Ultraconservation identifies a small subset of extremely constrained developmental enhancers

Axel Visel[1], Shyam Prabhakar[1], Jennifer A. Akiyama[1], Malak Shoukry[1], Keith D. Lewis[1], Amy Holt[1], Ingrid Plajzer-Frick[1], Veena Afzal[1], Edward M. Rubin[1,2], and Len A. Pennacchio[1,2,*]

[1]Genomics Division, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA.

[2]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598 USA.

* To whom correspondence should be addressed: Len A. Pennacchio, Genomics Division, One Cyclotron Road, MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94720. Email: LAPennacchio@lbl.gov, Phone: (510) 486-7498, Fax: (510) 486-4229.

First paragraph

While experimental studies have suggested that non-coding ultraconserved DNA elements are central nodes in the regulatory circuitry that specifies mammalian embryonic development, the possible functional relevance of their >200bp of perfect sequence conservation between human-mouse-rat remains obscure [1,2]. Here we have compared the *in vivo* enhancer activity of a genome-wide set of 231 non-exonic sequences with ultraconserved cores to that of 206 sequences that are under equivalently severe human-rodent constraint (ultra-like), but lack perfect sequence conservation. In transgenic mouse assays, 50% of the ultraconserved and 50% of the ultra-like conserved elements reproducibly functioned as tissue-specific enhancers at embryonic day 11.5. In this *in vivo* assay, we observed that ultraconserved enhancers and constrained non-ultraconserved enhancers targeted expression to a similar spectrum of tissues with a particular enrichment in the developing central nervous system. A human genome-wide comparative screen uncovered ~2,600 non-coding elements that evolved under ultra-like human-rodent constraint and are similarly enriched near transcriptional regulators and developmental genes as the much smaller number of ultraconserved elements. These data indicate that ultraconserved elements possessing absolute human-rodent sequence conservation are not distinct from other non-coding elements that are under comparable purifying selection in mammals and suggest they are principal constituents of the *cis*-regulatory framework of mammalian development.

Main Text

The last common ancestor of human and rodents lived ~75 million years ago [3] and yet the human genome contains 256 non-coding "ultraconserved" elements of ≥200bp that are perfectly conserved in mouse/rat presumably due to extreme purifying selection [1]. Their depletion in segmental duplications and copy number variant regions [4] as well as their reduced frequency of derived alleles in the human population [2,5] further point toward a pivotal functional role of these elements. In sharp contrast, the identity of ultraconserved elements as a distinct class of genomic function has been challenged by the observation that more rigorous comparative genomic methods (e.g., 6,7) can identify additional sequences with similar conservation properties by some measures, but lacking extended perfect sequence conservation. Moreover, while human-rodent, human-mouse-dog and human-chicken genome comparisons each identify several hundred ultraconserved elements, there is limited overlap in the catalogs of elements identified by these comparisons [4]. Another feature of ultraconserved elements that undercuts their relevance as a distinct class is that they are almost invariably embedded in larger blocks of constrained sequence, suggesting that they exist not as independent units of biological function but as somewhat arbitrary fragments of larger functional modules. In the absence of comprehensive experimental data it remains unclear whether the absolute sequence conservation of ultraconserved elements is indicative of a unique role or if they are merely a functionally indistinct fraction of a much larger set of extremely constrained elements.

To explore the functional uniqueness of non-coding ultraconserved elements, we identified a large number of human-rodent conserved elements that are under similar evolutionary constraint as regions containing ultraconservation. We compared the entire set of these

elements, the majority of which lack perfectly conserved regions of ≥200bp, to the small subset that overlap ultraconserved elements to identify possible properties specifically associated with ultraconservation, including their degree of constraint in other mammalian species and their enrichment near genes with certain functions. Moreover, we examined the ability of a genome-wide set of non-exonic ultraconserved elements and more than 200 ultra-like constrained elements to drive tissue-specific *in vivo* expression in transgenic mouse embryos, a property that has previously been proposed to be a predominant function associated with non-coding ultraconservation [8-11].

In an initial comparative genomic assessment of ultraconservation, we found substitutions in 79% of these elements in other mammalian species (Fig. 1), indicating that their absolute conservation between human and rodents is at least partially a matter of ascertainment bias, rather than absolute intolerance of nucleotide substitutions. This finding further challenges the possible uniqueness of ultraconserved elements and raises the possibility that they represent only a subset of a larger group of elements with similar properties. In an attempt to identify elements with ultra-like conservation, we used a statistical measure of human-mouse-rat constraint [12] with scoring parameters optimized through multiple genome-wide scans (see methods) to generate a constraint-ranked set of conserved non-coding sequences. When we compared these elements to the distribution of non-exonic ultraconserved elements, we found that the constraint scores of ultraconserved regions are distributed over a surprisingly wide range and a much larger number of elements appear similarly constrained (Fig. 2). We identify a population of 2,614 human-rodent constrained elements that overlap or include 234 (91%) of all 256 non-exonic ultraconserved elements. To ascertain the ultra-like conservation of these elements independently from the scoring scheme used for their

identification, we determined their branch length and rejected substitution counts [6] in human, rodents and five additional mammalian species (Suppl. Fig. 1). We find that extremely constrained elements that contain or do not contain regions of ultraconservation have similar characteristics by these two widely used comparative genomic measures, confirming their ultra-like nature. While an order of magnitude more numerous than non-exonic ultraconserved elements, the highly constrained non-coding regions identified here are enriched near genes of a small subset of functional categories. As for ultraconserved elements, these functions include transcriptional regulation and development [1] and, in particular, development of the nervous system (Fig. 3; see suppl. table 4 for a list of all significantly enriched functions). Taken together, comparative analysis as well as the genome-wide distribution suggest that ultraconservation identifies a small subset of genome regions that are equally constrained and have similar properties, but the majority of which lack regions of ultraconservation.

