Supplementary materials for:

28-Way Vertebrate Alignment and Conservation Track in the UCSC Genome Browser

Miller, Rosenbloom, Hardison, Hou, Taylor, Raney, Burhans, King, Baertsch, Blankenberg, Kosakovsky Pond, Nekrutenko, Giardine, Harris, Tyekucheva, Diekhans, Pringle, Murphy, Lesk, Weinstock, Lindblad-Toh, Gibbs, Lander, Siepel, Haussler, Kent

The 28-way alignment. Figure S1 augments Figure 1 (main text) by explicitly giving the branch lengths.

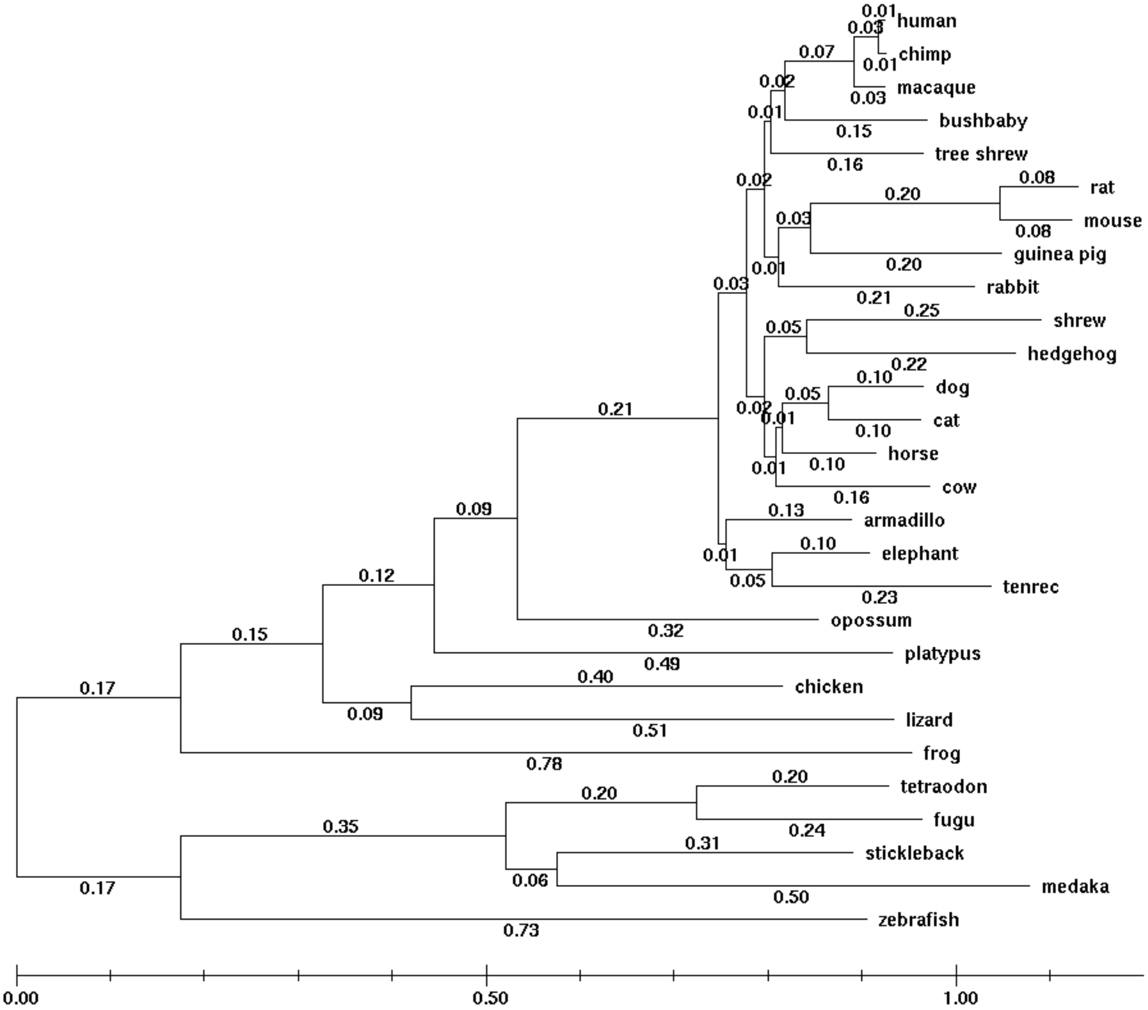


Figure S1. Phylogenetic tree of species in the 28-way alignment, giving the average number of substitutions per site on each branch.

