16. L. Shen, K. L. Rock, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3035 (2004).
17. S. P. Schoenberger *et al.*, *J. Immunol.* **161**, 3808 (1998).
18. M. L. Albert, B. Sauter, N. Bhardwaj, *Nature* **392**, 86 (1998).
19. M. Bellone *et al.*, *J. Immunol.* **159**, 5391 (1997).
20. J. W. Yewdell, C. C. Norbury, J. R. Bennink, *Adv. Immunol.* **73**, 1 (1999).
21. A. Serna, M. C. Ramirez, A. Soukhanova, L. J. Sigal, *J. Immunol.* **171**, 5668 (2003).
22. N. P. Restifo *et al.*, *J. Immunol.* **154**, 4414 (1995).
23. We thank B. Buschling, D. Tokarchick, and A. Schell for technical assistance. We are grateful to M. Epler and S. Tevethia for their generous gift of D^b-$NP_{366-374}$ tetramers. This work was supported in part by a Wellcome Prize Traveling Fellowship and U.S. Public Health Service grants, and NIH grant AI-056094-01 to C.C.N.

Supporting Online Material
www.sciencemag.org/cgi/content/full/304/5675/1318/DC1
Materials and Methods
Figs. S1 and S2
References and Notes

3 February 2004; accepted 14 April 2004

Ultraconserved Elements in the Human Genome

Gill Bejerano,[1]* Michael Pheasant,[3] Igor Makunin,[3] Stuart Stephen,[3] W. James Kent,[1] John S. Mattick,[3] David Haussler[2]*

There are 481 segments longer than 200 base pairs (bp) that are absolutely conserved (100% identity with no insertions or deletions) between orthologous regions of the human, rat, and mouse genomes. Nearly all of these segments are also conserved in the chicken and dog genomes, with an average of 95 and 99% identity, respectively. Many are also significantly conserved in fish. These ultraconserved elements of the human genome are most often located either overlapping exons in genes involved in RNA processing or in introns or nearby genes involved in the regulation of transcription and development. Along with more than 5000 sequences of over 100 bp that are absolutely conserved among the three sequenced mammals, these represent a class of genetic elements whose functions and evolutionary origins are yet to be determined, but which are more highly conserved between these species than are proteins and appear to be essential for the ontogeny of mammals and other vertebrates.

Although only about 1.2% of the human genome appears to code for proteins (*1*–*3*), it has been estimated that as much as 5% is more conserved than would be expected from neutral evolution since the split with rodents, and hence may be under negative or "purifying" selection (*4*–*6*). Several studies have found specific noncoding segments in the human genome that appear to be under selection, using a threshold for conservation of 70 or 80% identity with the mouse over more than 100 base pairs (bp) (*7*–*13*). A study of these elements on human chromosome 21 found that those that were very highly conserved in multiple species contained significant numbers of noncoding elements (*13*). Similar results were found when comparing the human, mouse, and rat genomes (*14*, *15*) in a study of the 1.8–megabase (Mb) *CFTR* region (*16*, *17*) and in a functional study of the *SIM2* locus in a number of mammalian species (*18*).

We determined the longest segments of the human genome that are maximally conserved with orthologous segments in rodents: those showing 100% identity and with no insertions or deletions in their alignment with the mouse and rat. Exclusive of ribosomal RNA (rRNA) regions, there are 481 such segments longer than 200 bp that we call ultraconserved elements (table S1). They are widely distributed in the genome (on all chromosomes except chromosomes 21 and Y) and are often found in clusters (Fig. 1). The probability is less than 10^{-22} of finding even one such element in 2.9 billion bases under a simple model of neutral evolution with independent substitutions at each site, using the slowest neutral substitution rate that is observed for any 1-Mb region of the genome (supporting text, section S1). Nearly all of these elements also exhibit extremely high levels of conservation with orthologous regions in the chicken genome [467 out of 481 (467/481) = 97% of the elements aligning at an average of 95.7% identity, 29 at 100% identity] and about two-thirds of them with the fugu genome as well (324/481 = 67.3% of the elements aligning at an average of 76.8% identity), despite the fact that only about 4% of the human genome can be reliably aligned to the chicken genome (at an average of 62.9% identity where an alignment is found), and less than 1.8% of the human genome aligns to fugu (at an average of 60% identity). In addition, nearly all exhibit extremely high levels of conservation with the dog genome, which was estimated using reads from the National Center for Biotechnology Information (NCBI) trace archive (477/481 = 99.2% of the elements aligning at an average of 99.2% identity).Thus, it appears that nearly all of these ultraconserved elements may have been under extreme negative selection in many species for more than 300 million years, and some of them for at least 400 million years.