To test whether such apparent equivalence at the sequence level is also associated with similar functional properties, we focused on transcriptional enhancer activity during embryonic development. We used a transgenic mouse assay to determine the embryonic *in vivo* enhancer activities of 155 human genome regions that include non-coding ultraconserved elements and combined these data with a previously reported smaller data set [10] to establish a genome-wide compendium of enhancer activities for this class of non-coding elements (suppl. table 1). A total of 231 transgenic assays was considered, in which the tested human genome fragments included 245 of all 256 non-exonic ultraconserved elements (12 constructs contained 2 or 3 adjacent ultraconserved regions). Only elements that drove reporter gene expression reproducibly in the same anatomical structure in at least

three e11.5 mouse embryos resulting from independent transgenic integration events were considered enhancers. We found that half (115/231) of the ultraconserved regions drove reporter gene expression in various tissues of the developing mouse embryo, often in a tightly spatially restricted manner and with subregions of the central nervous system among the most frequently targeted structures (Fig. 4a).

To determine whether such an enrichment in embryonic enhancers is specifically associated with the presence of ultraconserved regions in highly human-rodent constrained sequences, we also tested the enhancer activities of 206 non-coding sequences that have ultra-like human-rodent constraint scores, but lack regions of ultraconservation. Of note, these regions were selected blind to evolutionary conservation depth (i.e. detectable sequence conservation in non-mammalian species), but purely based on their human-rodent constraint scores. Using identical scoring criteria as for the ultraconserved elements, we found that 102 of these 206 elements (50%) are tissue-specific enhancers at e11.5. As with ultraconserved elements, the patterns driven by these enhancers are highly reproducible among embryos resulting from different transgene integration events and often highly restricted in their spatial boundaries (Suppl. Fig. 2). We did not find significant differences between the ultraconserved and non-ultraconserved elements regarding the overall distribution of the targeted anatomical structures (Fig. 4a). We observed multiple cases of ultraconserved and non-ultraconserved elements driving virtually identical patterns when scrutinized at higher resolution (Fig. 4b), as well as dozens of patterns driven by non-ultraconserved elements for which no counterpart was found among ultraconserved elements (Suppl. Fig. 2), highlighting the value of ultra-like constraint for the discovery of tissue-specific reagents. Our findings indicate that extreme human-rodent constraint

identifies genome regions that are in their entirety highly enriched in embryonic enhancers, while the ultraconserved subset within this population was neither found to be enriched in enhancers targeting specific tissues nor to be generally more enriched in developmental enhancers.

Ultraconserved elements appear to have become virtually "frozen" during mammalian evolution [1] and their perfect, uninterrupted sequence identity between human and rodents is suggestive of them representing the pinnacle of extreme non-coding sequence conservation in mammals. The identification of several thousand elements with ultra-like human-rodent constraint indicates, however, that the relatively small number of ultraconserved elements may be more likely due to their definition by a simple percent-identity-plot approach [13] than to a uniquely high degree of constraint. If enrichment in enhancer activity is considered as a measure, a direct comparison within the population of extremely human-rodent constrained elements identified in this study indicates that non-coding ultraconserved elements do not represent the very tail of a distribution spectrum of human-rodent conservation, but merely a subset of a ten-fold larger population of elements under similar constraint and with equivalent regulatory function. Through the analysis of embryonic *in vivo* enhancer activities of more than 400 of these elements, this study provides a window into a portion of the human *cis*-regulome that appears to be severely constrained throughout the mammalian clade. Since these elements are defined independent of their conservation in non-mammalian vertebrate species, we expect that hundreds of additional tissue-specific enhancers remain to be discovered in this category of extreme conservation and many will be unique to mammals and thus not be detectable by comparison with evolutionarily more distant vertebrate species. The association of extreme human-rodent sequence conservation

with enhancer activity, independent from presence of ultraconservation, suggests that the population of ultra-like constrained elements identified here constitutes a core *cis*-regulatory framework of mammalian development.

Acknowledgements:

The authors thank Inna Dubchak, Alexander Poliakov and Simon Minovitsky for help with genome alignments and database programming; Sumita Bhardwaj and Sengthavy Phouanenavong for technical assistance; Nadav Ahituv, Marcelo Nobrega, James Noonan and members of the Pennacchio and Rubin laboratories for discussion and critical comments on the manuscript. L.A.P. was supported by grant HL066681, Berkeley-PGA, under the Programs for Genomic Applications, funded by National Heart, Lung, & Blood Institute, and HG003988 funded by National Human Genome Research Institute. Research was performed under Department of Energy Contract DE-AC02-05CH11231, University of California, E.O. Lawrence Berkeley National Laboratory. A.V. was supported by an American Heart Association postdoctoral fellowship.

Methods

Substitutions in mammalian species. We examined the 256 non-coding human-rodent ultraconserved elements for substitutions relative to human in chimpanzee, rhesus, dog, horse or cow using the 28-species vertebrate multiz alignment [14] available from the UCSC Genome Bioinformatics website [15]. This alignment includes the following genome assemblies used in this analysis: hg18 (human, NCBI Build 36.1), panTro2 (Chimpanzee Sequencing and Analysis Consortium, March 2006 assembly), rheMac2 (Macaque Genome Sequencing Consortium, January 2006 assembly), canFam2 (Broad Institute and Agencourt Bioscience, May 2005 assembly), bosTau3 (Baylor College of Medicine, August 2006 assembly) and equCab1 (Broad Institute, January 2007 assembly). Positions in the human genome that were substituted in multiple other lineages were nevertheless counted only once. Only aligned positions of high sequence quality (quality score ≥30) were included in the count of human-rodent ultraconserved positions substituted in other mammalian lineages (Fig. 1). Although this quality-score filter eliminates the majority of errors, it is still possible that a small fraction of the human-rodent ultraconserved positions that appear to be substituted in other mammals are actually sequencing artifacts. To estimate the impact of sequencing errors at high-quality positions, we assumed that each such non-human position generated a spurious mismatch with probability given by $10^{-(\text{quality-score}/10)}$ (see [16]) which yielded an estimated total of 2.5 spurious mismatches over all 256 non-exonic ultraconserved elements, far smaller than the observed total of 673 mismatched positions.