Table S1. Additional information about the species in the 28-way alignment.

Dist = Distance from human computed from branch lengths of 4d tree
(substitutions per 100mya)

Align = %Human bases that chain blocks of this species covers

Covg = Assembly sequencing coverage

Filter: N = none, S = syntenic, R = reciprocal best

Species	Common Name	Dist	Align	Covg	Filter	UCSC ID	Sequencing Center, ID
Homo sapiens	human	0.0	100.0%	F	N	hg18	Mar 2006 NCBI Build 36.1
Pan troglodytes	chimpanzee	13.5	94.9%	6X	S	panTro2	Mar 2006 Chimp GSC Build 2.1
Macaca mulatta	rhesus	63.9	85.6%	5.1X	S	rheMac2	Jan 2006 Macaque GSC V1.0
Otolemur garnetti	bushbaby	25.6	44.6%	2X	R	otoGar1	Dec 2006 Broad 1.0
Tupaia belangeri	tree shrew	28.4	37.3%	1.5X	R	tupBel1	Dec 2006 Broad 1.0
Equus caballus	horse	28.5	57.1%	6.8X	S	equCab1	Feb 2007 Broad 1.0
Dasypus novemcinctus	armadillo	31.9	32.2%	2X	R	dasNov1	May 2005 Broad 1.0
Felis catus	cat	33.2	35.9%	2X	R	felCat3	Mar 2006 Broad 3.0
Canis familiaris	dog	33.5	52.9%	7.6X	S	canFam2	May 2005 Broad 2.0
Loxodonta africana	elephant	33.8	34.4%	2X	R	loxAfr1	May 2006 Broad 1.0
Bos taurus	cow	34.2	46.7%	7.1X	S	bosTau3	Aug 2006 Baylor 3.1
Oryctolagus cuniculus	rabbit	35.0	34.0%	2X	R	oryCun1	May 2005 Broad 1.0
Cavia porcellus	guinea pig	37.9	29.7%	2X	R	cavPor2	Oct 2005 Broad 2.0
Erinaceus europaeus	hedgehog	43.4	19.6%	2X	R	eriEur1	June 2006 Broad 1.0
Mus musculus	mouse	45.3	34.5%	F	S	mm8	Feb 2006 NCBI Build 36
Sorex araneus	shrew	46.0	20.1%	2X	R	sorAra1	June 2006 Broad 1.0
Rattus norvegicus	rat	46.1	32.9%	7X	S	rn4	Nov 2004 Baylor v3.4
Echinops telfairi	tenrec	46.8	23.6%	2X	R	echTel1	July 2006 Broad 1.0
Monodelphis domestica	opossum	71.1	12.4%	6.5X	S	monDom4	Jan 2006 Broad 4.0
Ornithorhychus anatinus	platypus	96.7	7.9%	6X	N	ornAna1	Dec 2006 WUSTL 6.0
Gallus gallus	chicken	108.7	3.6%	6.6X	N	galGal3	May 2006 Chicken GSC v2.1
Anolis carolinensis	lizard	120.6	4.8%	6.8X	N	anoCar1	Jan 2007 Broad 1.0
Xenopus tropicalis	frog	152.8	2.6%	7.9X	N	xenTro2	Aug 2005 JGI 4.1
Danio rerio	zebrafish	181.4	1.9%	6.5X	N	danRer4	Mar 2006 Sanger Zv6

Gasterosteus aculeatus	stickleback	182.9	2.0%	6X	N	gasAcu1	Feb 2006 Broad 1.0
Tetraodon nigroviridis	tetraodon	185.2	2.0%	7.9X	N	tetNig1	Feb 2004 Genoscope/Broad V7
Takifugu rubripes	fugu	188.8	1.8%	8.5X	N	fr2	Oct 2004 JGI 4.0
Oryzias latipes	medaka	200.2	2.0%	6.7X	N	oryLat1	Oct 2005 NIG 1.0

Quality values in the 28-way alignment. We added sequence quality scores, represented as extra MAF rows, to the alignment for twenty of the twenty-eight species. The quality values are based on data provided by the organization that assembled each genome. We compressed the data so that each quality value is represented by a single character and emphasized low quality values (our real concern) with the following formula:

min(floor(Original Quality Value/5), 9)