As expected, the ultraconserved elements exhibit almost no natural variation in the human population. Only 6 out of 106,767 bases examined in the ultraconserved elements (excluding the first and last 20 bases in each element) are at validated single-nucleotide polymorphisms (SNPs) in NCBI's SNP database (dbSNP) (table S2a). For this much DNA, we would have expected 119 validated sites, so validated SNPs are underrepresented by 20-fold ($P < 10^{-42}$). The 48 unvalidated SNPs we found revealed many likely errors in the unvalidated portion of dbSNP (table S2b). These same 106,767 bases exhibit very few differences with the chimp genome as well, showing only 38 single base changes where the chimp base has a Phred quality score at least 45, whereas the expected number would be 716 (roughly a 19-fold reduction, $P < 10^{-200}$; supporting text, section S2). This low level of variation within the human population and in comparison with the chimp suggests that these elements are currently changing at a rate that is roughly 20 times slower than the average for the genome. Only 4.3% of the bases are different in the chicken, which is also consistent with a roughly 20-fold reduction from neutral substitution rates (supporting text, section S2).

Of the 481 ultraconserved elements, 111 overlap the mRNA of a known human protein-coding gene [including the untranslated regions (UTRs)], 256 show no evidence of transcription from any matching expressed sequence tag (EST) or mRNA from any species, and for the remaining 114 the evidence for transcription is inconclusive. We call these partly exonic (or exonic for short), nonexonic, and possibly exonic ultraconserved elements, respectively. A hundred nonexonic elements are located in introns of known genes and the rest are intergenic. The non-exonic elements, both intronic and intergenic, tend to congregate in clusters near transcription factors and developmental genes, whereas the exonic and possibly

[1]Department of Biomolecular Engineering, [2]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA. [3]ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia.

*To whom correspondence should be addressed. E-mail: jill@soe.ucsc.edu (G.B.); haussler@soe.ucsc.edu (D.H.).

exonic elements are more randomly distributed along the chromosomes (Fig. 1).

There are 93 known genes that overlap with exonic ultraconserved elements; we call these type I genes. The 225 genes that are near the non-exonic elements we call type II genes (methods in supporting text, section S3). We looked for categories of biological process and molecular function defined in the Gene Ontology (GO) database (*19*) that are significantly enriched in type I and II genes and also searched InterPro (*20*) for enrichment in particular structural domains (Fig. 2). The type I genes show significant functional enrichment for RNA binding and regulation of splicing ($P < 10^{-18}$ and 10^{-9}, respectively, against all GO annotated human genes) and are uniquely abundant in the RNA recognition motif RRM ($P < 10^{-17}$, against all InterPro annotated human genes). In contrast, the type II genes are devoid of enrichment for RNA binding or splicing or the RRM ($P =$ 0.39, 0.44, and 0.77, respectively). However, type II genes are strongly enriched for regulation of transcription and DNA binding ($P < 10^{-19}$ and 10^{-14}, respectively), as well as DNA binding motifs, in particular the Homeobox domain ($P < 10^{-14}$). These three attributes are enriched in type I genes as well but 16, 8, and 9 orders of magnitude less significantly, respectively. This suggests that exonic ultraconserved elements may be specifically associated with RNA processing and non-exonic elements with regulation of transcription at the DNA level.

Non-exonic ultraconserved elements are often found in "gene deserts" that extend more than a megabase. In particular, of the non-exonic elements, there are 140 that are more than 10 kilobases (kb) away from any known gene, and 88 that are more than 100 kb away. The set of 156 annotated genes that