Identification of human-rodent constrained elements. Whole-genome global alignment of human (hg16, NCBI Build 34), mouse (mm4, NCBI Build 32) and rat (rn3, Rat Genome Sequencing

Consortium, June 2003 assembly) was performed using MLAGAN [17] and the VISTA alignment pipeline [18]. Evolutionarily conserved regions in this alignment were identified using Gumby [12], and conserved regions were filtered for any overlap with UCSC genes, RefSeq genes, RNA genes, sno/microRNAs, human mRNAs or spliced ESTs (as annotated at http://genome.ucsc.edu) to define a genome-wide set of human-rodent conserved non-coding sequences. This procedure was tested at multiple settings of the Gumby R-ratio parameter (expected local-neutral/conserved branch length ratio). To identify a genome-wide set of elements with ultra-like conservation, a value of R=50 was chosen (see supplemental methods), and a P-value threshold of 1e-40 yielded 2,614 non-coding human-rodent ultra-like elements genome-wide.

Branch length, rejected substitution counts and distribution correction. Phylogenetic branch lengths (substitution rates) were estimated for each conserved element using fastDNAml [19] and the above-mentioned 28-way whole-genome alignment, after masking nucleotides of low sequence quality in the draft genome assemblies. Local "neutral" substitution rates were similarly estimated based on aligned non-coding non-conserved positions within the 10-kb flanks of conserved elements (see supplemental methods). Rejected substitutions in each lineage were calculated as the product of the length of the conserved element in human and the difference between the within-element and local neutral substitution rates. When binned by conservation P-value, the ultra-like conserved elements broadly resembled the subset within the same bin that overlap ultraconserved elements (data not shown). A simple average of evolutionary rates across all 2,614 conserved elements would not reflect this, since their P-value distribution is skewed towards the high end, relative to the subset that overlaps ultraconserved regions (Fig. 2). We therefore corrected for this distribution bias by dividing

ultra-like conserved elements among 10 P-value bins, and calculating the average of bin averages, rather than the average over all individual measurements. In effect, distribution-corrected statistics summarize the average relation within any given P-value bin between all ultra-like elements and the subset overlapping ultraconserved elements.

Gene Ontology (GO) analysis of neighboring genes. Each non-coding element was assigned to the nearest neighboring RefSeq gene based on distance to the 5' or 3' end of the transcript, resulting in 851 unique genes neighboring ultra-like constrained elements and 162 unique genes neighboring non-exonic ultraconserved elements. The expected number of neighbor genes with any particular GO biological process annotation, as well as the binomial enrichment P-value relative to this expectation, was calculated using L2L [20]. Standard deviations about the expected value were based on the approximation that random sampling yields a Poisson distribution of genes in any particular GO category.

Cloning of highly constrained regions. All enhancer candidate regions were PCR amplified from human genomic DNA (Clontech) using primers designed to amplify the regions listed in supplemental tables 1 and 2. Where possible, primers were designed to include several hundred base pairs of the sequence flanking the ultraconserved and/or highly constrained core region. In 12 cases, neighboring ultraconserved regions were amplified and assayed in a single construct, resulting in 231 constructs encompassing 245 of the 256 non-exonic ultraconserved elements originally described [1]. 206 additional highly constrained elements were selected based on their extreme P-values (average genome-wide rank 343) and the absence of overlap with regions of ultraconservation, but blind to the identity of neighboring genes or their conservation in other species than human/mouse/rat. All PCR fragments were cloned into pENTR (Invitrogen), transfered into an Hsp68 promoter-LacZ reporter

vector containing a Gateway cassette using LR recombination (Invitrogen; [10,21,22]) and sequence validated.

Transgenic enhancer assay. Transgenic mice were generated as previously described [21] in accordance with protocols approved by the Lawrence Berkeley National Laboratory. Embryos were collected at e11.5 and stained for LacZ activity. A minimum reproducibility of 3 embryos resulting from independent transgenic integration events with the same staining pattern in at least one anatomical structure was required for positive elements. If no consistent pattern was observed although a minimum of 5 transgenic embryos was obtained (in absence of LacZ activity confirmed by yolk sac genotyping), elements were defined as negative. Detailed imagery and anatomical annotations for all enhancers are available at http://enhancer.lbl.gov.

Supplemental Methods

Gumby R parameter. The test statistic used by the Gumby algorithm to assess the statistical significance (P-value) of an evolutionarily constrained element is a heuristic likelihood-ratio score. This log-odds score compares the likelihood of the observed aligned segment under constrained (slower than neutral) evolution to that under the local neutral substitution rate. It is therefore necessary to define in advance the degree of constraint one expects to observe by specifying the value of the Gumby R parameter, which is the expected factor by which the local neutral rate exceeds the rate of constrained evolution in functional sequences (strictly speaking, Gumby is parametrized by the ratio of mismatch frequencies, which is not exactly the same as the ratio of substitution rates, though the difference is small in the case of eutherian sequence comparisons). The actual constrained sequences identified in the human genome at any given setting of R will not necessarily evolve R times slower than the local neutral rate. In other words, the R-value of any particular constrained sequence identified by Gumby (R^{obs}) is not the same as the parametric R used to score all constrained sequences in the genome (R^{par}). However, setting R^{par} to a higher value will indeed shift the spectrum of identified constrained elements towards higher R^{obs}. In the limit, as R^{par} tends to infinity, no substitutions or indels will be allowed inside constrained elements, and ultraconserved elements will be detected as the most significantly conserved set in the genome.