The relationship between the original quality value for each base and the value in the 28-way alignment is summarized in the following table:

Character in 28-way alignment	Original Quality Value	Probability of Error
0	$0 <= Q < 5$	$0.3162 < P <= 1$
1	$5 <= Q < 10$	$0.1 < P <= 0.3162$
2	$10 <= Q < 15$	$0.03162 < P <= 0.1$
3	$15 <= Q < 20$	$0.01 < P <= 0.03162$
4	$20 <= Q < 25$	$0.003162 < P <= 0.01$
5	$25 <= Q < 30$	$0.001 < P <= 0.003162$
6	$30 <= Q < 35$	$0.0003162 < P <= 0.001$
7	$35 <= Q < 40$	$0.0001 < P <= 0.0003162$

8	40 <= Q < 45	0.00003162 < P <= 0.0001
9	45 <= Q < 98	0.000000000158 < P <= 0.00003162

Table S2. Translation of original quality values to characters in the 28-way alignment.

In addition, special characters are used to indicate that the base has been edited by hand, namely `F’ for bases considered to be high-quality and `0’ for low-quality bases. Aligned bases in an assembly gap are assigned quality `.’.

For each of the twenty species having quality data, the following table summarizes how often each of our twelve quality values is used. As expected, species with a lower sequence coverage tend to have more low-quality bases.

Common Name	Seq. Cov.	Total Bases	F	9	8	7	6	5	4	3	2	1	0	.
Chimp	6.0x	2,704,068,384	0.000%	96.545%	0.336%	0.259%	0.212%	0.174%	0.150%	0.135%	0.137%	0.065%	1.981%	0.008%
Rhesus	5.1x	2,445,536,595	0.000%	96.264%	0.143%	0.568%	2.272%	0.331%	0.162%	0.110%	0.088%	0.047%	0.005%	0.011%
Bushbaby	2.0x	1,236,144,958	0.000%	77.706%	4.275%	3.815%	3.964%	3.839%	2.265%	1.937%	1.327%	0.727%	0.099%	0.045%
Tree shrew	1.5x	1,060,099,301	0.000%	79.600%	3.927%	3.453%	3.591%	3.517%	2.060%	1.754%	1.210%	0.668%	0.132%	0.087%
Rat	7.0x	944,090,409	0.000%	98.897%	0.372%	0.212%	0.126%	0.134%	0.080%	0.062%	0.045%	0.020%	0.041%	0.012%
Guinea pig	2.0x	837,174,260	0.000%	76.478%	4.972%	4.372%	4.281%	3.882%	2.239%	1.889%	1.217%	0.585%	0.056%	0.030%
Rabbit	2.0x	950,706,952	0.000%	80.638%	4.103%	3.623%	3.454%	3.188%	1.760%	1.507%	1.033%	0.587%	0.064%	0.043%
Shrew	2.0x	570,467,216	0.000%	83.845%	3.435%	3.120%	2.960%	2.586%	1.490%	1.244%	0.810%	0.399%	0.093%	0.017%
Hedgehog	2.0x	553,147,628	0.000%	84.636%	3.231%	2.916%	2.739%	2.426%	1.448%	1.245%	0.848%	0.459%	0.039%	0.014%
Dog	7.6x	1,537,051,022	0.000%	98.032%	0.789%	0.432%	0.236%	0.156%	0.102%	0.081%	0.068%	0.038%	0.062%	0.003%
Cat	2.0x	1,002,774,833	0.000%	87.929%	2.644%	2.152%	2.320%	1.903%	1.077%	0.875%	0.690%	0.302%	0.068%	0.040%
Horse	6.8x	1,643,549,942	0.000%	97.680%	0.601%	0.512%	0.394%	0.288%	0.163%	0.140%	0.095%	0.048%	0.075%	0.005%
Cow	7.1x	1,346,289,412	0.000%	99.306%	0.195%	0.114%	0.110%	0.093%	0.056%	0.046%	0.037%	0.017%	0.010%	0.014%
Armadillo	2.0x	900,934,775	0.000%	77.659%	4.596%	3.947%	3.945%	3.876%	2.116%	1.839%	1.266%	0.634%	0.069%	0.053%
Elephant	2.0x	969,022,725	0.000%	78.116%	4.598%	3.826%	3.841%	3.786%	2.068%	1.764%	1.227%	0.632%	0.085%	0.058%
Tenrec	2.0x	677,674,931	0.000%	79.719%	4.482%	3.841%	3.668%	3.292%	1.805%	1.527%	0.999%	0.516%	0.093%	0.058%
Opossum	6.5x	320,665,980	0.000%	96.901%	1.598%	0.522%	0.323%	0.227%	0.140%	0.113%	0.088%	0.046%	0.039%	0.002%
Lizard	6.8x	131,303,519	0.000%	97.367%	0.751%	0.549%	0.426%	0.305%	0.162%	0.127%	0.074%	0.034%	0.206%	0.000%
Stickleback	6.0x	52,104,953	0.000%	96.444%	0.852%	0.713%	0.580%	0.444%	0.298%	0.267%	0.190%	0.097%	0.116%	0.000%
Medaka	6.7x	46,565,709	0.000%	99.257%	0.301%	0.181%	0.091%	0.055%	0.044%	0.025%	0.020%	0.002%	0.008%	0.016%

