4/22/2004

Title:

Ultra-conserved elements in the human genome

Authors and affiliations:

Gill Bejerano*, Michael Pheasant**, Igor Makunin**, Stuart Stephen**, W. James Kent*, John S. Mattick** and David Haussler***

*Department of Biomolecular Engineering and ***Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

**ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia

Corresponding authors: Gill Bejerano (jill@soe.ucsc.edu) and David Haussler (haussler@soe.ucsc.edu)

Supporting on-line material:

Separate figures, Like Figure 1 but for each individual chromosome are available in postscript and PDF format, at http://www.cse.ucsc.edu/~jill/ultra.html.

Table S1. A table listing all 481 ultra conserved elements and their properties can be found at http://www.cse.ucsc.edu/~jill/ultra.html.

The elements were extracted from an alignment of NCBI Build 34 of the human genome (July 2003, UCSC hg16), mouse NCBI Build 30 (February 2003, UCSC mm3), and rat Baylor HGSC v3.1 (June 2003, UCSC rn3). This table does not include an additional, probably ultra conserved element (uc.10) overlapping an alternatively spliced exon of FUSIP1, which is not yet placed in the current assembly of human chromosome 1. Nor does the list contain the ultra conserved elements found in ribosomal RNA sequences, as these are not currently present as part of the draft genome sequences. The small subunit 18S rRNA includes 3 ultra conserved regions of sizes 399, 224, 212bp and the large subunit 28S rRNA contains 3 additional regions of sizes 277, 335, 227bp (the later two are one base apart). Also excluded is a genomic rDNA fragment (uc.32) of length 328bp on human chromosome 1, with perfect, but non-syntenic matches in the draft genomes of mouse and rat.

The table lists for each element (1) name (2) length of absolutely conserved segment in bp (3) type of element – exonic, non-exonic or possibly-exonic, as defined in main text. (4) position in the assembled genome (July 2003 version), (5-6) Distance to nearest gene upstream of element (on leading strand), and name of gene, (7) Name of gene element resides in, or "\N" for intergenic elements, (8-9) name and distance of nearest downstream gene, (10-11) Number of bases overlapping GenBank human mRNA/EST records, (12-13) Base overlap with any species mRNA/EST records, (14-16) number of bases that overlap a UTR region, coding region, or

intron of a known gene (as defined in the known genes track of the UCSC browser (*39*). In case of multiple gene isoforms we combine overlaps from all isoforms. (17-19) 10Kb upstream, 10Kb downstream, or 10Kb away from any known gene, (20) RNAfold (*24*) prediction based fraction of 10,000 shuffled versions of its sequence with minimal energy lower than that of the element itself, (21-22) Number of bases that align in the best chicken draft match, plus how many of these are identical in chicken, (23-24) likewise for fugu,

In addition, for interactive exploration, a direct link is provided to the UCSC genome browser for each of the 481 elements, which shows the extent of the element, and allows access to all the information available about this region of the genome from the UCSC genome browser, including, mapped mRNAs, ESTs, known genes and gene predictions, as well as detailed DNA alignments of it and the surrounding region to other species.

Table S2a.

ELEMENT	dbSNP ACCESSION	LOCATION OF ELEMENT
uc.53	rs1861100	intergenic
uc.140	rs2056116	intergenic
uc.252	rs1538101	intergenic
uc.295	rs7092999	intergenic
uc.353	rs9572903	intergenic
uc.374	rs7143938	in intron of MIPOL1

Legend of Supplementary Table 2a. The six validated SNPs found searching the 481 ultra-conserved elements, ignoring the first and last 20bp of each element. The columns list the ultra-conserved element that contains the SNP, the dbSNP Accession for information about the SNP (http://www.ncbi.nlm.nih.gov/SNP/), and the location of the ultra-conserved element that contains the SNP. See Table 1 for further information about each ultra-conserved element.

Table S2b.

ELEMENT	dbSNP ACCESSION	LOCATION OF ELEMENT
uc.478	rs1132303	"flop" exon of GRIA3
uc.478	rs1052539	"flop" exon of GRIA3
uc.478	rs1052540	"flop" exon of GRIA3
uc.478	rs1052541	"flop" exon of GRIA3
uc.478	rs1052542	"flop" exon of GRIA3
uc.478	rs1052543	"flop" exon of GRIA3
uc.478	rs1052544	"flop" exon of GRIA3
uc.478	rs1052545	"flop" exon of GRIA3
uc.478	rs1052546	"flop" exon of GRIA3

Legend of Supplementary Table 2b. A cluster of nine unvalidated SNPs found in one ultra-conserved element. These appear to be errors in dbSNP caused by confusing the "flip" exon of GRIA3 as a polymorphic variant of the "flop" exon.

Table S3.