Extremely conserved elements at different settings of R^{par}. We set an extreme P-value threshold of 1e-40, and evaluated the whole-genome sets of constrained non-coding elements obtained at 6 different settings of R^{par}, ranging from 5 to 10,000 (Supp. Table 5). The 6 whole-genome

sets of extremely constrained non-coding elements are available at http://pga.jgi-psf.org/Gumby/CNS_sets. The number of extremely conserved elements detected at this P-value threshold decreased monotonically from 5,467 at $R^{par} = 5$ to 919 at $R^{par} = 10{,}000$. In contrast, the number of the 256 non-exonic human-rodent ultraconserved sequences overlapped by these elements remained approximately the same, fluctuating in the range from 215 to 234.

Comparison of extremely conserved elements to human-rodent non-exonic ultraconserved elements (nUCs). In order to compare the evolutionary properties of all extremely conserved non-coding elements to those of the subset that overlap human-rodent ultraconserved sequences, we examined three measures of evolutionary constraint: human element length L, R^{obs} and rejected substitutions [6]. R^{obs} was calculated for each conserved element as the local "neutral" (background) substitution rate S^{BG} divided by the substitution rate S^c within the element. S^{BG} was calculated for each element by maximum likelihood from the alignment of all non-conserved non-coding positions within 10 kb of either edge of the conserved element, where "non-coding" means not contained within human UCSC genes, RefSeq genes, RNA genes, sno/microRNAs, mRNAs or spliced ESTs and "non-conserved" means not contained within "most conserved" elements defined by phastCons [7] in the 28-way multiz alignment available from the UCSC Genome Browser (http://genome.ucsc.edu). Overlapping flanking regions of neighboring conserved elements were merged. Since the extreme and ultraconserved elements were identified on the basis of human-rodent alignments, evolutionary rates estimated in those lineages are subject to ascertainment bias. S^{BG}, S^c and R^{obs} were therefore estimated solely on the basis of branches in the human-rhesus-mouse-rat-dog-horse-cow phylogenetic tree that are not contained in the human-mouse-rat tree.

Positions of low sequence quality (quality score < 30) in rhesus, dog, horse and cow were masked in all sequence alignments. The rejected substitution count S^{rej} for each conserved element was estimated as $S^{rej} = (S^{BG}-S^{C}) \times L$.

For each set of extreme conserved elements defined by a particular value of R^{par}, the distribution-corrected (see methods) average length, R^{obs} and rejected-substitution count S^{rej} were calculated, both for the entire set of conserved elements and for the subset of conserved elements that overlaps nUCs. In order to maintain consistency in the accuracy of bin averages used in the distribution correction procedure, conserved elements were divided among 10 P-value bins in such a way that each bin contained approximately the same number of nUC-overlapping conserved elements. We shall from now on refer exclusively to distribution-corrected averages of non-coding element length, R^{obs} and S^{rej}.

As expected, the average length of conserved elements (i.e., the average of bin averages) decreased monotonically as R^{par} was increased from 5 to 10,000, while the degree of constraint R^{obs} of the conserved elements increased (supp. table 5). The average rejected substitution score decreased significantly with increasing R^{par}, since rejected substitutions depend more strongly on element length than on degree of constraint, once the degree of constraint R^{obs} exceeds 2. Conserved elements as a whole were generally 3-6% longer than their nUC-overlapping subset though 14-28% less constrained in terms of substitutions per nucleotide. The disparity in per-nucleotide constraint between all ultra-like elements and those that overlap nUCs decreased as R^{par} was increased from 5 to 10,000, with the average R^{obs} of all ultra-like elements reaching 86% of the value for ultra-like elements overlapping nUCs. This trend towards convergence of the two sets at high values of R^{par} is as expected, since ultra-like elements become 100% conserved, i.e. ultraconserved as R^{par} tends to infinity.

Finally, the set of all ultra-like conserved elements is marginally more constrained on average than the subset overlapping nUCs by the rejected substitution criterion (supp. table 5), with the difference ranging from 1-6%. This is largely due to the shorter length of nUC-overlapping conserved elements relative to all conserved elements in any given P-value bin, but also partly because ultraconserved elements tend to lie in regions of marginally slower neutral evolutionary rate (data not shown).

Optimizing R^{par} for Gumby whole-genome run. To determine a suitable value of R^{par} for defining a single whole-genome set of extremely conserved elements, two criteria were considered: a) similarity of constraint between all conserved elements and nUC-overlapping elements and b) enhancer predictivity. Here, enhancer predictivity was defined as the correlation coefficient between enhancer status of a conserved element (0 for negatives and 1 for positives in the transgenic assay) and its Gumby conservation score ($\log_{10}(1/\text{P-value})$). To quantify enhancer predictivity, we examined the results of all enhancer assays of ultraconserved elements, together with additional elements reported in ref. 10. Enhancer predictivity of the conservation score was found to decrease monotonically from 0.29 to 0.23 as R^{par} was increased from 5 to 10,000 (supp. table 5). Thus, enhancer predictivity favors low values of R^{par}. Indeed, it is evident from the lack of a significant correlation between ultraconserved element length and success in the enhancer assay (data not shown) that extremely high values of R^{par}, which have the effect of scoring conserved elements by the length of the largest perfectly conserved block, result in poorer enhancer predictivity. On the other hand, similarity of constraint between all ultra-like elements and the subset that overlaps nUCs favors high settings of R^{par}, if one uses R^{obs} as the measure of constraint. As a tradeoff between enhancer predictivity and per-nucleotide constraint, an intermediate value

of R^{par}=50 was therefore selected for defining the reference set of extremely human-rodent-constrained non-coding sequences in the human genome.

