Evolutionary Discrimination of Mammalian Conserved Non-Genic Sequences (CNGs)

Emmanouil T. Dermitzakis,1,6 Alexandre Reymond,1† Nathalie Scamuffa,1 Catherine Ucla,1 Ewen Kirkness,2 Colette Rossier,1 Stylianos E. Antonarakis1,*

Analysis of the human and mouse genomes identified an abundance of conserved non-genic sequences (CNGs). The significance and evolutionary depth of their conservation remain unanswered. We have quantified levels and patterns of conservation of 191 CNGs of human chromosome 21 in 14 mammalian species. We found that CNGs are significantly more conserved than protein-coding genes and noncoding RNAs (ncRNAs) within the mammalian class from primates to monotremes to marsupials. The pattern of substitutions in CNGs differed from that seen in protein-coding and ncRNA genes and resembled that of protein-binding regions. About 0.3% to 1% of the human genome corresponds to a previously unknown class of extremely constrained CNGs shared among mammals.

Until recently, the extent of nucleotide conservation between human and other mammalian species has been unclear. Small-scale analyses between human and mouse genomes suggested conservation outside of gene regions (1–6). Comparison with the draft of the mouse genome indicated that at least 5% of the human genome was under selective constraint; surprisingly, the majority of these highly conserved sequences did not correspond to known genetic sequences, and experimental attempts to test the hypothesis that they are previously unidentified genes showed that this is unlikely (7–10). In addition, a method was recently described for the identification of primate-specific functional elements (11). Computational and mathematical efforts have attempted to distinguish the conserved regulatory portion of the genome from neutrally evolving sites (12). However, no highly accurate methodology that can discriminate between different functional classes of highly conserved sequences has been developed.

In this report, we analyze 220 sequences of the 2262 CNGs initially identified as highly conserved between human chromosome 21 and mouse syntenic regions and presented no evidence for transcription potential (7). We subsequently compared their evolutionary properties with protein coding sequences (CODs) from past studies (13–15) and noncoding RNA gene sequences (ncRNAs) obtained here. To perform polymerase chain reaction (PCR) from genomic DNA of green monkey, ring-tailed lemurs, brush-tailed porcupine, rabbit, pig, cat, greater mouse-eared bat, white-toothed shrew, nine-banded armadillo, African elephant, tammar wallaby, and platypus, we designed oligonucleotides on CNG and ncRNA human sequences in highly conserved regions between human and mouse. The selection of ncRNAs has its basis in criteria of orthology and sufficient conservation to design primers. Only a small subset of known ncRNAs could be used because of characteristics such as antisense to genes, small size, and unknown function.

After PCR, we obtained at least one sequence from the other 12 species eligible to human and mouse for 191 out of 220 CNGs (87%) and 14 out of 16 ncRNAs (88%). The 19 nuclear protein-coding genes had been analyzed previously (15); we aligned 12 of the 44 original species (human, strepsirrhine, mouse, hystricid, rabbit, pig, cat, free-tailed bat, shrew, armadillo, elephant, and opossum). In that study, CODs were chosen to have 80 to 95% nucleotide identity between human and mouse (14), and they were selected from a larger set because a PCR product could be obtained from all 44 species (15). Therefore, these sequences are biased for high success of amplification in other species and high conservation, issues that become relevant below. Our analyses were performed in multiple alignments of 55,519 base pairs (bp) of CNGs, 17,028 bp of CODs, and 5599 bp of ncRNAs.

To minimize biases from missing data resulting from PCR failure, we considered two CNG data sets, one of all 191 CNGs (CNG-all, fig. S1A) and another of 63 CNGs for which the sequences of at least 10 species were available, including at least one of armadillo, elephant, wallaby, or platypus (16). This second data set (CNG-high, for high species coverage; fig. S1B) is directly comparable to the CODs, which contain all 12 species’ sequences. With the use of the same criteria, we considered the complete data set for all 14 ncRNAs (ncRNA-all, fig. S1C) and a subset of 5 ncRNAs with high alignment coverage (ncRNA-high, fig. S1D). Both data sets of CNGs and ncRNAs (all and high) are used below to illustrate that the missing data do not influence the observed patterns.