ELEMENT	LENGTH	GENE NAME	GO/InterPro ATTRIBUTE
uc.13	237	EIF2C1	
uc.28	355	SFRS11	RNA binding, RRM
uc.33	312	PTBP2	RRM
uc.45	203	HNRPU	RNA binding
uc.46	217	HNRPU	RNA binding
uc.48	298	PUM2	RNA binding
uc.49	207	BC060860	
uc.50	222	SFRS7	RNA binding, RRM
uc.61	326	BCL11A	
uc.77	296	ZFHX1B	
uc.97	442	HAT1	
uc.102	338	PTD004	
uc.129	212	MBNL1	RNA binding
uc.135	201	AK096400	RNA binding
uc.138	419	SFRS10	RNA binding, RRM
uc.143	218	AB014560	RNA binding, RRM
uc.144	205	HNRPDL	RNA binding, RRM
uc.151	214	ZFR	RNA binding
uc.174	260	MATR3	RNA binding, RRM
uc.183	236	FBXW1B	
uc.184	230	CPEB4	RRM
uc.185	411	CLK4	
uc.186	305	HNRPH1	RNA binding, RRM
uc.189	573	SFRS3	RNA binding, RRM
uc.193	319	SYNCRIP	RNA binding, RRM
uc.194	201	EPHA7	
uc.203	203	AB067798	
uc.208	218	TRA2A	RNA binding, RRM
uc.209	250	TRA2A	RNA binding, RRM
uc.233	266	CENTG3	
uc.263	207	HNRPK	RNA binding
uc.264	267	HNRPK	RNA binding
uc.280	220	PBX3	
uc.282	207	GRIN1	
uc.285	232	CARP-1	
uc.292	217	MLR2	
uc.313	231	TIAL1	RNA binding, RRM
uc.324	225	C11orf8	
uc.330	207	RBM14	RNA binding, RRM
uc.331	218	DLG2	
uc.333	270	FLJ25530	
uc.338	223	PCBP2	RNA binding
uc.339	252	ATP5G2	
uc.356	251	MBNL2	

ELEMENT	LENGTH	GENE NAME	GO/InterPro ATTRIBUTE
uc.375	300	MIPOL1	
uc.376	290	PRPF39	
uc.377	217	PRPF39	
uc.378	251	NRXN3	
uc.393	275	CLK3	
uc.395	249	RBBP6	
uc.406	211	NFAT5	
uc.409	244	L32833	
uc.413	272	BC060758	
uc.414	246	THRA	
uc.419	289	SFRS1	RNA binding, RRM
uc.436	210	TCF4	
uc.443	239	HNRPM	RNA binding, RRM
uc.454	208	SLC23A1	
uc.455	245	RNPC2	RNA binding, RRM
uc.456	320	SFRS6	RNA binding, RRM
uc.471	239	DDX3X	RNA binding
uc.473	222	NLGN3	
uc.474	210	ZNF261	
uc.475	397	OGT	
uc.477	209	RAB9B	
uc.478	252	GRIA3	
uc.479	302	GRIA3	

Legend of Supplementary Table 3. A curated list of ultra conserved elements implicated in alternative splicing. For each of the 67 elements we show its name and length, which gene it resides in, and whether that gene is annotated with the most enriched GO annotation (RNA binding, $p < 8.1 \times 10^{-18}$ in this set), and/or InterPro annotation (RNA recognition motif, $p < 9.1 \times 10^{-19}$ in this set).

Table S4.

RANK	NAME (STRAND)	LENGTH	MINIMAL ENERGY	FRACTION OF SHUFFLES WITH LOWER ENERGY	BRIEF DESCRIPTION
1	uc.193+	319	-82.00	0/10000	in 3' UTR of SYNCRIP, RNA binding
2	uc.281-	238	-69.43	0/10000	in intron of DDX31, RNA helicase, evidence it is separately transcribed
3	uc.189-	573	-196.41	1/10000	in alt-spliced 3' UTR in SFRS3, RNA splicing factor
4	uc.275-	255	-45.17	1/10000	in intron of transcription factor PBX3
5	uc.338-	223	-107.80	1/10000	alt-spliced exon of PCBP2
6	uc.397-	311	-104.10	1/10000	in intron of transcription factor OAZ
7	uc.334-	222	-72.80	6/10000	in intron of HNT, a cell adhesion molecule family member
8	uc.214+	243	-57.84	7/10000	20Kb upsteram of transcription factor NEUROD6

9	uc.93-	263	-65.60	7/10000	about 200kb upstream from FIGN, unknown function
10	uc.445-	310	-63.64	8/10000	in intron of mRNA AK098372, unknown function
11	uc.433-	206	-47.80	9/10000	in gene desert 1mb ustream from RNA binding gene BRUNOL4
12	uc.475-	397	-114.20	9/10000	in an alternative 5'UTR of OGT (nuclear localized)
13	uc.355+	228	-53.70	15/10000	overlaps unspliced transcript (EST AI359363, three others too)
14	uc.111+	296	-106.04	18/10000	alt spliced exon of KIAA1757
15	uc.116+	206	-54.80	18/10000	in gene desert 500Kb from transcription factor FOXP1.
16	uc.357+	242	-67.70	27/10000	near transcription factor/homeobox SOX1
17	uc.468+	489	-146.98	27/10000	next to uc.469 between POLA and homeobox ARX
18	uc.198-	307	-97.70	28/10000	in intron of gene with unknown function near transcription factor POU3F2
19	uc.143+	218	-61.40	32/10000	contains alt-spliced coding exon of G3BP2
20	uc.354-	235	-65.00	33/10000	may be part of gene represented by mouse mRNA AK051163 near transcription factor POU4F1
21	uc.479-	302	-75.92	34/10000	flip alt-exon of GRIA3
22	uc.29-	219	-57.50	45/10000	in intron of uncharacterized gene near transcription factor LMO4
23	uc.157-	207	-59.02	48/10000	in cluster of 3 elements upstream of the ortholog of the fly transcription factor orthopedia
24	uc.327+	268	-69.12	48/10000	in intron of transcription factor ELP4
25	uc.224+	295	-72.64	52/10000	in intron of FOXP2
26	uc.166+	310	-92.50	53/10000	overlaps transcript of uncharacterized gene near transcription factor MEF2C
27	uc.406+	211	-49.65	61/10000	overlaps alt-spliced exon of NFAT5
28	uc.335-	214	-68.60	62/10000	in intron of neuronal specific transcription factor DAT
29	uc.83+	296	-66.00	62/10000	in intron of uncharacterized gene defined by mRNA BC032407
30	uc.120+	270	-67.16	63/10000	in intron of transcription factor ZNF288
31	uc.268-	251	-65.30	63/10000	in intron of MNAB, has 1 validated SNP and 3 unvalidated
32	uc.461-	397	-99.70	82/10000	in intron of POLA (near transcription factor ARX)
33	uc.431-	230	-50.10	83/10000	in intron of BRUNOL4
34	uc.283-	277	-70.80	84/10000	in intron of transcription factor DRG11
35	uc.245-	339	-95.34	91/10000	in intron of transcription factor ZFPM2
36	uc.88-	312	-110.60	95/10000	upstream of uncharacterized gene