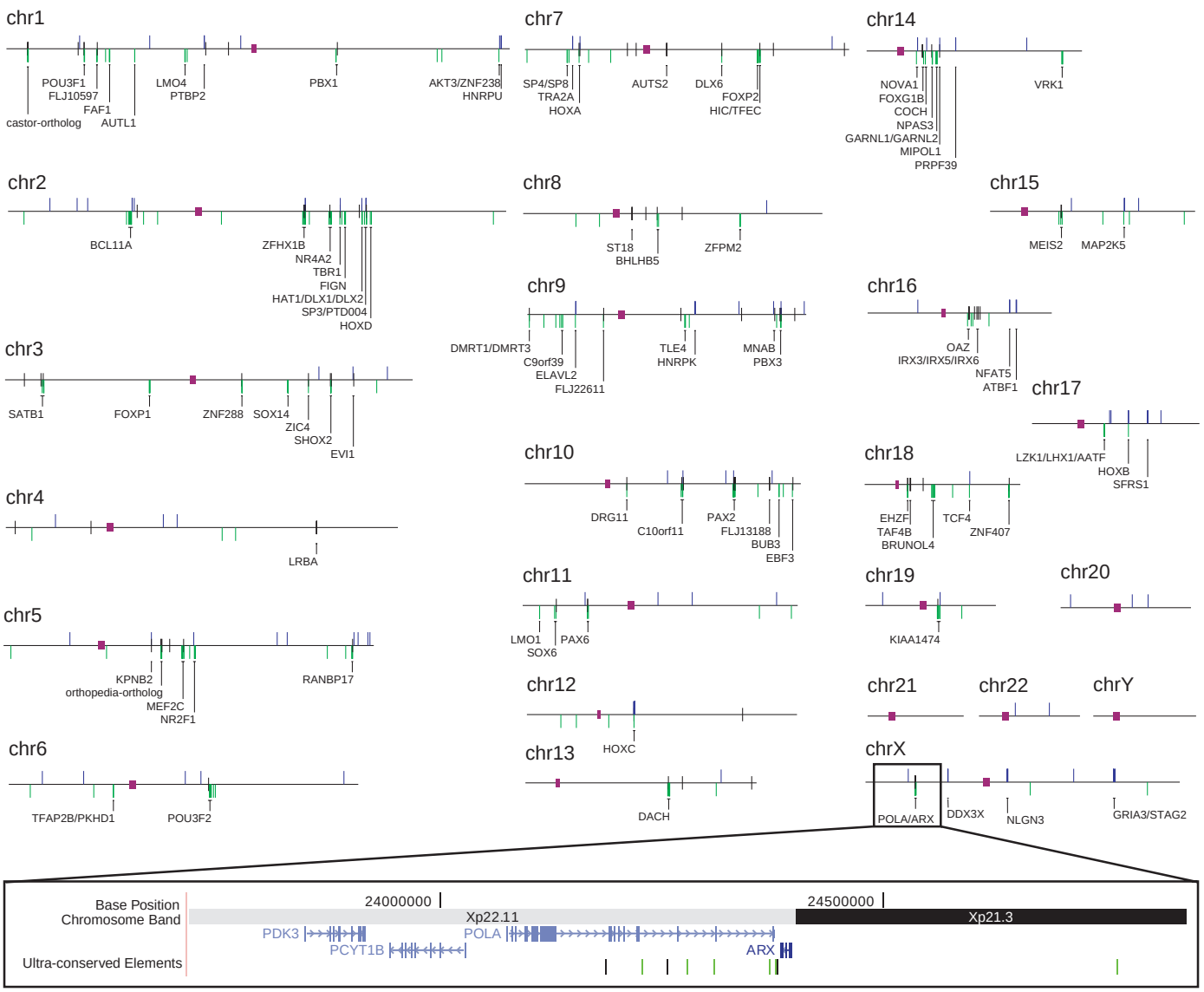

Fig. 1. Locations of the 481 ultraconserved elements on the 24 human chromosomes. Each partly exonic element is represented by a thin blue tick mark extending above the chromosome, each non-exonic element by a green tick mark extending below the chromosome, and each possibly exonic element by a black tick mark centered on the chromosome. Purple boxes represent centromeres. By joining two elements into a cluster when they are separated by less than 675 kb, we obtained 89 local clusters of two or more elements, each of which is boxed and named. Names are taken from a prominent gene or gene family co-located with the cluster or from a *Drosophila* ortholog or mRNA entry if no Human Genome Organization (HUGO)–named gene was available. Among the cluster representatives, there is a distinct enrichment for non-exonic elements and for developmental genes, suggesting that many of these clusters may be part of distal enhancers or "global control loci" analogous to those studied in association with *HOXD* (*38*) or *DACH* (*21*). One possible such cluster, near the *ARX* gene, is shown in more detail in the inset at the bottom of the figure. There known genes are shown in blue (tall boxes for coding exons, shorter boxes for UTRs, and hatched lines for introns), and ultraconserved elements are shown below them.

flank intergenic ultraconserved elements is significantly enriched for developmental genes ($P < 10^{-6}$) and in particular for genes involved in early developmental tasks ($P = 2.7 \times 10^{-5}$), suggesting that many of the associated ultraconserved elements may be distal enhancers of these early developmental genes. Indeed, one of these elements (uc.351 in table S1) is contained in an enhancer situated about 225 kb upstream of *DACH* (homolog of the *Drosophila* dachshund gene, known to be involved in the development of brain, limbs, and sensory organs), which has been shown to reproducibly drive expression in the retina when cloned upstream of a mouse heat shock protein 69 minimal promoter coupled to β-galactosidase and injected into a mouse oocyte (*21*). Non-exonic ultraconserved elements that lie in introns are also often associated with developmental genes. These include the neuroretina-specific enhancer in the fourth intron of *PAX6* (uc.328), investigated in quail but shown to also be functionally conserved in mouse (*22*).