A large fraction of the 191 successfully amplified CNGs were highly conserved in multiple mammalian species (fig. S1; A, B, and E). Specifically, we could retrieve more than 43% of the orthologous sequences from wallaby and/or platypus. High sequence conservation was evident even in the absence of species-specific substitution biases [e.g., A-T to G-C bias in mouse, porcupine, rabbit, and elephant (17)] that increase the substitution rate, providing additional support for the significant role of CNGs. The divergence values of CNGs were much lower than those of CODs and ncRNAs for each species pair (Fig. 1 and table S1), illustrating strong selective constraint.

To quantify the levels of conservation, we estimated the amount of sequence divergence per unit of evolutionary time. For each of the 191 CNGs, 14 ncRNAs, and 57 CODs (18), we calculated sequence change per million years (D/my) assuming the phylogenetic tree described in (13, 19). Ancestral states were derived with maximum likelihood with the use of PAML3 (20), and inferred substitutions were placed on the branches of the phylogenetic tree to account for all detectable substitution events. Divergence times were derived from (19). We calculated the sequence change for each tree branch and divided by the number of millions of years each branch covered. Figure 2A shows that CNG-all and CNG-high are significantly more constrained than CODs, ncRNA-all, and ncRNA-high. These observations are not a result of amplification bias, because multiple species sequences for CODs were obtained with stricter criteria (see above) than CNGs and ncRNAs. The low D/my values of CNGs show that they are under a stronger selective pressure than other functional genomic elements. To confirm that the higher substitution rate of CODs is not an artifact of the selection of the CNGs, we performed a
similar analysis by searching for all the Hsa21 2262 CNGs and 1229 CODs identified in (7) in the 1.5X genome of the dog (Canis familiaris) available from TIGR (18). For the set of 2262 CNGs, 1674 (74%) had a reciprocal best dog hit (E value < 0.001), and 1406 (62%) satisfied additional criteria of at least 90% coverage and at least 70% nucleotide identity. For the set of 1229 CODs, 994 (81%) had a reciprocal best dog hit (E < 0.001), and 749 (61%) satisfied additional criteria of at least 90% coverage and at least 70% nucleotide identity. This result, together with the fact that we expect to find by chance about 70% of any sequence in a 1.5X genome, suggests that the vast majority of CNGs are conserved in dog and likely in many placental mammals. We subsequently aligned 1638 CNGs and 976 CODs in human, mouse, and dog and compared their dog-specific divergence (Fig. 2B). CNGs showed a significantly lower rate of substitution than CODs, confirming the result with multiple species.

We conclude that a large fraction of these CNGs, originally found conserved between human and mouse, are highly conserved in multiple mammals, strongly supporting functional importance. The CNGs studied here represent 10% of the total number of CNGs on Hsa21. Even if only the CNG-high set (29% of the whole) can be considered functionally important, there are at least 656 such highly conserved elements (CNGs) on Hsa21 and at least 65,600 in the human genome, twice as many as the genes (Hsa21 is ~1% of the human genome). Moreover, the 2262 CNGs of Hsa21 cover 1% of the Hsa21 sequence, and the CNG-high constitutes 29% of this 1%. Therefore, we estimate that at least 0.3% of the Hsa21 sequence (~90 kbp) or of the whole human genome (~9 Mbp) is under very strong selective pressure so that there is minimal sequence change across hundreds of millions of years of evolutionary time. In addition, the fact that there is extensive conservation of CNGs in the dog strongly supports the idea that the majority of CNGs are conserved in most placental mammals.

In order to identify characteristics that could distinguish CNGs from other functional genomic elements, we devised three metrics that are described below. CODs tend to have a uniform distribution of substitutions, because most third positions of the codons are free to change (silent changes). In protein-binding DNA regions, substitutions usually occur between highly conserved binding sites (21–23). For ncRNAs, we have no prior evidence for one or the other pattern. To derive a measure for the distribution and clustering of variable sites within the sequence, we used a modified method from (24) to infer significance (18). For each of the 191 CNGs, 14 ncRNAs, and 57 CODs, we calculated the P value of clustering of variable sites along the sequence. Figure 3A illustrates the highly significant separation of P values of CNG-all and CNG-high from the CODs, whereas the ncRNAs covered have a wide distribution. As expected, the majority of CODs have significantly uniform distribution of substitutions along the sequence. In contrast, many of the CNGs have statistically significant clustering, strongly suggesting the presence of motifs for protein-binding or other interactions.