Legend of Supplementary Table 4. The table shows the top ranking ultra conserved elements, with respect to a minimal energy computed against 10,000 random shuffles of each element. The strand with lower ranking (smallest fraction of shuffles with lower energy) is shown. If we interpret these rankings as p-values, then the top twelve elements in this table show significant

evidence of secondary structure at a false discovery rate of 0.05 (*40*). However, there may be serious limitations with the random permutation null model, so this must be viewed with caution. The results are only suggestive, and do not confirm or reject the presence of RNA structure in these sequences.

Table S5.

PARALOGOUS ELEMENTS	PARALOGOUS HOST GENES	LOCATION
uc.175, uc.235, uc.318	EBF, EBF2, EBF3	intron
uc.150, uc.403, uc.404	IRX1, IRX3, IRX6	distal
uc.40, uc.275	PBX1, PBX3	intron
uc.129, uc.356	MBNL1, MBNL2	intron
uc.123, uc.355	SOX14, ABCC4	distal
uc.138, uc.208+209	SFRS10, TRA2A	alt-exon
uc.185, uc.393	CLK1, CLK2	alt-exon
uc.213, uc.344, uc.416	HOXA5, HOXC5, HOXB5	5' exon
uc.342, uc.417	HOXC6, HOXB6	5' UTR
uc.257, uc.298	PAX5, PAX2	distal
uc.397, uc.425	OAZ, EHZF	intron
uc.478, uc.479	GRIA3	flop and flip exons

Legend of Supplementary Table 5. The 12 clusters of paralogous ultra-consevred elements founds by Blastz comparison among the elements. All clusters were found to be associated with paralogous "host genes" as indicated in columns 2 and 3, which either contained the elements in the corresponding introns ("intron"), had corresponding exons overlapping with the elements ("alt-exon", "5' exon", or "5' UTR"), or were consistently positioned upstream or downstream of the elements ("distal").

Table S6.

LENGTH	# OF CONSERVED ELEMENTS	# OF CONSERVED CODING ELEMENTS	% CODING ELEMENTS
50 to 799	18,391	5,596	30.4%
100 to 779	5,412	1,482	27.4%
200 to 779	482	108	22.4%
300 to 779	97	18	18.6%
400 to 779	18	2	11.1%

Legend of Supplementary Table 6.

Number of perfectly conserved elements of 50bp or more. For lengths between 50bp and 779bp (the length of the longest contiguous ultra-conserved element), "# of conserved elements" gives the number of elements in the human genome of size in the indicated length range that are

absolutely conserved (100% identity with no insertions or deletions) between orthologous regions in the mouse and rat genomes. The remaining columns give the number of such elements that overlap a known coding region and the percent these constitute of the total number of conserved elements. The fraction of the elements overlapping coding sequence tends to drop as the length of the element increases.

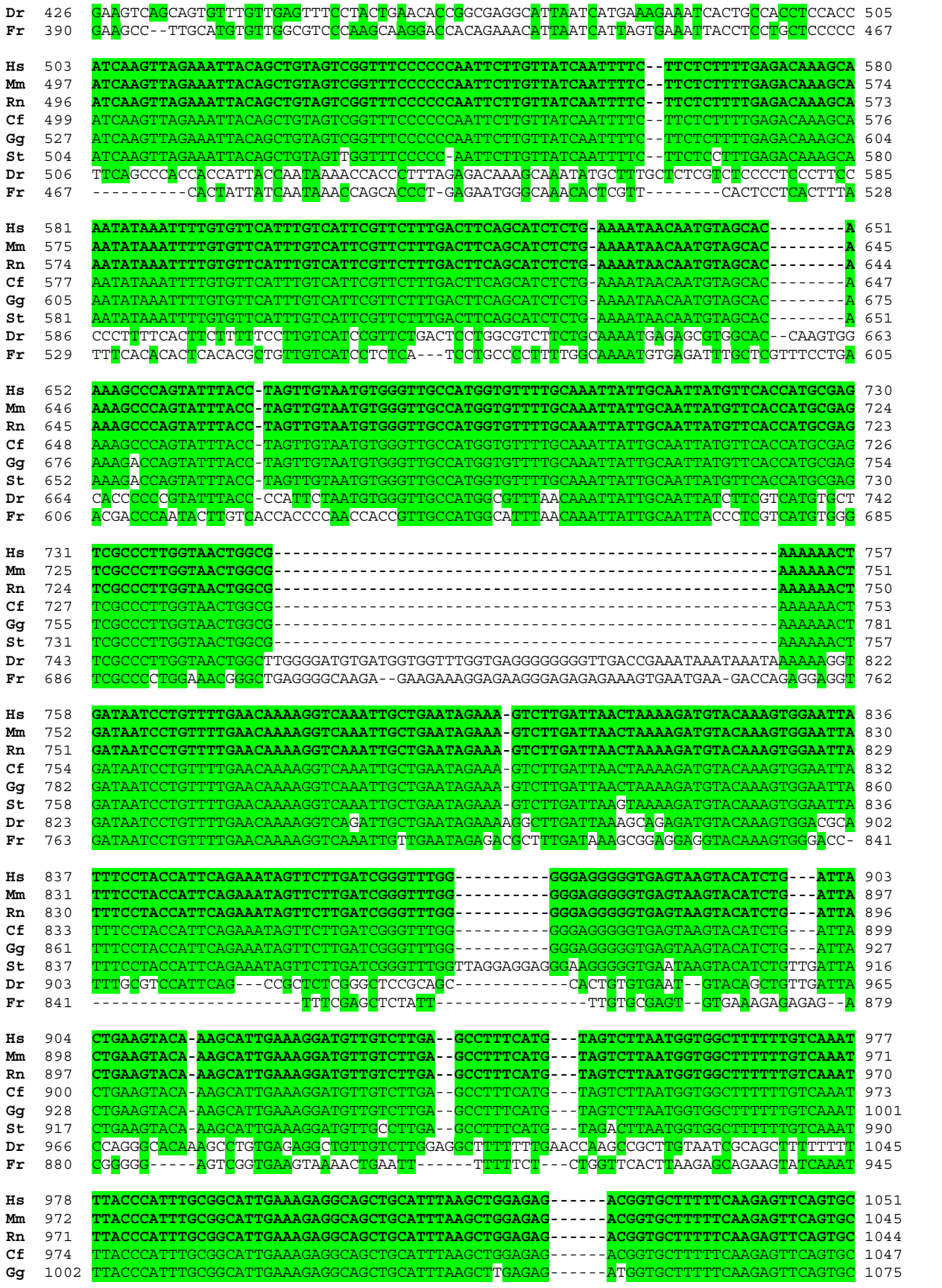
Table S7.