Type I genes (harboring exonic elements) include many genes encoding well-known RNA-binding proteins, such as *HNRPK*, *HNRPH1*, *HNRPU*, *HNRPDL*, *HNRPM*, *SFRS1*, *SFRS3*, *SFRS6*, *SFRS7*, *SFRS10*, *SFRS11*, *TRA2A*, *PCBP2*, and *PTBP2*. All of the above are among the 59 type I genes annotated by GO that exhibit clear mRNA/EST evidence of alternative splicing overlapping the ultraconserved element [out of 66 elements in all, from a total of 111 exonic elements (table S3)]. Many of the above, including the six members of the *SFRS* family, contain the RNA recognition motif. The ultraconserved elements associated with alternative splicing events often contain small coding exons that are skipped in the mRNA in some tissues, but the elements extend well into the flanking intronic regions on one or both sides of the exon. Such is the case for one explicitly studied ultraconserved element (uc.33) in *PTBP2*, a polypyrimidine tract–binding protein (*23*). *PTBP2* contains a 312-bp ultraconserved segment that is mostly intronic but includes a small (34-bp) exon that is included in the mRNA only in brain tissue. The 203 bases at the 3′ end of the element, including the 34-bp exon, are 100% conserved in the chicken as well.

The *PTBP2* element may form an RNA structure in the pre-mRNA that participates in the regulation of splicing through interactions with the spliceosome (*23*). We used the program RNAfold (*24*) to further assess the potential of this and other ultraconserved elements to form an RNA secondary structure, comparing the energy of the best folded structure for both the positive and negative strand element to that of 10,000 random permutations of the same sequence (table S4). No statistically significant structure was found for the *PTBP2* element, but the energy of the fold for the 573-bp ultraconserved self-regulated alternatively-spliced UTR element (uc.189) in arginine/serine-rich splicing factor *SFRS3* (*25*) was lower than that of all but one of the 10,000 randomized versions of this sequence, indicating that it may form an important RNA secondary structure (fig. S2).

In addition to alternative splicing, the exonic ultraconserved elements also include the consecutive mutually exclusive "flop" and "flip" exons (uc.478/9) from the glutamate receptor *GRIA3* (*26*), which exhibits RNA editing as well as alternate splicing (*27*). The "flop" ultraconserved element extends into the ~600-bp intron 13 of the gene. At the other end (adjacent to the previous exon), intron 13 contains a much shorter highly conserved RNA hairpin structure that guides the essential and highly regulated editing of adenosine to inosine (*27*). Although the element containing the "flop" exon does not have detectable RNA secondary structure preferences, the minimal energy of the secondary structure of the element containing the "flip" is less than that of 34 out of 10,000 permuted versions, indicating possible structure.

Although the minimal region of 100% conservation between human, mouse, and rat that was required to be included in the ultraconserved set was 200 bp, many elements were considerably longer. The longest elements (779, 770, and 731 bp) all lie in the last three introns in the 3′ portion of *POLA*, the DNA polymerase alpha catalytic subunit (EC 2.7.7.7) on chromosome X, along with other shorter ultraconserved elements (Fig. 1). A similar-sized conserved region, 711 bp formed by concatenation of uc.468 and uc.469 (separated by a single base), lies in the ~7-kb intergenic region between the 3′ end of *POLA* and its downstream neighbor, the *ARX* homeobox gene. *ARX* is involved in central nervous system development and is associated with a host of X-linked Mendelian diseases, including epilepsy, mental retardation, autism, and cerebral malformations (*28*). Because this group of elements lies at the 3′ end of the 303-kb *POLA* gene, nearer to the 3′ end of *ARX* than to the rest of *POLA* (Fig. 1), it is possible that their function is not related to *POLA* but that they instead form a cluster of enhancers of *ARX*. The longest of these ultraconserved elements, 779 bp, is actually adjacent to a 275-bp element, which together form a 1046-bp region with only one change in rodents. As a calibration, note that these *POLA*/*ARX* elements are considerably longer than the ultraconserved portions of the human, mouse, and rat rRNA genes, which harbor six ultraconserved segments, three each in the 18*S* and 28*S* rRNA genes, the longest of which is 563 bp (table S1).