References

1. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* 304, 1321-5 (2004).
2. Katzman, S. et al. Human genome ultraconserved elements are ultraselected. *Science* 317, 915 (2007).
3. Waterston, R.H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-62 (2002).
4. Derti, A., Roth, F.P., Church, G.M. & Wu, C.T. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nature Genetics* 38, 1216-1220 (2006).
5. Drake, J.A. et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* 38, 223-7 (2006).
6. Cooper, G.M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15, 901-13 (2005).
7. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 1034-50 (2005).
8. de la Calle-Mustienes, E. et al. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* 15, 1061-72 (2005).
9. Poulin, F. et al. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* 85, 774-81 (2005).
10. Pennacchio, L.A. et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499-502 (2006).
11. Ghanem, N. et al. Distinct cis-regulatory elements from the Dlx1/Dlx2 locus mark different progenitor cell populations in the ganglionic eminences and different subtypes of adult cortical interneurons. *J Neurosci* 27, 5012-22 (2007).
12. Prabhakar, S. et al. Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res* 16, 855-63 (2006).

13. Hardison, R.C., Oeltjen, J. & Miller, W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res* 7, 959-66 (1997).

14. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14, 708-15 (2004).

15. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* 31, 51-4 (2003).

16. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8, 186-94 (1998).

17. Brudno, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 13, 721-31 (2003).

18. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32, W273-9 (2004).

19. Olsen, G.J., Matsuda, H., Hagstrom, R. & Overbeek, R. fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput Appl Biosci* 10, 41-8 (1994).

20. Newman, J.C. & Weiner, A.M. L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol* 6, R81 (2005).

21. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* 302, 413 (2003).

22. Kothary, R. et al. Inducible expression of an hsp68-lacZ hybrid gene in transgenic mice. *Development* 105, 707-14 (1989).

Figures

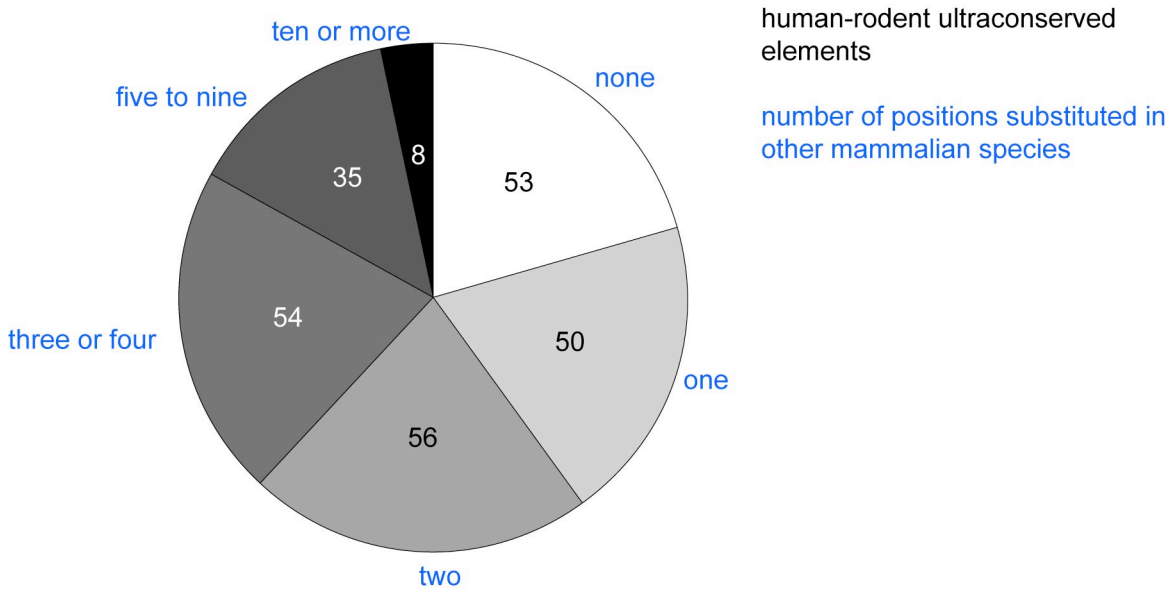


Fig. 1: Most ultraconserved elements are not perfectly conserved in other mammals. Nucleotide substitutions in 256 non-exonic human-rodent ultraconserved elements [1] in five additional placental mammalian genomes were considered (chimpanzee, rhesus, dog, horse, cow). 203 elements (79%) have at least one position substituted in other mammals, 153 (60%) have two or more substituted positions and excessive substitutions at five or more positions were observed in 43 (17%) cases. Additional cases of imperfect sequence conservation due to insertions and deletions were not considered.