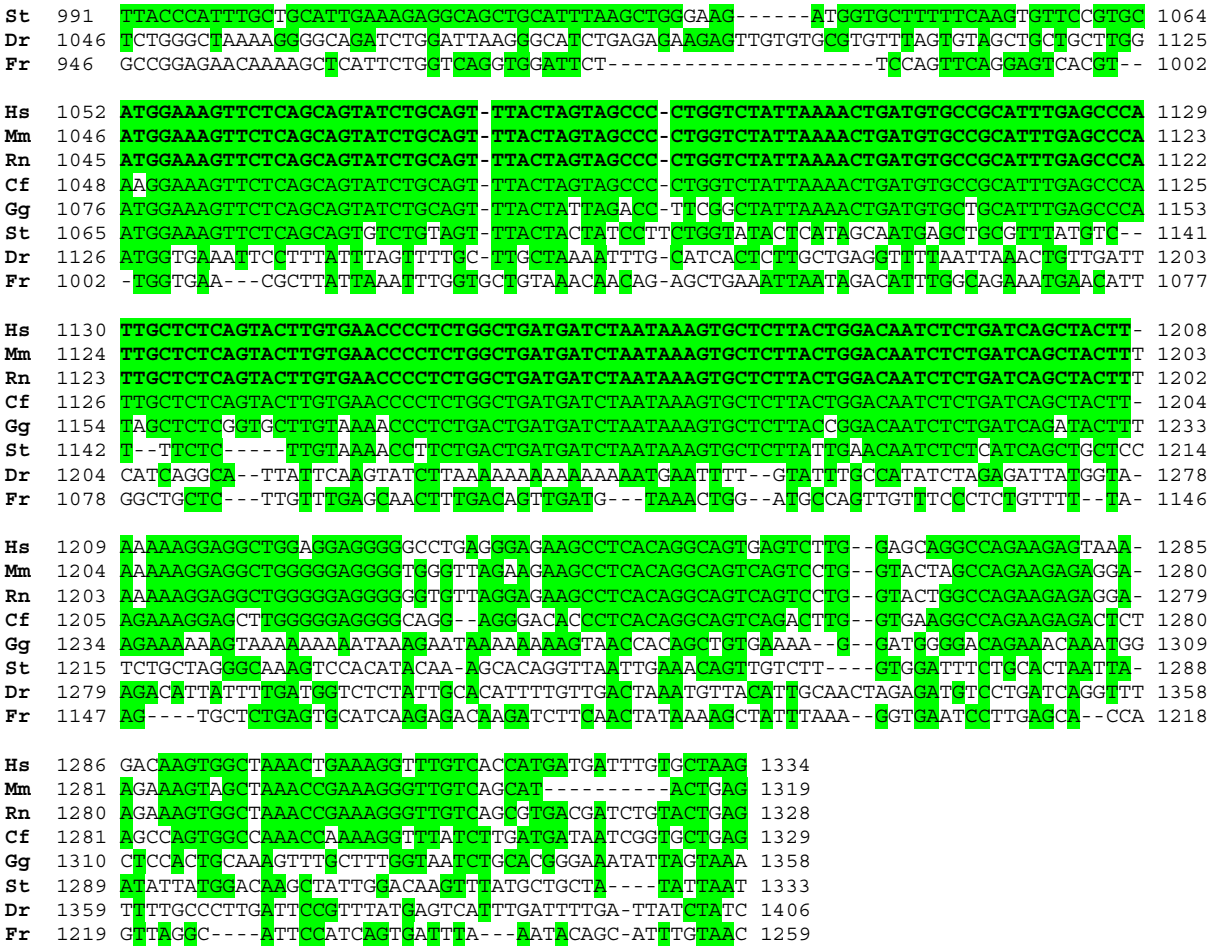
ELEMENT	GENE NAME	*C. elegans*	*Ciona intestinalis*	*Drosophila melanogaster*
uc.13	EIF2C1	3.80E-14	-	4.30E-23
uc.61	BCL11A	1.80E-33	4.50E-17	1.00E-40
uc.97	HAT1	-	9.10E-06	-
uc.102	PTD004	Blastz	-	-
uc.135	EVI1	-	-	2.40E-10
uc.151	ZFR	-	translated Blat	-
uc.153	KPNB2	5.20E-05	3.30E-06	1.00E-09
uc.169	NR2F1	2.00E-27	1.40E-22	1.60E-31
uc.185	CLK4	-	5.10E-05	-
uc.186	HNRPH1	7.90E-07	-	3.60E-08
uc.194	EPHA7	-	9.80E-12	1.50E-08
uc.280	PBX3	4.20E-06	7.40E-11	2.00E-12
uc.292	MLR2	8.90E-05	-	8.00E-06
uc.299	PAX2	1.40E-18	4.40E-20	1.70E-21
uc.324	C11orf8	8.60E-09	1.10E-16	2.10E-07
uc.331	DLG2	-	-	3.60E-05
uc.341	HOXC10	3.30E-18	1.20E-25	6.90E-19
uc.356	MBNL2	-	-	6.80E-06
uc.419	SFRS1	4.40E-09	5.00E-09	1.40E-06
uc.420	DDX5	8.50E-15	6.40E-15	5.00E-14
uc.457	HIRA	-	-	1.90E-08
uc.459	CNK2	-	1.00E-07	-
uc.478	GRIA3	1.30E-09	-	1.40E-06
uc.479	GRIA3	1.30E-07	-	1.70E-07