In sharp contrast to rRNA and most human coding regions, there were only 24 out of 481 cases (5%) where an ortholog of an ultraconserved element could be partially traced back by sequence similarity search as far as *Ciona intestinalis*, *Drosoplila melanogaster*, or *Caenorhabditis elegans* (table S7). All of these were among the 68 elements (14%) that overlapped coding exons from

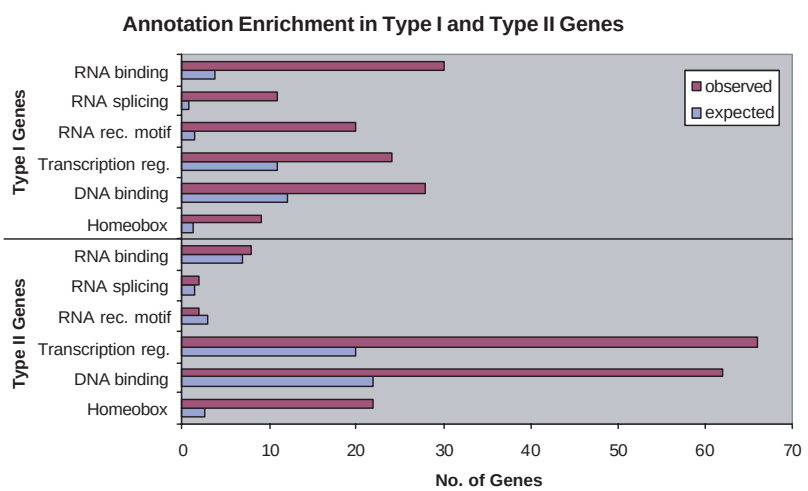

Fig. 2. Annotation enrichment in type I and type II genes. In the top half of the figure, the maroon bars ("observed") give the numbers of type I genes that are annotated in the GO database (*19*) with molecular function "RNA binding" or "DNA binding" or biological process "RNA splicing" or "transcription regulation," or are annotated in InterPro (*20*) as containing the domains "RNA recognition motif" or "homeobox." The blue bars ("expected") give the number of genes that one would expect to obtain if the same number of genes (111 genes for type I) were chosen at random among all genes annotated in the relevant database. The bottom half of the figure gives similar information for type II genes. It is apparent that type I genes are enriched for RNA-related functions, whereas type II genes are not. Both types are enriched for DNA-related functions, but the type II genes are more enriched.

known genes. In 17 of these 24 "ancient" cases, there is clear mRNA or EST evidence that the coding region overlapped by the element is alternatively spliced in humans. These include alternatively spliced exons of genes *EIF2C1*, *BCL11A*, *EVI1*, *ZFR*, *CLK4*, *HNRPH1*, and *DDX5*, as well as *GRIA3*. In none of the other cases could we find evidence that any element that was intronic in humans was coding in another species, although in some cases there was EST evidence for a retained intron that presumably has a function other than protein coding. Moreover, indels of noncoding ultraconserved elements relative to their alignments with the chicken and other species are often not in multiples of three, giving further evidence that these sequences are noncoding (fig. S1, A and B).

The ultraconserved elements we found in introns seem to have been at one time rather fast-evolving as compared to the known coding exons in their genes. We tried to map selected introns containing ultraconserved intronic elements to more distant species using protein/translated DNA matches to their enclosing exons. Often only a "core" conserved region was recognizable in fish, and this had very different flanking DNA, suggesting that additional parts of the ultraconserved region were innovations after the common ancestor with fish, as observed in the analysis of uc.108 near *HOXD* (*29*). In cases where we could trace beyond vertebrates, we always found that the orthologous intron in the more distant species was either very small with apparently unrelated sequence, or was nonexistent. For example, tracing the intron that contains the first (most 5′) ultraconserved element in *POLA* (uc.460), we find that although it is an approximately 50-kb intron in humans, its ortholog in *Fugu rubripes* is only ~7500 bp (which is still large relative to most fugu introns), only about 335 bp in *C. intestinalis*, and does not exist (the flanking exons abut) in *D. melanogaster* and *C. elegans*. The human element is not recognizably similar to anything in the orthologous intron of *Ciona*. Yet, like the other *POLA* ultraconserved elements discussed above, this element is more than 99% identical between human and chicken. Similar results were found for the three longest *POLA* intronic elements. Another similar case was a cluster of seven ultraconserved elements (uc.273 to uc.279) with sizes ranging from 237 to 432 bp, all contained in an ~165-kb intron of *PBX3*, pre-B cell leukemia transcription factor 3, a member of the *TALE*/*PBX* homeobox family. This was one of the largest introns we found, and it contained one of the largest collections of ultraconserved elements in a single intron. The orthologous intron in *F. rubripes* is ~38 kb, in *C. intestinalis* it appears to be ~1 kb, in *Drosophila* it is ~200 bp (ortholog exd), and the flanking exons abut in *C. elegans*. Despite the inability to trace most of the vertebrate ultraconserved elements to distant species, the possibility that processes similar to those that produced ultraconserved elements in vertebrates also exist in other classes of species remains open. In one tantalizing example, it has been observed that the mating type gene *MATa2* in yeast shows 100% conservation over 357 bp in four yeast species (*30*). The mechanism of this conservation is not known.