In constrained sequences, there are recurrent substitutions in the nucleotide positions that are free to evolve. When the entirety of the sequence is constrained, we observe sparse events of substitution, because almost none of the nucleotides are neutral. The former case is true for CODs because about one-third of the substitutions lead to a silent change, whereas regions with a high density of binding sites will resemble the latter pattern. The ncRNAs may have either pattern depending on the fraction of nucleotides that are functional. The average number of substitutions per variable site was calculated for each sequence and corrected for the substitution rate in the sequence (18). This correction was done to exclude the effect of global selective constraint within the sequence and to consider a normalized estimate of the number of substitutions per variable site. CNG-all and CNG-high have significantly smaller residual values than CODs (Fig. 3B). This suggests that even the fraction of variable nucleotides in CNGs are more constrained than that in CODs.

One of the properties that distinguishes a transcribed sequence (COD and ncRNA) from a nontranscribed one (CNG) is that the function of the former is expressed in one of the two strands whereas for the latter both strands may be important. Therefore,
selection may be acting in only one strand for the CODs and ncRNAs, which could be detected by asymmetries of substitutions (e.g., A→T compared with T→A). Such asymmetry has been shown to exist, and selection due to transcription is one explanation (25, 26). We quantified this asymmetry for CNGs, CODs, and ncRNAs with substitutions that maintain the G + C content (A→T compared with T→A and C→G compared with G→C), because G + C content may be under different selective forces. We first counted the number of A→T substitutions compared with T→A and C→G compared with G→C in the phylogenetic tree for each of the sequences. We then calculated the probability of the data assuming a binomial distribution and obtained probabilities for the AT and CG asymmetries. The AT asymmetry (Fig. 3C) was more pronounced in CNG-high, ncRNA-all, and CODs than in CNG-all and CNG-high, illustrating that transcription generates a preferential accumulation of substitutions in one strand. The CG asymmetry (Fig. 3D) was stronger in CODs than in both CNGs and ncRNAs, indicating a protein-coding-specific bias for this type of substitutions.

The results presented here demonstrate that a large fraction of CNGs on Hsa21 belong to a distinct class of highly constrained functional sequences. At least 29% of them were highly conserved in multiple mammalian species as distant as human, mouse, pig, elephant, wallaby, and platypus and generally more conserved than protein-coding and ncRNA sequences. High levels of conservation of almost all CNGs in dog suggest that the majority of CNGs are conserved in placental mammals. CNGs also have characteristics typical of protein-binding regions with alternating clusters of high- and low-constraint nucleotides, suggesting that some of them are indeed protein-binding and likely regulatory regions. Functional analysis of CNGs will require extensive protein-binding assays, reporter construct experiments, mouse knockouts, and other intensive experimental efforts. Nevertheless, understanding the role of CNGs in genome function and regulation and their involvement in phenotypic variation and human diseases should be a high priority in future genomic studies.

References and Notes
3. I. Dubchak et al., Genome Res. 10, 1304 (2000).
12. L. Elnitski et al., Genome Res. 13, 64 (2003).
16. Characteristics are theoretically influenced by the alignment coverage. Therefore, we consider both CNG and ncRNA data sets (CNG-all, CNG-high, ncRNA-all, and ncRNA-high). Comparisons involving CNG-high and ncRNA-high are statistically sound, whereas comparisons with CNG-all and ncRNA-all are indicative.
17. E. T. Dermitzakis et al., unpublished data.
18. Materials and methods are available as supplemental material on Science Online.
27. We are grateful to M. Caselli, F. M. Catzeflis, D. L. Dittmann, J. A. Marshall Graves, M. Ruedi, F. Shel- don, P. Vogel, C. Wenker, and M. Westerman for DNA samples; to W. Murphy and S. O’Brien for sharing sequence data; and to M. Chapuisat and D. Sanchez Ruiz for help and advice. This project was supported by the Fonds National de la Recherche Scientifique (Switzerland), NCCR Frontiers in Genetics, the European Union, and the ChildCare and Lejeune Foundations. Tissues were provided by Institut des Sciences de l’Evolution, Montpellier University; the Swiss National Foundation; the Basel Zoo; the Museum of Natural History, Louisana State University; the Comparative Genomics Research Group, Australian National University; the Department of Genetics, La Trobe University; the Basel Zoo; the Museum of Natural History of Geneva; and the Institute of Ecology, University of Lausanne.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1087047/DC1
Materials and Methods
SOM Text
Figs. S1 to S4
Tables S1 to S3
References
21 May 2003; accepted 22 September 2003
Published online 2 October 2003; 10.1126/science.1087047
Include this information when citing this paper.