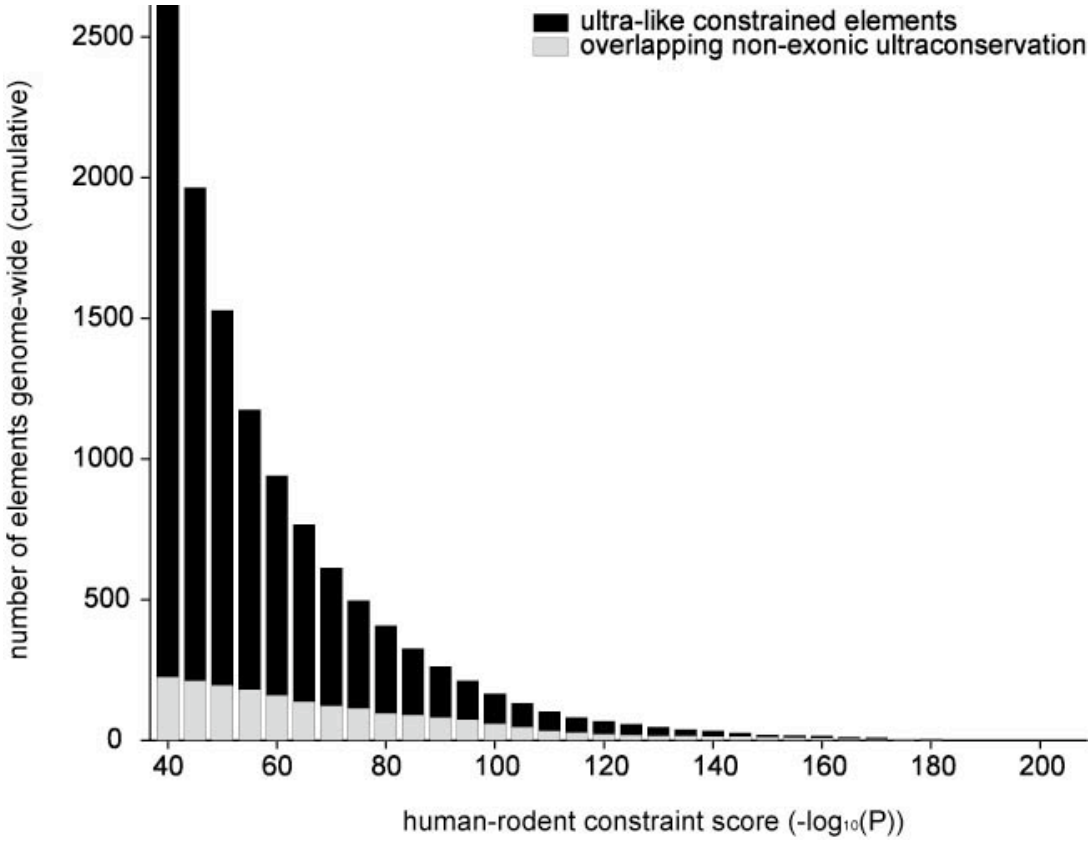


Fig. 2: Ultraconservation identifies a small fraction of elements that are under similar constraint. More than 2,600 extremely human-rodent constrained elements are identified at a constraint score threshold of ≥40, of which more than 2,300 are not defined as ultraconserved. Of the 500 most human-rodent constrained non-coding elements (score ≥74.7), 350 (70%) do not contain or overlap regions of ultraconservation. Overlap with possibly exonic [1] ultraconserved regions is not indicated in the graph.

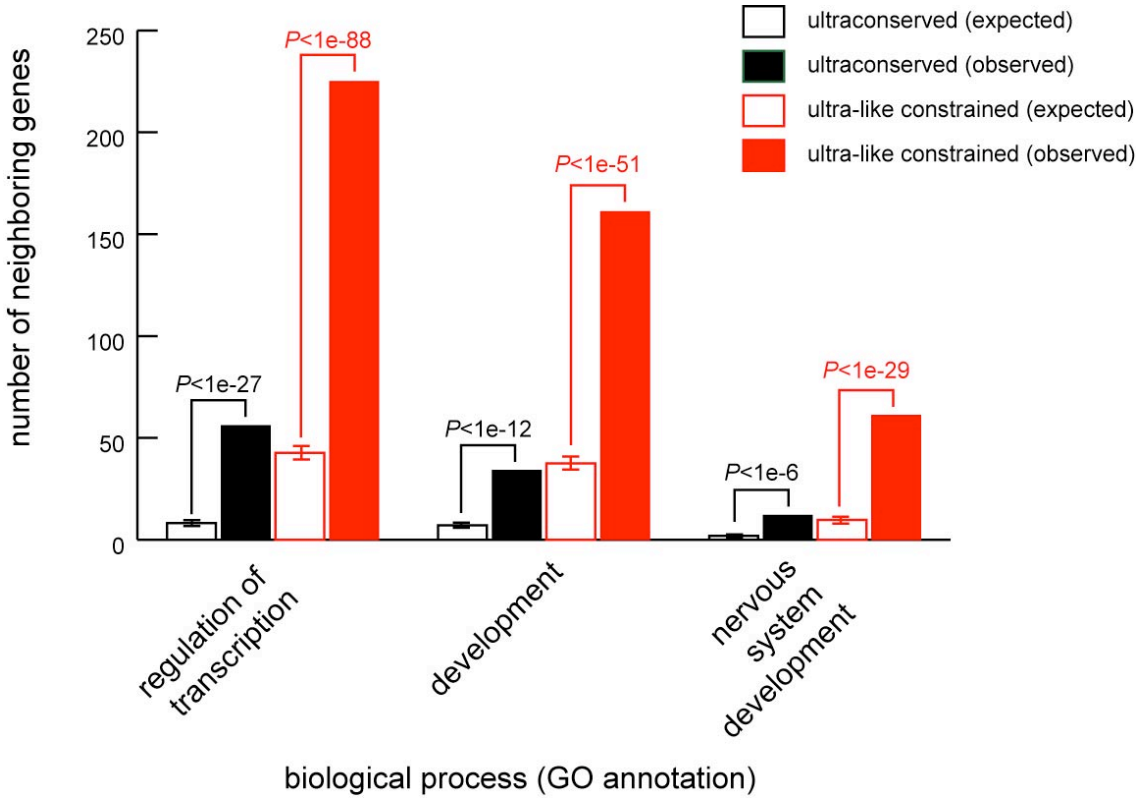


Fig. 3: Enrichment near genes involved in transcriptional regulation, general development and nervous system development. The function (GO, biological process) of the closest neighboring gene of each conserved element was considered. Observed numbers of genes in each category were compared to the number expected based on all annotated RefSeq genes. Additional significantly enriched categories are listed in supplemental table 4. Enrichment P-values are based on the binomial distribution. Error bars indicate ±1 standard deviation.