We found only 12 paralogous sets, each consisting of two or three elements, among all 481 ultraconserved elements (table S5). Each paralogous set is consistent with the paralogy relationship between the enclosing or nearby "host" genes. All paralogs (except, currently, uc.344 overlapping *HOXC5*) have highly conserved matches in the chicken, providing more opportunities for evolutionary analysis of these duplication events that predate the divergence from birds. In each of the clusters, we found significant divergence between the paralogs, which must have occurred in the early part of their evolution (fig. S3), because each individual instance in a paralogous set has changed very little in the past 300 million years in birds and mammals. This, combined with the above analysis, suggests that the bulk of the ultraconserved elements represent chordate innovations that evolved fairly rapidly at first but then slowed down considerably, becoming effectively "frozen" in birds and mammals.

A more extensive analysis of paralogs, based on a recent global clustering of highly conserved noncoding human DNA (*31*), reveals several further highly conserved intronic and intergenic elements in functionally equivalent positions relative to paralogous genes. These were not classified as ultraconserved by our stringent criteria. Indeed, if we merge alignment blocks of 200 bases, each with at least 99% identical columns, we obtain 1974 "highly conserved" elements up to 1087 bp long in the human. Four of the five longest elements are the aforementioned *POLA*/*ARX* elements, along with a 906-bp element (encompassing uc.326/7) in an intron of *ELP4*, adjacent to *PAX6*. If instead we demand at least a 100-bp exact match between humans and rodents, we get more than 5000 highly conserved elements. Tens of thousands more are found at lower cutoffs; for example, there is a 57-bp exactly conserved sequence overlapping an alternatively spliced exon of the *WT1* gene which is invariant in mammals and in chickens and is largely conserved in fishes (fig. S1). The percentage of the conserved elements that overlap with a known coding region steadily rises from 14 to 34.7% as the length criteria defining these elements is reduced from 200 to 50 bp (table S6).

If experiments with less conserved elements in recent studies (*13*, *18*) are any indication, many of these shorter elements are also functional. Compared to the ultraconserved elements, a greater percentage of these shorter conserved elements are significantly different in birds but are highly conserved in mammals. This suggests that the process of the evolution of new elements, followed by near-"freezing" of their DNA sequences, is probably still ongoing in vertebrates. Lineage-specific specializations of these elements may reflect regulatory changes that are important to the ontogeny and physiology of the clade.

The patterns of conservation exhibited in the ultraconserved elements must result from the onset during chordate evolution of either a highly elevated negative selection rate in these regions (about a 20 times smaller chance of mutations becoming fixed in the population), a highly reduced mutation rate (about 20 times fewer mutations), or some combination of these effects. The possibility of strong negative selection is intriguing, because selection to maintain protein coding, protein-nucleic acid interactions, or RNA-RNA interactions does not result in near total conservation over long stretches of bases unless multiple functions are overlaid on the same DNA, such as in regions of coding exons that also bind splicing factors or in regions of rRNA that must form RNA structures as well as bind proteins. If the exonic ultraconserved elements form pre-mRNA structures that are under selection to preserve interaction with the spliceosome or editing machinery (*23*, *27*), then these interactions must be extremely constraining over hundreds of bases of DNA, much like those of the anciently derived rRNAs, making them potentially quite novel objects for molecular study. The same holds true if the conservation in the non-exonic elements is associated with selection for molecular interactions involved in the regulation of transcription, which could be in cis over long genomic distances, or in trans, perhaps also involving RNA (*29*, *32*, *33*).