Legend of Supplementary Table 7. A curated list of 24 ultra conserved elements that could be traced back to worm, sea squirt or fly. In all cases the match is between coding exons. The majority of matches were found using NCBI tblastx (with matrix Blosum45). For these we

report tblastx E-value scores. Two additional matches were obtained using Blastz and translated Blat. As these tools have no E-value associated with their matches, we give the tool's name instead.

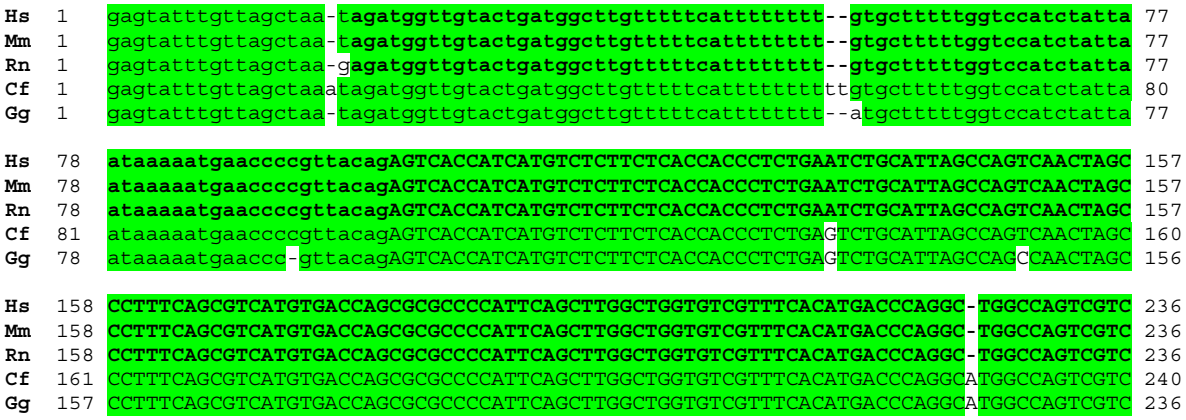
Figure S1a.





Legend of Supplementary Figure 1a. Multiple alignment of the 779 bp ultra-conserved element (uc.462, shown in bold), which occurs in an intron of DNA polymerase alpha (POLA), along with flanking sequence. Orthologous sequences are taken from human (Hs), mouse (Ms), rat (Rn), dog (Cf), chicken (Gg), frog (St), zebrafish (Dr) and fugu (Fr). The consensus base is highlighted in columns with over 50% identity.