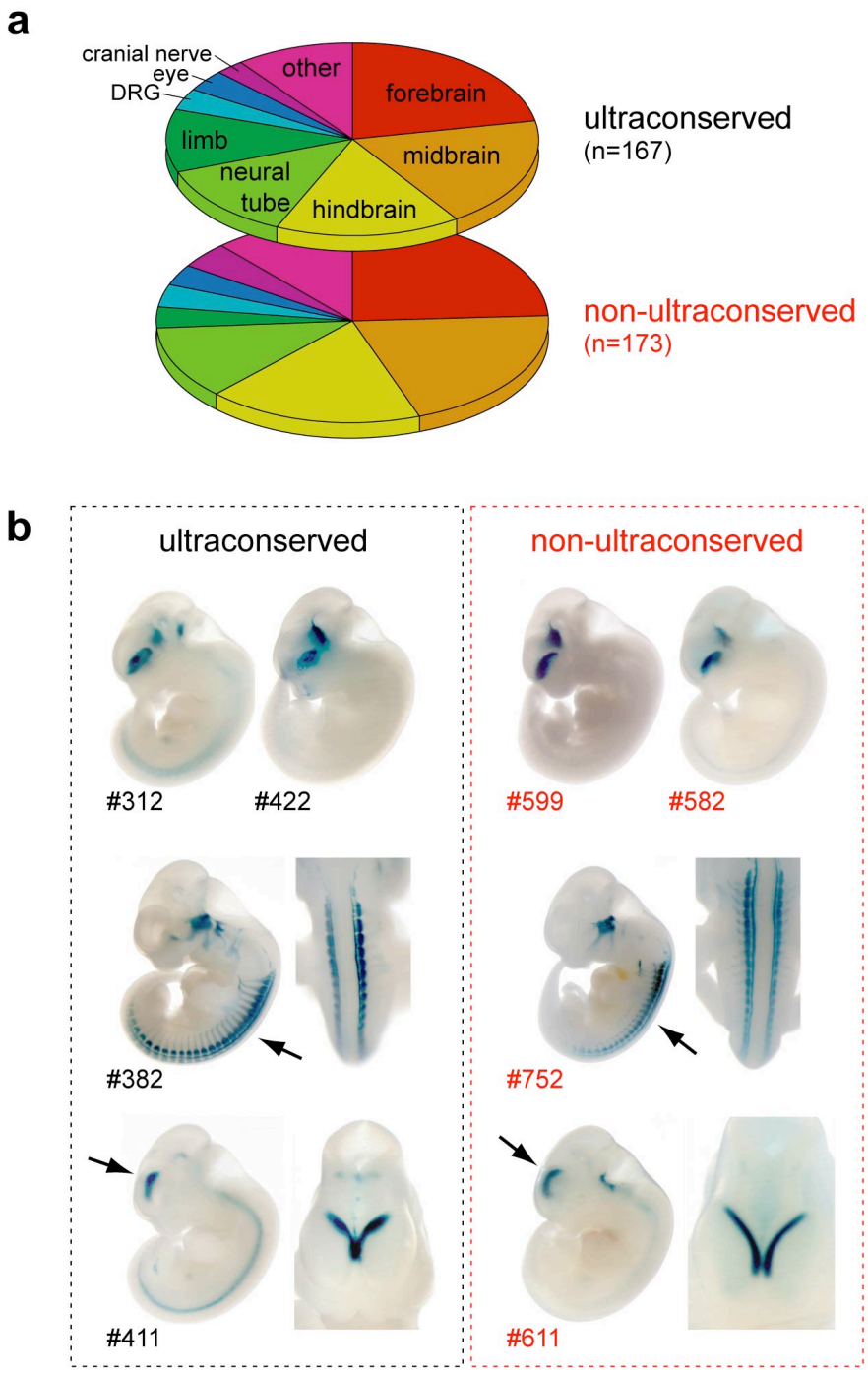
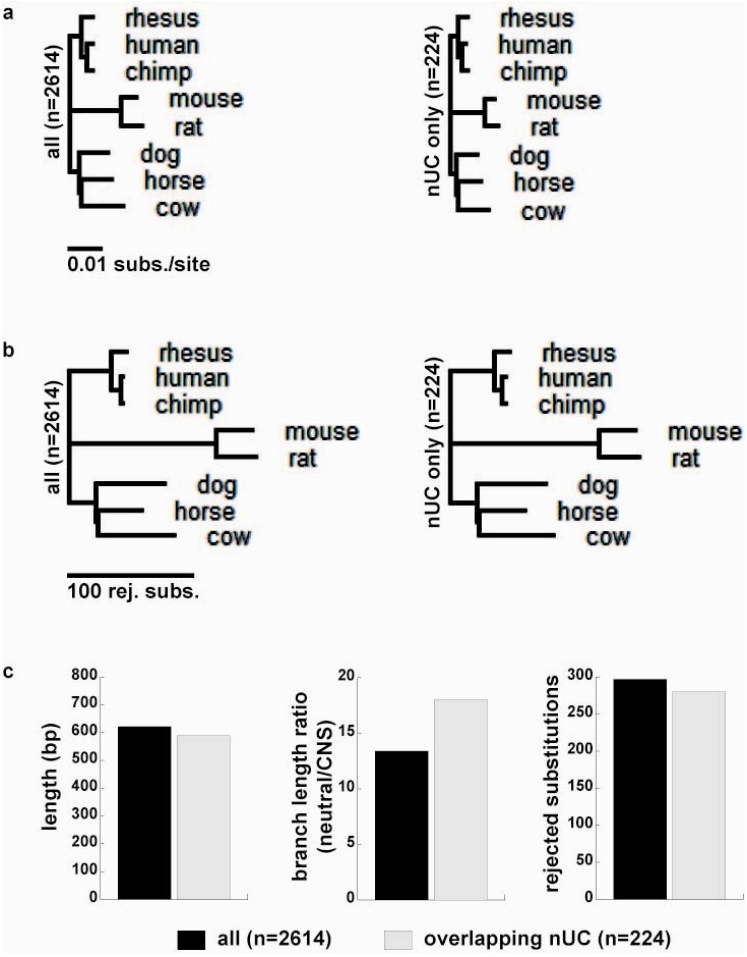