Table S3. Percentages of bases assigned various quality scores.

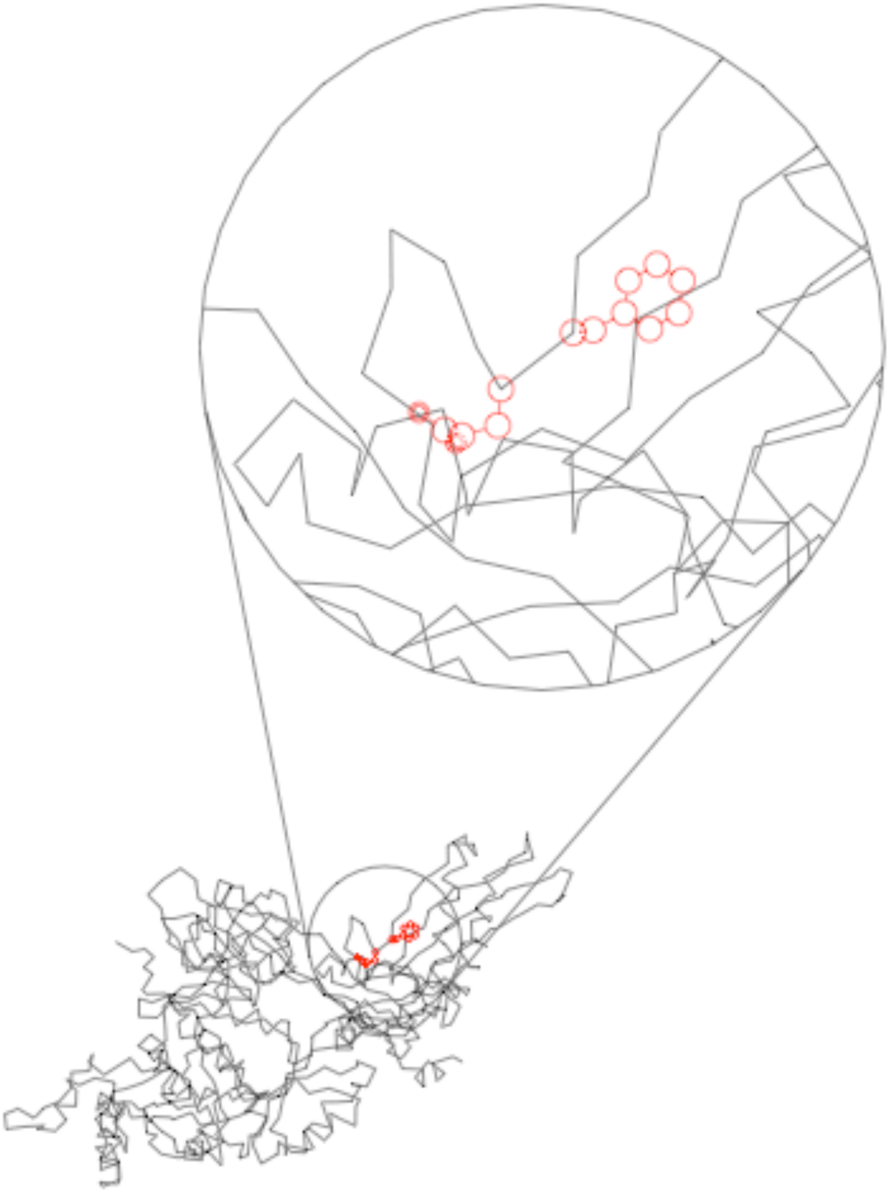
Insertions/deletions in coding regions. Figure S2 greatly strengthens the evidence given in Figure 2 (main paper) to support the claim that humans diverged from dogs (and many other mammals) before they diverged from rodents. The 2-amino-acid deletion in the signal peptide of the *PRNP* gene appears to have happened only once in eutherian evolution.