On the other hand, if reduced mutation rates are the explanation, then the existence of regions of a few hundred bases with 20-fold reduced mutation rates would itself be quite novel. Although neutral mutation rates may vary depending on chromosomal location on a megabase scale (*34–36*), there is to our knowledge no evidence or precedent for the existence of short "hypomutable" or "hyperrepaired" neutral regions. Finally, the answer could also be a combination of negative selection and better repair in these regions, owing to some vital role that these elements play, such as self-regulating networks of RNA processing control in the case of exonic elements and self-regulatory networks of transcriptional control for non-exonic ele-

ments. In any case, the questions remain: What kind of elements associated with these processes would have arrived relatively early in chordate evolution and then become practically frozen in birds and mammals? And what mechanisms would underlie this, allowing them to resist virtually all further change?

Note added in proof: We recently became aware of related observations made by Boffelli *et al*. (*37*).

References and Notes

1. E. S. Lander *et al*., *Nature* **409**, 860 (2001).
2. J. C. Venter *et al*., *Science* **291**, 1304 (2001).
3. Human Genome Sequencing Consortium, in preparation.
4. R. H. Waterston *et al*., *Nature* **420**, 520 (2002).
5. K. M. Roskin, M. Diekhans, D. Haussler, in *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology* (ACM, New York, NY, 2003), pp. 257–266.
6. F. Chiaromonte *et al*., *Cold Spring Harbor Symp. Quant. Biol*. **68**, 245 (2003).
7. R. C. Hardison, *Trends Genet*. **16**, 369 (2000).
8. G. G. Loots *et al*., *Science* **288**, 136 (2000).
9. L. A. Pennacchio, E. M. Rubin, *Nature Rev. Genet*. **2**, 100 (2001).
10. K. A. Frazer *et al*., *Genome Res*. **11**, 1651 (2001).
11. U. DeSilva *et al*., *Genome Res*. **12**, 3 (2002).
12. E. T. Dermitzakis *et al*., *Nature* **420**, 578 (2002).
13. E. T. Dermitzakis *et al*., *Science* **302**, 1033 (2003).
14. Rat Genome Sequencing Consortium, *Nature* **428**, 493 (2004).
15. G. M. Cooper *et al*., *Genome Res*. **14**, 539 (2004).
16. J. W. Thomas *et al*., *Nature* **424**, 788 (2003).
17. E. H. Margulies, M. Blanchette, D. Haussler, E. D. Green, *Genome Res*. **13**, 2507 (2003).
18. K. A. Frazer *et al*., *Genome Res*. **14**, 367 (2004).
19. M. Ashburner *et al*., *Nature Genet*. **25**, 25 (2000).
20. N. J. Mulder *et al*., *Nucleic Acids Res*. **31**, 315 (2003).
21. M. A. Nobrega, I. Ovcharenko, V. Afzal, E. M. Rubin, *Science* **302**, 413 (2003).
22. S. Plaza, C. Dozier, M. C. Langlois, S. Saule, *Mol. Cell. Biol*. **15**, 892 (1995).
23. L. Rahman, V. Bliskovski, F. J. Kaye, M. Zajac-Kaye, *Genomics* **83**, 76 (2004).
24. I. L. Hofacker, *Nucleic Acids Res*. **31**, 3429 (2003).
25. H. Jumaa, P. J. Nielsen, *EMBO J*. **16**, 5077 (1997).
26. B. Sommer *et al*., *Science* **249**, 1580 (1990).
27. P. J. Aruscavage, B. L. Bass, *RNA* **6**, 257 (2000).
28. E. H. Sherr, *Curr. Opin. Pediatr*. **15**, 567 (2003).
29. C. Sabarinadh, S. Subramanian, R. Mishra, *Genome Biol*. **4** (2003).
30. M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E. S. Lander, *Nature* **423**, 241 (2003).
31. G. Bejerano, D. Haussler, M. Blanchette, *Bioinformatics* (Suppl.), in press.
32. J. S. Mattick, M. J. Gagen, *Mol. Biol. Evol*. **18**, 1611 (2001).
33. E. T. Dermitzakis *et al*., *Genome Res*. **14**, 852 (2004).
34. K. H. Wolfe, P. M. Sharp, W. H. Li, *Nature* **337**, 283 (1989).
35. R. C. Hardison *et al*., *Genome Res*. **13**, 13 (2003).
36. J. H. Chuang, H. Li, *PLoS Biol*. **2**, E29 (2004).
37. D. Boffelli, M. Nobrega, E. M. Rubin, *Nature Rev. Genet.*, in press.
38. F. Spitz, F. Gonzalez, D. Duboule, *Cell* **113**, 405 (2003).

