicken synteny blocks from Bourque *et al*. (40). The collections of annotated genes in the human, mouse and rat genomes, the hg16 mm3rn3 PhastCons score files and transposon annotation files for the human genome were downloaded from the GoldenPath database (http://hgdownload.cse.ucsc.edu/goldenPath/).

HCEs and test genes (CDSs) were aligned against the genomes using BLAT (24) with default parameters. Among the resulting alignments for each sequence, only the hit with best alignment identity was kept. The data obtained from the various studies were derived from different version of the human genome sequence and had to be mapped onto a common version for comparison, thus the total number of HCEs, TFRs, synteny blocks and CDSs differ slightly from the figures published by the original studies.

Assignments of homologous elements pairs

Two HCEs or CDSs were regarded as a pair if they were found as neighbours in the genomes of both (or all) the species compared. We got 12 001 CDS–CDS pairs, 4824 UCE–UCE pairs, 1064 CNE–CNE pairs and 2658 UCR–UCE pairs which occur in a 1:1:1 relationship among man, mouse and dog.

Conservation of UCE and gene order

The measure of UCE or gene order conservation between two genomes used here is the ratio between the number of UCEs or genes linked in conserved runs and the total number of related UCEs or genes (26).

UCE density within three-way synteny blocks and evolutionary break regions

Using the coordinates of the midpoint of synteny blocks and inter-synteny block regions in the human genome, we created discrete intervals of 1 Mb windows surrounding the midpoints and counted the number of UCEs. As a test of the statistical significance of our results, we randomly selected the same number of windows in the three-way synteny blocks and the evolutionary break regions. To evaluate the statistical significance, each analysis was repeated 1000 times with independent, randomly sampled data sets. The fraction of times in which the random sample set average UCE density was lower than the average density of UCEs located in inter-synteny block regions provided the basis for the statistical significance.

Calculation of distance differences between pairs of genomic elements

To investigate the conservation of distances between the HCEs and CDSs pairs, we used two relative distance measures. One, termed 'RDD', relates a difference in sequence length between a pair of genetic elements in two different genomes to the

average sequence length between the pair of elements, i.e. $RDD = (d_q - d_h)/[(d_q + d_h)/2]$, d_q and d_h being the distance between the midpoints of two consecutive sequence element pairs in the query (non-human) and human genomes, respectively. The other measure calculates the difference in sequence length as a fraction of the length in one of the two genomes (in practice, the human genome), thus termed 'percent of human (%oH)', defined as $\%oH = 100\times\ (d_q - d_h)/d_h$. In principle, the measures are identical. All basic calculations were carried out using both measures, giving similar conclusions; however, for simplicity, the data presented in the text are all based on the RDD measure.

RDD distributions were drawn using R function Density (41). The algorithm used in computing probability density disperses the mass of the empirical distribution function over a regular grid of at least 512 points, then uses the fast Fourier transform to convolve this approximation with a discretized version of the kernel and then linear approximation to evaluate the probability density at the specified points. The corresponding frequency of UCE (or CDS and exon) pairs within any RDD interval can thus be approximated as the size of the interval times the 'average' probability density for the interval.

We tested the null hypothesis that there is no difference in the RDD values between HCE pairs and CDS pairs, or between HCE and exon pairs, using the Wilcoxon's unpaired test, which is appropriate for RDDs of these groups of pairs because of their highly skewed distribution.

Gene ontology annotation analysis

We compared gene ontology (GO) annotations of genes associated with UCEs in the human genome against the background of all annotated human genes, using the tail of the hypergeometric distribution to calculate *P*-values and adjusted for multiple testing using the FDR method (42). GO molecular function analysis was performed by using the GOMINER tool (43). Statistical analyses were carried out using the R language and software (41).

Correlations between transposon density, TFRs and RDD value

Spearman's rho statistic method was used to estimate a rank-based measure of association between RDD values and transposon frequency or fraction.

We randomly selected same number of regions with same size as TFRs (>5 kb) as a negative control to evaluate the enrichment of TFRs in UCE distance-conserved regions. Each analysis was repeated 1000 times with independent, randomly sampled data sets.

Sequence conservation in the UCE–UCE distance-conserved regions

Spearman's rho statistic method was used to measure relationship between PhastCons score and UCE–UCE lengths and |RDD| values. Wilcoxon's test was used to estimate the significance of PhastCons score for sequences between UCE pairs with different genomic location.

All algorithms used will be made available on request.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

We thank Gill Bejerano and Adam Woolfe for kindly making available the UCE and CNE data sets and Lisa Caviglia for careful correction of the manuscript. We also thank two anonymous reviewers for a number of constructive suggestions. This work was supported by the National High Technology Development Program of China under Grant No. 2002AA231031, National Key Basic Research & Development Program (973) under Grant Nos 2002CB713805 and 2003CB715907.

Conflict of Interest statement. The authors declare no conflict of interest.