Figure S1b.



```
Hs  237 AGGTTGCACCGCCCTTTGGTTCCCGAGCATGCTGTTTTCTCTCAGCCTTCTCTCCAACCTTAACCAAATCGGCAGCAGCC 316
Mm  237 AGGTTGCACCGCCCTTTGGTTCCCGAGCATGCTGTTTTCTCTCAGCCTTCTCTCCAACCTTAACCAAATCGGCAGCAGCC 316
Rn  237 AGGTTGCACCGCCCTTTGGTTCCCGAGCATGCTGTTTTCTCTCAGCCTTCTCTCCAACCTTAACCAAATCGGCAGCAGCC 316
Cf  241 AGGTTGCACCGCCCTTTGGTTCCCGAGCATGCTGTTTTCTCTCAGCCTTCTCTCCAACCTTAACCAAATCGGCAGCAGCC 320
Gg  237 AGGTTGCACCGCCCTTTGGTTCCCGAGCATGCTGTTTTCTCTCAGCCTTCTCTCCAACCTTAACCAAATCGGCAGCAGCC 316

Hs  317 ACCTCGACCGCCCACACATTCCTGGCCAATCAGCTCAGCTGTTTATTTACCAAATGTCTTCACAACAACTACAGCAGCAG 396
Mm  317 ACCTCGACCGCCCACACATTCCTGGCCAATCAGCTCAGCTGTTTATTTACCAAATGTCTTCACAACAACTACAGCAGCAG 396
Rn  317 ACCTCGACCGCCCACACATTCCTGGCCAATCAGCTCAGCTGTTTATTTACCAAATGTCTTCACAACAACTACAGCAGCAG 396
Cf  321 ACCTCGACCGCCCACACATTCCTGGCCAATCAGCTCAGCTGTTTATTTACCAAATGTCTTCACAACAACTACAGCAGCAG 400
Gg  317 ACCTCGACCGCCCACACATTCCCGGCCAATCAGCTCAGCTGTTTATTTACCAAATGTCTTCACAACAACTACAGCAGCAG 396

Hs  397 CCTTCGGCTAACAAAAAAGCAGGAAAAATCCACAACACCCCCTTCGCCAACCAACTAAATCCAACGCAACATCTGGCAAA 476
Mm  397 CCTTCGGCTAACAAAAAAGCAGGAAAAATCCACAACACCCCCTTCGCCAACCAACTAAATCCAACGCAACATCTGGCAAA 476
Rn  397 CCTTCGGCTAACAAAAAAGCAGGAAAAATCCACAACACCCCCTTCGCCAACCAACTAAATCCAACGCAACATCTGGCAAA 476
Cf  401 CCTTCGGCTAACAAAAAAGCAGGAAAAATCCACAACACCCCCTTCGCCAACCAACTAAATCCAACGCAACATCTGGCAAA 480
Gg  397 CCTTCGGCTAACAAAAAAGCAGGAAAAATCCACAACACCCCCTTCGCCAACCAACTAAATACAACGCAACATCTGGCAAA 476

Hs  477 ACCTTTTCAGCAAATTCTTCCTGGCCGTCAGTCCGGCAGCCTCACCTCACCATTTCTAGCTTGTTGAAACCCAAAACTAg 556
Mm  477 ACCTTTTCAGCAAATTCTTCCTGGCCGTCAGTCCGGCAGCCTCACCTCACCATTTCTAGCTTGTTGAAACCCAAAACTAg 556
Rn  477 ACCTTTTCAGCAAATTCTTCCTGGCCGTCAGTCCGGCAGCCTCACCTCACCATTTCTAGCTTGTTGAAACCCAAAACTAg 556
Cf  481 ACCTTTTCAGCAAATTCTTCCTGGCCGTCAGTCCGGCAGCCTCACCTCACCATTTCTAGCTTGTTGAAACCCAAAACTAg 560
Gg  477 ACCTTTTCAGCAAATTCTTCTTGGCAGTCAGTCCGGCAGCCTCACCTCACCATTTCTAGCTTGTTGAAACCAAAAACTAg 556

Hs  557 taagtttttcctgcttatacagtttactgctggttaaaa-taaggagtaagcggctta 613
Mm  557 taagtttttcctgcttatacagtttactgctggttaaaaataaggagtaagcggctta 614
Rn  557 taagtttttcctgcttatacagtttactgctggttaaaaataaggagtaagcggctta 614
Cf  561 taagtttttcctgcttatacagtttactgctggttaaaa-taaggagtaagcggctta 617
Gg  557 taagtttttcctgcttatacagtttactgctggttaaaa-taaggagtaagcggctta 613
```

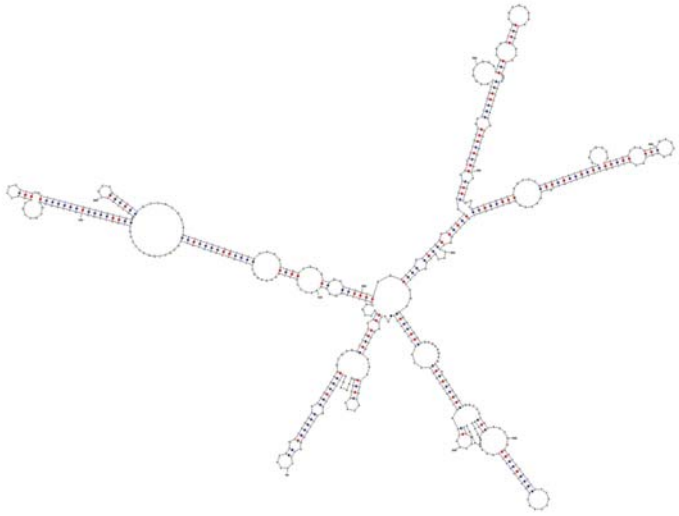
Legend of Supplementary Figure 1b. Alignment of then ultra-conserved sequence from the SFRS3 gene (uc.189). Sequence identical in human, mouse and rat (bold) includes an alternatively spliced 3' UTR exon (upper case). Neighboring introns (lower case) show transcriptional evidence of retention. The consensus base is highlighted in columns with over 50% identity. Note that all indels are one or two nucleotides long. Species acronyms as above.