`MA--NLGCWMLFLFVATWSDLGLCKKRPKPG`	*Callithrix jacchus*	Common marmoset
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Cebus apella*	Brown capuchin monkey
`MA--NLGCWMLVVFVATWSDLGLCKKRPKPG`	*Cercopithecus aethiops*	Vervet monkey
`MA--NLGCWMLVVFVATWSDLGLCKKRPKPG`	*Cercopithecus diana*	Diana monkey
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Colobus guereza*	Mantled guereza monkey
`ME--NLGCWMLILFVATWSDIGLCKKRPKPG`	*Cynocephalus variegatus*	Malayan flying lemur
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Gorilla gorilla*	Western gorilla
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Homo sapiens*	Human
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Hylobates lar*	White-handed gibbon
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Macaca arctoides*	Stump-tailed macaque
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Macaca fascicularis*	Long-tailed macaque
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Macaca fuscata*	Japanese macaque
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Macaca mulatta*	Rhesus monkey
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Macaca nemestrina*	Pigtail macaque
`MA--NLGCWMLVVFVATWSDVGLCKKRPKPG`	*Microcebus murinus*	Gray mouse lemur
`MA--RLGCWMLVLFVATWSDIGLCKKRPKPG`	*Otolemur garnettii*	Small-eared galago
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Pan troglodytes*	Chimpanzee
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Papio hamadryas*	Hamadryas baboon
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Pongo pygmaeus*	Bornean orangutan
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Presbytis francoisi*	Francois' langur
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Saimiri sciureus*	South American squirrel monkey
`MA--NLGCWMLVLFVATWSDLGLCKKRPKPG`	*Symphalangus syndactylus*	Siamang gibbon
`MA--QLGCWLMVLFVATWSDVGLCKKRPKPG`	*Tupaia belangeri*	Northern tree shrew
`MA--NLGYWLLALFVTTWTDVGLCKKRPKPG`	*Apodemus sylvaticus*	Long-tailed field mouse
`MA--NAGCWLLVLFVATWSDTGLCKKRPKPG`	*Cavia porcellus*	Guinea pig
`MA--HLSYWLLVLFVAAWSDVGLCKKRPKPG`	*Ochotona princeps*	American pika
`MA--NLGCWLLVLFVATWSDLGLCKKRTKPG`	*Dipodomys ordii*	Ord's kangaroo rat
`MA--NLSYWLLAFFVTTWTDVGLCKKRPKPG`	*Clethrionomys glareolus*	Bank vole
`MA--NLSYWLLALFVATWTDVGLCKKRPKPG`	*Cricetulus griseus*	Chinese hamster
`MA--NLSYWLLALFVATWTDVGLCKKRPKPG`	*Cricetulus migratorius*	Gray dwarf hamster
`MA--NLGYWLLALFVTMWTDVGLCKKRPKPG`	*Meriones unguiculatus*	Mongolian gerbil
`MA--NLSYWLLALFVAMWTDVGLCKKRPKPG`	*Mesocricetus auratus*	Golden hamster
`MA--NLGYWLLALFVTMWTDVGLCKKRPKPG`	*Mus musculus*	House mouse
`MA--HLGYWMLLLFVATWSDVGLCKKRPKPG`	*Oryctolagus cuniculus*	European rabbit
`MA--NLGYWLLALFVTTCTDVGLCKKRPKPG`	*Rattus norvegicus*	Brown rat
`MA--NLGYWLLALFVTTCTDVGLCKKRPKPG`	*Rattus rattus*	Black rat
`MA--NLGYWLLALFVATWTDVGLCKKRPKPG`	*Sigmodon fulviventer*	Tawny-bellied cotton rat
`MA--NLGYWLLALFVATWTDVGLCKKRPKPG`	*Sigmodon hispiedis*	Hispid cotton rat
`MV--NPGCWLLVLFVATLSDVGLCKKRPKPG`	*Spermophilus tridecemlineatus*	13-lined ground squirrel
`MV--NPGYWLLVLFVATLSDVGLCKKRPKPG`	*Sciurus vulgaris*	Eurasian red squirrel
`MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG`	*Bos taurus*	Domestic cow
`MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG`	*Bison bison*	American bison

MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Rangifer tarandus*	Caribou
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Alces alces*	Eurasian elk
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Capreolus capreolus*	Western roe deer
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Kobus megaceros*	Nile lechwe antelope
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Connochaetes taurinus*	Blue wildebeest
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Ammotragus lervia*	Barbary sheep
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Hippotragus niger*	Sable antelope
MVKSHMGSWILVLFVVTWSDVGLCKKRPKPG	*Camelus dromedarius*	Dromedary
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Capra hircus*	Domestic goat
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Cervus elaphus*	Elk
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Cervus elaphus nelsoni*	Nelson's elk
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Dama dama*	Fallow deer
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Odocoileus hemionus*	Mule deer
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Odocoileus virginianus*	White-tailed deer
MVKSHIGSWILVLFVAMWSDVALCKKRPKPG	*Oryx leucoryx*	Arabian oryx
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Ovibos moschatus*	Muskox
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Ovis aries*	Domestic sheep
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Ovis canadensis*	Bighorn sheep
MVKSHIGGWILVLFVAAWSDIGLCKKRPKPG	*Sus scrofa*	Pig
MVKSHIGSWILVLFVAMWSDVALCKKRPKPG	*Tragelaphus strepsiceros*	Greater kudu
MVKSHIGGWILLLFVATWSDVGLCKKRPKPG	*Canis lupus familiaris*	Dog
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Felis catus*	Domestic cat
MVKSHIGSWLLVLFVATWSDIGFCKKRPKPG	*Mustela putorius*	European polecat
MVKSHIGSWLLVLFVATWSDIGFCKKRPKPG	*Mustela vison*	American mink
MVKSHIGSWILVLFVAMWSDVGLCKKRPKPG	*Ailuropoda melanoleuca*	Giant panda
MVKSHVGGWILVLFVATWSDVGLCKKRPKPG	*Equus caballus*	Horse
MVRSHVGGWILVLFVATWSDVGLCKKRPKPG	*Diceros bicornis*	Black rhinoceros
MVKNHVGCWLLVLFVATWSEVGLCKKRPKPG	*Erinaceus europaeus*	Western European hedgehog
MVTGHLGCWLLVLFMATWSDVGLCKKRPKPG	*Sorex araneus*	Eurasian shrew
MVKSLVGGWILLLFVATWSDVGLCKKRPKPG	*Myotis lucifugus*	Little brown bat
MVKNYIGGWILVLFVATWSDVGLCKKRPKPG	*Pteropus vampyrus*	Large flying fox (a fruit bat)
MVKSHIANWILVLFVATWSDMGFCKKRPKPG	*Tursiops truncatus*	Bottlenose dolphin
MVKSHLGCWIMVLFVATWSEVGLCKKRPKPG	*Cyclopes didactylus*	Silky anteater
MVKGTVSCWLLVLVVAACSDMGLCKKRPKPG	*Echinops telfairi*	Lesser hedgehog tenrec
MVKSSLGCWILVLFVATWSDMGLCKKRPKPG	*Elephas maximus*	Asiatic elephant
MVKSSLGCWILVLFVATWSDMGLCKKRPKPG	*Loxodonta africana*	African bush elephant
MVKSSLGCWMLVLFVATWSDVGLCKKRPKPG	*Procavia capensis*	Rock hyrax
MMKSGLGCWILVLFVATWSDVGLCKKRPKPG	*Orycteropus afer*	Aardvark
MVKSGLGCWILVLFVATWSDVGVCKKRPKPG	*Trichechus manatus*	West Indian manatee
MVRSRVGCWLLLLFVATWSELGLCKKRPKPG	*Dasypus novemcinctus*	9-banded armadillo
MAKIQLGYWILALFIVTWSELGLCKKPKTRPG	*Macropus eugenii*	Tammer wallaby
MGKIHLGYWFLALFIMTWSDLTLCKKPKPRPG	*Monodelphis domestica*	Gray short-tailed opossum
MGKIQLGYWILVLFIVTWSDLGLCKKPKPRPG	*Trichosurus vulpecula*	Common brushtail possum
MARLLTTCCLLALLLAACTDVALSKKGKGKPS	*Gallus gallus*	Chicken
MAKLPGTSCLLLLLLLLGADLASCKKGKGKPG	*Taeniopygia guttata*	Zebra finch
MGKHQMTCWLAIFLLLIQANVSLAKK-KPKPS	*Anolis carolinensis*	Green anole
MRRFLVTCWIAVFLILLQTDVSLSKKGKNKPG	*Gekko gekko*	Tokay gecko
MGRYRLTCWIVVLLVVMWSDVSFSKKGKGKGG	*Trachemys scripta*	Pond slider (a turtle)
MGRHLISCWIIVLFVAMWSDVSLAKKGKGKTG	*Pelodiscus sinensis*	Chinese softshell turtle
MPQSLWTCLVLISLICTLTVSSKKSGGGKSKTG	*Xenopus laevis*	African clawed frog
MLRSLWTSLVLISLVCALTVSSKKSGSGKSKTG	*Xenopus tropicalis*	(a related frog; model amphibian)