REFERENCES

1. 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.
2. Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. *et al*. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
3. Sandelin, A., Bailey, P., Bruce, S., Engstrom, P., Klos, J., Wasserman, W., Ericson, J. and Lenhard, B. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.
4. Kamal, M., Xie, X. and Lander, E.S. (2006) A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci. USA*, **103**, 2740–2745.
5. Xie, X., Kamal, M. and Lander, E.S. (2006) A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl Acad. Sci. USA*, **103**, 11659–11664.
6. Glazov, E.A., Pheasant, M., McGraw, E.A., Bejerano, G. and Mattick, J.S. (2005) Ultraconserved elements in insect genomes: A highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.*, **15**, 800–808.
7. Shashikant, C.S., Kim, C.B., Borbely, M.A., Wang, W.C.H. and Ruddle, F.H. (1998) Comparative studies on mammalian Hoxc8 early enhancer sequence reveal a baleen whale-specific deletion of a *cis*-acting element. *Proc. Natl Acad. Sci. USA*, **95**, 15446–15451.
8. Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
9. Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, **12**, 1725–1735.
10. Gomez-Skarmeta, J.L., Lenhard, B. and Becker, T.S. (2006) New technologies, new findings, and new concepts in the study of vertebrate *cis*-regulatory sequences. *Dev. Dyn.*, **235**, 870–885.
11. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.
12. Phinchongsakuldit, J., MacArthur, S. and Brookfield, J.F.Y. (2004) Evolution of developmental genes: molecular microevolution of enhancer sequences at the ubx locus in drosophila and Its impact on developmental phenotypes. *Mol. Biol. Evol.*, **21**, 348–363.
13. Carroll, S.B. (2000) Endless forms: The evolution of gene regulation and morphological diversity. *Cell*, **101**, 577–580.
14. Mainguy, G., In der Rieden, P.M.J., Berezikov, E., Woltering, J.M., Plasterk, R.H.A. and Durston, A.J. (2003) A position-dependent organisation of retinoid response elements is conserved in the vertebrate Hox clusters. *Trends Genet.*, **19**, 476–479.

15. 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.
16. Neznanov, N., Umezawa, A. and Oshima, R.G. (1997) A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *J. Biol. Chem.*, **272**, 27549–27557.
17. Sandrelli, F., Campesan, S., Rossetto, M.G., Benna, C., Zieger, E., Megighian, A., Couchman, M., Kyriacou, C.P. and Costa, R. (2001) Molecular dissection of the 5′ region of no-on-transientA of drosophila melanogaster reveals *cis*-regulation by adjacent dGpi1 sequences. *Genetics*, **157**, 765–775.
18. de la Calle-Mustienes, E., Feijoo, C.G., Manzanares, M., Tena, J.J., Rodriguez-Seguel, E., Letizia, A., Allende, M.L. and Gomez-Skarmeta, J.L. (2005) A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.*, **15**, 1061–1072.
19. Boffelli, D., Nobrega, M.A. and Rubin, E.M. (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.*, **5**, 456–465.
20. 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.
21. Vavouri, T., McEwen, G.K., Woolfe, A., Gilks, W.R. and Elgar, G. (2006) Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends Genet.*, **22**, 5–10.
22. Lettice, L.A., Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N. *et al*. (2002) Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA*, **99**, 7548–7553.
23. McEwen, G.K., Woolfe, A., Goode, D., Vavouri, T., Callaway, H. and Elgar, G. (2006) Ancient duplicated conserved noncoding elements in vertebrates: A genomic and functional analysis. *Genome Res.*, **16**, 451–465.
24. Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
25. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., III, Zody, M.C. *et al*. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.
26. Tamames, J. (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol.*, **2**, RESEARCH0020.
27. Murphy, W.J., Larkin, D.M., der Wind, A.E.-v., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L. *et al*. (2005) Dynamics of mammalian chromosome rvolution inferred from multispecies comparative maps. *Science*, **309**, 613–617.
28. Bourque, G., Pevzner, P.A. and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, **14**, 507–516.
29. Simons, C., Pheasant, M., Makunin, I.V. and Mattick, J.S. (2006) Transposon-free regions in mammalian genomes. *Genome Res.*, **16**, 164–172.
30. Rat Genome Sequencing Project Consortium (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
31. 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.
32. Vinogradov, A.E. (2006) 'Genome design' model: evidence from conserved intronic sequence in human–mouse comparison. *Genome Res.*, **16**, 347–354.
33. Bishop, C.E., Whitworth, D.J., Qin, Y., Agoulnik, A.I., Agoulnik, I.U., Harrison, W.R., Behringer, R.R. and Overbeek, P.A. (2000) A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse. *Nat. Genet.*, **26**, 490–494.
34. 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.
35. Frith, M.C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayshizaki, Y. and Sandelin, A. (2006) Evolutionary turnover of mammalian transcription start sites. *Genome Res.*, **16**, 713–722.
36. Kmita, M. and Duboule, D. (2003) Organizing axes in time and space; 25 years of colinear tinkering. *Science*, **301**, 331–333.
37. Kepes, F. (2003) Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J. Mol. Biol.*, **329**, 859–865.
38. Lercher, M.J. and Hurst, L.D. (2006) Co-expressed yeast genes cluster over a long range but are not regularly spaced. *J. Mol. Biol.*, **359**, 825–831.
39. Hurst, L.D., Pal, C. and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.*, **5**, 299–310.
40. Bourque, G., Zdobnov, E.M., Bork, P., Pevzner, P.A. and Tesler, G. (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.*, **15**, 98–110.
41. Ihaka, R. and Gentleman R. (1996) R: A language for data analysis and graphics. *J. Comput. Graph. Statist.*, **5**, 299–314.
42. Storey, J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Stat.*, **31**, 2013–2035.
43. Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.