39. We thank the Genome Sequencing Consortia for the human, mouse, rat, dog, chicken, and other genome sequences we used in this analysis; W. Miller, M. Diekhans, A. Hinrichs, K. Rosenbloom, D. Thomas, and the members of the University of California Santa Cruz (UCSC) browser team for providing the genome alignments and other tracks of genome annotation available on the UCSC genome browser; M. Blanchette, S. Salama, T. Lowe, M. Ares, K. Pollard, and B. Cohen for helpful discussions; A. Siepel for the neutral substitution rate analysis involving chicken and chimp; K. Roskin for the calculation of the percent identity in ancestral repeat sites for 1-Mb windows; and S. Walton for help in preparing the manuscript. G.B., W.J.K., and D.H. were supported by National Human Genome Research Institute grant 1P41HG02371 and National Cancer Institute contract 22XS013A, and D.H. was additionally supported by the Howard Hughes Medical Institute. S.S., M.P., I.M., and J.S.M. were supported by the Australian Research Council and the Queensland State Government.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1098119/DC1

SOM Text
Figs. S1 to S3
Tables S1 to S7
References and Notes

19 March 2004; accepted 27 April 2004
Published online 6 May 2004;
10.1126/science.1098119
Include this information when citing this paper.

A Family with Severe Insulin Resistance and Diabetes Due to a Mutation in *AKT2*

Stella George,[1]* Justin J. Rochford,[1]* Christian Wolfrum,[3] Sarah L. Gray,[1] Sven Schinner,[1] Jenny C. Wilson,[1] Maria A. Soos,[1] Peter R. Murgatroyd,[1] Rachel M. Williams,[2] Carlo L. Acerini,[2] David B. Dunger,[2] David Barford,[4] A. Margot Umpleby,[5] Nicholas J. Wareham,[6] Huw Alban Davies,[7] Alan J. Schafer,[8] Markus Stoffel,[3] Stephen O'Rahilly,[1]† Inês Barroso[8,9]

Inherited defects in signaling pathways downstream of the insulin receptor have long been suggested to contribute to human type 2 diabetes mellitus. Here we describe a mutation in the gene encoding the protein kinase AKT2/PKBβ in a family that shows autosomal dominant inheritance of severe insulin resistance and diabetes mellitus. Expression of the mutant kinase in cultured cells disrupted insulin signaling to metabolic end points and inhibited the function of coexpressed, wild-type AKT. These findings demonstrate the central importance of AKT signaling to insulin sensitivity in humans.

Most forms of diabetes are likely to be polygenic in origin, although a number of monogenic forms are being recognized (*1*, *2*). Although rare, these monogenic examples offer insight into the function of the affected gene in humans as well as offering important clues to understanding more common forms.

We have been screening genomic DNA from 104 unrelated subjects with severe insulin resistance for mutations in genes that are implicated in insulin signaling. We identified a missense mutation in the serine/threonine kinase gene *AKT2* in one Caucasian proband. AKT2 (also known as PKBβ) is highly expressed in insulin-sensitive tissues and is activated in response to growth factors and related stimuli (*3*, *4*), a process that requires its phosphorylation by the phosphoinositide-3 phosphate-dependent kinase activities designated PDK1 and PDK2 (*3*). The proband (Fig. 1D, iii/1) is a nonobese 34-year-old female who developed diabetes mellitus at 30 years of age. The proband, her nonobese mother, her maternal grandmother, and a maternal uncle were all heterozygous for a G-to-A substitution predicted to result in an R-to-H substitution at amino acid 274 (Fig. 1, A and B) (*5*). All were markedly hyperinsulinemic (table S1), and the mother and maternal grandmother developed diabetes mellitus in their late thirties. Three other first-degree relatives available for study were all clinically normal, with normal fasting glucose and insulin, and were homozygous for the wild-type AKT2 sequence (Fig. 1D and table S1). This mutation was not found in the genomic DNA of 1500 Caucasian control subjects from the United Kingdom.

R^{274} forms part of an RD sequence motif within the catalytic loop of the AKT2 kinase

[1]Department of Clinical Biochemistry, [2]Department of Paediatrics, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [3]Laboratory of Metabolic Diseases, Rockefeller University, New York, NY 10021, USA. [4]Section of Structural Biology, Institute of Cancer Research, Chester Beatty Laboratories, London SW3 6JB, UK. [5]Department of Diabetes, Endocrinology and Internal Medicine, Guy's, King's and St. Thomas' School of Medicine, London, UK. [6]Medical Research Council Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. [7]Darent Valley Hospital, Darenth Wood Road, Dartford, Kent DA2 8DA, UK. [8]Incyte, 3160 Porter Drive, Palo Alto, CA 94304, USA. [9]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: sorahill@hgmp.mrc.ac.uk