Figure S1c.

```
Hs   1  CGGTCCGACCACCTGAAGACCCACACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttcttcacatttatttttcattatttttttaaacta 97
Pt   1  CGGTCCGACCACCTGAAGACCCACACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttcttcacatttatttttcattatttttttaaacta 97
Mm   1  CGGTCCGACCATCTGAAGACCCACACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttcttcacgtttatttttcattatttttctaagcta 97
Rn   1  CGGTCCGACCACCTGAAGACCCACACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttcttcacatttattcttcattatttttctaagcta 96
Cf   1  CGGTCTGACCACCTGAAGACCCACACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttcttcacatttatttttcattatttttttaaaaaa 97
Bt   1  CGTTCCGACCACCTGAAGACCCACACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttcttcacatttatttttctttatttttttaaacta 97
Gg   1  AGATCTGATCATCTGAAGACTCATACCAGGACTCATACAGGTAAAACAAgtgcgtaaacttttattcacatttatatttcattatttttttaagttta 97
Xl1  1  AGGTCCGACCACCTGAAGACTCACACCAGGACTCATACAGGTAAAACAAgtgcgtatacttttcttcatatttatttttactatatttttgatatctc 97
Xl2  1  AGGTCCGACCATCTGAAGACTCACACCAGGACTCATACAGGTAAAACAAgtgcgtatacttttcttcatattt-ttattactacattgttgaaatct 96
                                                                                *
```

Legend of Supplementary Figure 1c. Alignment of genomic DNA from the WT1 locus. The intron is shown in small case, and the site of alternative splicing site is marked by an asterisk. Species acronyms are as above, with two additional frog (Xenopus laevis) variants.

Figure S2.



Legend of Supplementary Figure 2. A putative RNA secondary structure of element uc.189, whose minus strand sequence has a minimal folding energy lower all but 1/10000 random shuffles of itself. This element overlaps an alternatively spliced 3' UTR exon of SFRS3.

Figure S3.

```
102112576 ccaaggcttcctgctgtcagctggggaatagataaagataaatgatattatgttaaattc
>>>>>>>>> || ||||    ||||||||     | | || ||||||||||||||| |||||| |||||||
037205225 cccaggcccattgctgtcactgatgaactatataaagataaatgacattatggtaaattc

102112516 cacttaatgacaaatttttaattttctgaacatggtcattttctggctagtgaatcaagt
>>>>>>>>> |||||||||||| |||||||||||||| |||||  | ||||||| |||||||||||| |||
037205285 cacttaatgacacatttttaattttcagaacacagacattttcaggctagtgaatgaag-

102112456 ggagggagctaattacatgaagatctgaacaaaaataactcctaattttcaaggataatg
>>>>>>>>>        ||||||||| | |||||| |||| |||||||||||| ||||  || ||||||
037205344 -------gctaattacctcaagatcagaac-aaaataactcctcatttcgaaagataata

102112396 gaagagaaatgttggagattaatggcactatttatcttt-----tttaaatttctatctt
>>>>>>>>> ||| |||||||||| |||||||||   |||||||||||        ||| || ||||||
037205396 gaaaagaaatgttgcagattaatgatgctatttatctttaaaggaaaaaaattttatctt

102112341 tctctgtgatagccgtgctccccaaggaaaatattcataaaatgaaattgaagtcgtaac
>>>>>>>>> | ||   ||||||   |||||| |||||   |  | ||||||||||  |||||||| |||
037205456 tatcca-gatagcatagctccctaaggagcgtggttataaaatgaatctgaagtcgcaac

102112281 ttaataggttattaaaatttttgagtgcatatcactttccttccgcagcactgtaaattt
>>>>>>>>> |||||||||||||||| || ||||||||        | |||  |     |  |||||||
037205515 ttaataggttattaaaa-ttgtgagtgcaggctctcctgctttagtctttccataaattt

102112221 aaattgaag 102112213
>>>>>>>>> ||||||||| <<<<<<<<<
037205574 aaattgaag 037205582
```

Legend of Supplementary Figure 3. Alignment of the similar portions of paralogous elements uc.257 and uc.298, located distal to host genes PAX5 and PAX2, respectively. Pairs of identical bases are joined by a vertical line. Recall that while these two elements differ markedly, each of them is perfectly conserved between human, mouse and rat.

Text Section S1.

Alignments of human, mouse, rat, fugu and chicken DNA were taken from the UCSC genome browser site, http://genome.ucsc.edu, built by Webb Miller, and the UCSC genome browser staff using Blastz (*41*). Chaining methods were used to remove non-orthologous matches, as described in (*42*). Regions overlapping segmental duplications were removed as well.

Calculation of p-value for finding any instance of 200 bases absolutely conserved between human, mouse and rat in the human genome: This calculation is done using a Poisson approximation. Each column in the orthologous multiple alignment is considered to be an independent observation of a Bernoulli random variable that is 1 ("heads") if the bases are completely conserved between the three species (a "3-way identity") and 0 ("tails") otherwise. Based on analysis of neutrally evolving (ancestral repeat) sites in each 1 Mb window in the human genome (*1, 35*), we estimated the mean of this Bernoulli variable (the probability of heads) to be at most 0.7. (The largest percent identity among ancestral repeat sites we obtained for any 1 Mb window with enough ancestral repeat sites to get a good estimate, i.e. at least 1000 sites, was actually 0.68.) The distribution of the number of runs of at least 200 heads in a series of 2.9 billion tosses of a biased coin with probability $p = 0.7$ of heads can be approximated quite well using a Poisson distribution with mean $(1-p)\cdot p^{200}$ (*43*), and the probability of one or more such runs is very close to the mean of the Poisson distribution in this case, which is at most 10^{-22}. This probability is small even if the neutral probability of 3-way identity is as high as 0.9.

Text Section S2. Calculation of the estimate neutral rate of substitution was done with genomic data from a 1.4 Mb region containing the human CFTR gene and orthologous regions in 12 other vertebrates (*16*). The phylogenetic position of chicken, dog, mouse, rat, chimp and human in this set was not in doubt.