Figure S2. Extension of Figure 2 to include many more species.

Human-specific protein indels. Figure S3 shows the structural model that was mentioned in the main paper.



gfm2 model

Figure S3. Model of the three dimensional conformation of the human *GFM2* gene, showing the location of a 6-bp insertion that is unique to humans. The insertion site looks unlikely to be constrained by secondary or tertiary structural interactions.

Pan-vertebrate UCEs. We studied the 481 human-rodent ultraconserved regions (Bejerano et al. 2004), looking at human-fish conservation in these regions. For each region, we computed the maximum identity percentage between human and any of the five sequenced fish (tetraodon, fugu, stickleback, medaka, and zebrafish). We then grouped the regions according to this value and looked for relationships among the annotated human genes nearest to each member in a group.

Comparing the UCEs with no conservation to fish (77 regions, 59 unique nearest genes) vs. those highly conserved to fish (>=80% identical, 127 regions, 95 genes), we found significant enrichment for two GO terms among the genes for the highly conserved UCEs

p-value	high vs. combined	GO term	
7.9e-4	57/89 vs. 70/131	GO:0065007	biological regulation
1.1e-3	52/89 vs. 63/131	GO:0003677	DNA binding

(Note that the denominators are reduced from 95 to 89 and 154 to 131 since GOstat does not recognize all gene/transcript names). Applying the Benjamini FDR correction to these sets raises both *p*-values to 0.28.

We also looked for relationships involving the distance to the nearest gene, and found that UCEs with no fish conservation were significantly further from genes than those with any fish conservation, and that fewer of them coincide with a gene. 59% of all UCEs coincide with a gene, while this drops to 51% for those with no fish conservation. This is under-representation with *p*-value 7.8e-2 (39/77 vs. 282/481).

For regions not intersecting with a gene, the average distance to the nearest gene is 309K bases for the regionw with zero fish-conservation, and 164K for the those above zero, an increase of 88%. This increase is dampened by the presence of a few outliers in the non-zero regions. If we instead compare median distance the ratio is 270K to 73K, an increase of 270%.

Phylogenetic extent of alignability of functional regions.

Table S4. Fits of alignability, *A(x)*, for different classes of genomic features to decay equations as functions of phylogenetic distance, *x*, for all comparison species.

Fit to	Feature Class	Decay fit begins with comparison to	a	b	c	R^2
$A(x) = a + b*e^{(-cx)}$	Background	Horse	0.01	2.10	3.74	0.99
	CodingExons	Horse	0.57	0.67	1.24	0.91
	Known Reg. Regions	Mouse	0.006	2.04	1.63	0.97
	PRPs	Mouse	-0.21	1.98	1.11	0.99
	pTRRs	Horse	-0.02	1.93	2.29	0.99
	CTCF binding sites	Horse	0.03	1.80	2.87	0.98
$A(x) = a + b*x + c*x^2$	UCEs	Horse	0.98	0.12	-0.16	0.96

The distance, *x*, is the REV-modeled substitutions per 4D site along the branches from human to the other species.