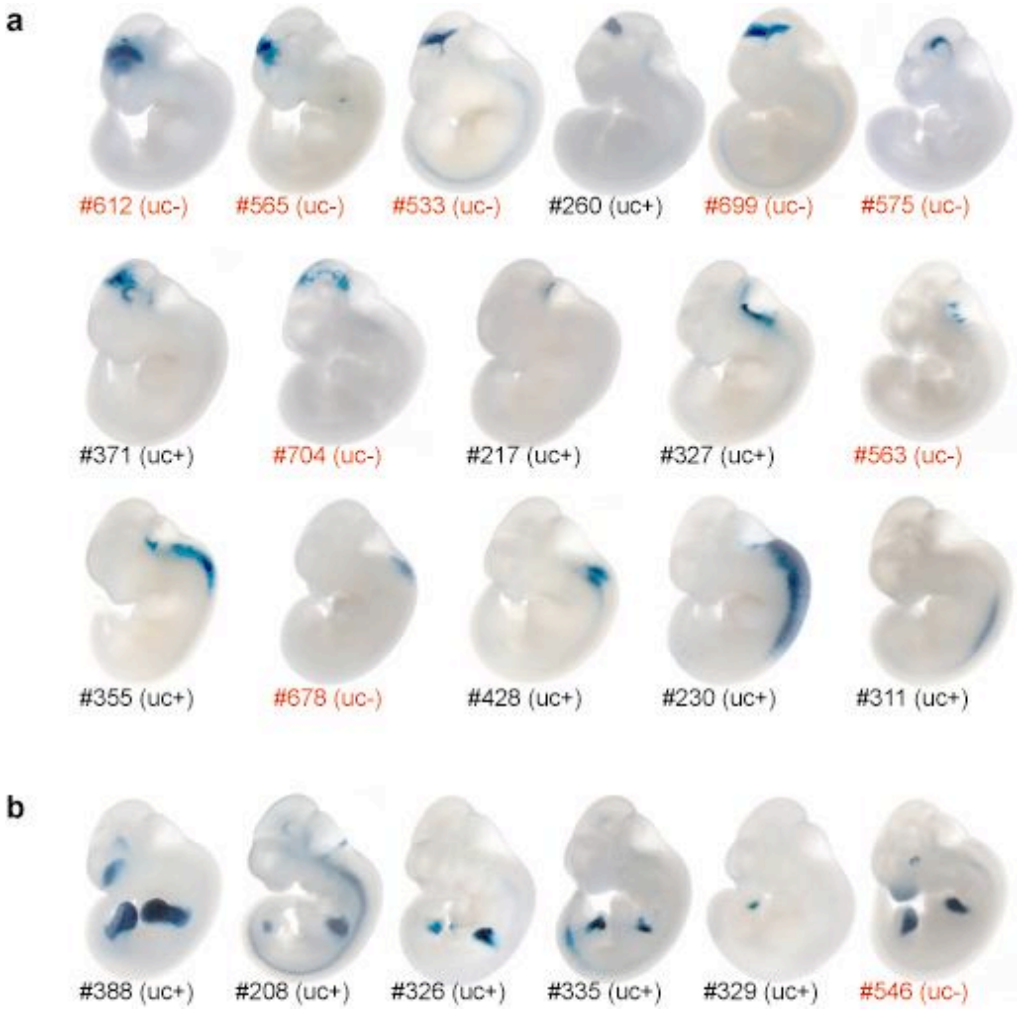
Fig. 4: Highly constrained enhancers target expression to similar tissues independent of ultraconservation. a) Binning of patterns driven by ultraconserved (top)

and ultra-like constrained (bottom) enhancers into broad anatomical domains does not reveal significant differences for any structure (all p-values >0.05, Fisher's exact test with Bonferroni correction for multiple hypothesis testing). Enhancers targeting expression to more than one region are reported in each respective category. b) Examples of extremely constrained enhancers that contain (left) or do not contain (right) regions of ultraconservation, but drive highly similar expression patterns. Arrows indicate viewing angle of insets, only one representative embryo per enhancer is shown; all patterns were reproducible in at least two additional embryos resulting from independent transgenic integration events. DRG, dorsal root ganglia. Genomic coordinates for all enhancers are provided in supplemental tables 1 and 2.

Supplemental Figures



Suppl. Fig. 1: Extreme conservation of ultra-like constrained elements throughout the mammalian clade. a) substitution rate and b) rejected substitutions of 2,614 non-coding elements with ultra-like constraint (left) and the subset that overlaps non-exonic ultraconserved regions (right). c) elements that overlap non-exonic ultraconserved regions are 5% shorter (left), have a 34% higher branch length ratio (center), and 6% less rejected substitutions in placental mammals (rhesus, dog, horse, cow, but excluding human, mouse, rat). All average values were corrected for distribution bias (see methods).



Suppl. Fig. 2: Ultra-like constraint identifies a human-rodent constrained core set of enhancers independent of ultraconservation. Examples of ultra-like constrained enhancers that contain (uc+) or do not contain (uc-) regions of ultraconservation and drive expression in a) subregions of the midbrain, hindbrain and neural tube and b) subdomains of the developing limb. Only one representative transgenic embryo per enhancer is shown; all patterns were reproducible in at least two additional embryos resulting from independent transgenic integration events.