For chicken estimates, third positions in aligned codons in the genes from this region were used to estimate a rate of substitution on each branch using the HKY model (version 3.13 of the PAML package), and a scaling factor of 1.2 was used to account for the effects of selection in some of these sites (estimated from similar experiments on mammals, where neutral sites from ancestral repeats could be used to calibrate). This gave an estimated neutral substitution rate of approximately 1 substitution per site in total on the branches between the chicken and the primate-rodent common ancestor. Very similar results were obtained by the REV model. At such large distances, the variance in these estimates can be considerable, so further work would have to be done to refine this rough estimate, but it seems unlikely to be much below about 0.85 substitutions per site. (Note that the substitution rate between human and mouse is about 0.5 substitutions per site, although the mouse is known to have a faster clock (*16*)). On the other hand, a 95.7% observed percent identity translates into an estimated substitution rate of 0.044 substitutions per site in the ultra-conserved elements between human and chicken, and the perfect identity between human and rodents suggests that most of these were on the branches separating chicken and the primate-rodent ancestor. Hence the substitution rate on these branches is likely to be reduced at least 20-fold in these sites.

For dog, a neutral rate of substitution between the dog and the primate-rodent common ancestor was estimated at 0.2 substitutions per site. The rate of observed changes with respect to the dog genome of 0.008 changes per site in the ultra-conserved regions translates into roughly the same rate of substitution, which is 25 times less than expected under the neutral estimate.

For the estimates of the expected number of differences between chimp and human, when the human base is identical to that of both rodents, we used the REV model, in a similar way to those described above, except that here we fully modeled the conditional probabilities, given that the human, mouse and rat bases where observed to be identical. This gave 716 expected changes in 106,767 ultra-conserved sites, compared to the 38 observed changes in high quality reads. This leads to an estimate of 19-fold slower substitution rate in the ultra-conserved regions. This estimate is fairly crude, because it does not take into account the local fluctuations in neutral rate, which could have a bigger effect on this calculation than they do on the calculation for dog and chicken, due to the smaller evolutionary distance.

Text Section S3.

The ultra-conserved elements can be classified as (1) lying inside known genes (defined by the "known genes" track on the UCSC Genome browser at http://genome.ucsc.edu, including the UTR and introns (277/481 = 57.6%), (2) lying within 10 Kb of a known gene, but not inside one (37/481 = 7.7%), or (3) lying more than 10 Kb away from any known gene (167/481 =34.7%). In all cases the orthologous ultra-conserved elements from humans overlap orthologous introns, or occur on the same side of orthologous genes in rodents. The 277 elements in class (1) lie in 172 distinct genes, an average of 1.6 elements per gene. Two distinct subsets can be defined within this set: 111 exonic elements, which lie in a known gene and overlap a processed transcript (mRNA/EST) in human (found in 93 distinct genes), and 100 elements (in 61 distinct genes) that do not match any known transcript in any organism. Only 7 genes contain elements from both subsets. Among the intergenic elements, only 15/204 overlap a known processed transcript in human.

However, the most useful division of the ultra-conserved elements is into the types exonic, non-exonic and possibly-exonic. Exonic elements are defined above. Non-exonic elements are all elements that show no evidence of transcription, in the sense that no mRNA or EST from any species that is mapped to the human genome on the UCSC browser overlaps with them. The possibly-exonic are the remainder.

We use these classifications of elements to define two sets of genes that are associated with ultra-conserved elements. Type I genes are all genes that overlap exonic elements, i.e. elements that at least partially overlap the processed transcript of a known gene. Type II genes are derived from non-exonic genes. They were chosen as flanking elements of these elements according to the following rules:

- if the element is in the intron of a known gene, include that gene.
- if the element is <10kb from a known gene, include that gene, but
  if there are known genes flanking <10kb away on each side, include only closest
- if the element is >=10kb away from any known gene, include both flanking known genes.

We compared GO (*19*) and InterPro (*20*) annotations of the genes the ultra-conserved elements lie within, or next to, against the background of all annotated human genes, using the tail of the hypergeometric distribution to calculate *P*. While we did not directly correct for multiple hypothesis testing, in practice we performed less than 1,000 individual tests, deeming the reported *P*'s highly significant.

In addition to the *P*s given in the main text, note that in a manually curated subset of 59/89 annotated type I genes whose elements appear to be involved in alt-splicing, which otherwise resembles the larger set, Homeobox is no longer pronounced ($P = 0.05$). Interestingly, of the sets we examined, the Homeobox is most significantly pronounced ($P < 10^{-31}$) in the set of 394 annotated genes that flank all ultra-conserved elements, genic and intergenic.

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* (2003).
6. F. Chiaromonte *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* (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.* (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, *Proc. Intelligent Systems in Molecular Biology and Bioinformatics*, in press.
32. J. S. Mattick, M. J. Gagen, *Mol. Biol. Evol.* **18**, 1611 (2001).
33. E. T. Dermitzakis *et al.*, *Genome Res.* (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. D. Karolchik *et al.*, *Nucleic Acids Res.* **32** (database issue), D493 (2004).
40. Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289 (1995).
41. S. Schwartz *et al.*, *Genome Res.* **13**, 103 (2003).
42. W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11484 (2003).

43. M. S. Waterman, *Introduction to Computational Biology* (Chapman & Hall--CRC Press, 1995).
44. We thank the Genome Sequencing Consortia for the human, mouse, rat and other genome sequences we used in this analysis. We thank W. Miller, M. Diekhans, A. Hinrichs, K. Rosenbloom, D. Thomas and the members of the UCSC browser team for providing the genome alignments and other tracks of genome annotation available on the UCSC genome browser. We also thank 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 NHGRI grant 1P41HG02371, NCI contract 22XS013A, and D.H. additionally 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.