**SUPPLEMENTARY ONLINE MATERIAL**

## Materials and Methods

### Species

Tissues or Genomic DNA from the following species were obtained from different sources (see Table S3): *Sus scrofa, Felis cattus, Myotis myotis, Crocidura russula, Atherurus africanus, Oryctolagus cuniculus, Cercopithecus aethiops, Lemur catta, Dasypus novemcinctus, Loxodonta africana, Macropus eugenii, Ornithorhynchus anatinus.* Sequences for human, mouse were obtained from (*1-5*) and rabbit (for the CNGs only) from (*5*).

### DNA preparation and PCR amplification

Ethanol fixed tissues were sliced, rinsed with distilled water and lyophilised. After rehydration in 80 mM NaCl/ 10mM pH8.0 EDTA, they were digested overnight at 50°C with 0.5 mg/ml Proteinase K in 150mM NaCl/10mM pH8.0 EDTA/10mM Tris pH8.0/1% SDS. Genomic DNA was extracted with phenol/chloroform and precipitated in the presence of Na-acetate and ethanol.

Primers were designed using the human sequence in regions highly conserved between human and mouse. PCR reactions were performed with 100 ng of genomic DNA, 25 ul of JumpStart REDTaq ReadyMix (Sigma) and 0.4 uM of each primer (Sigma-Genosys). The ten first cycles of PCR amplification were performed with a touchdown annealing temperatures decreasing from 60 to 50ºC, while the annealing temperature of the next 30 cycles was 50ºC. Amplimers were separated on 2% agarose gels and positives were directly sequenced on an ABI3100 (ABI). The average PCR product was 289 for CNGs and 400 for ncRNAs.

PCR products with a single amplicon from all other species were sequenced from both strands (>99% of the nucleotides were covered by sequencing from both strands) and the sequences were edited manually and aligned with MultiPipMaker (http://bio.cse.psu.edu/cgi-bin/multipipmaker).

The amplifications were done in pairs of oligonucleotide batches so that each 96-well plate of PCR reactions contained only one type of genomic DNA (species) to control for cross-contamination. A total of 12 platypus and wallaby sequences, that demonstrated high levels of identity with the human, were re-amplified using for platypus a batch of genomic DNA from a different source (see Table S3). They were sequenced two additional times in laboratories different from the one where w

originally performed the experiment. In both cases the sequences obtained were the same as the initial sequence.

**Selection of sequences to be analyzed**

Conserved Non-Genic sequences: The criteria for the selection of the 220 CNGs analysed here is described in (*5*).

Non-coding RNA genes: 16 such genes were selected from the non-coding RNA database (biobases.ibch.poznan.pl/ncRNA/). These genes were chosen as follows: (i) they were at least 100 bps ; (ii) we could identify their reciprocal best BLAST hit with the mouse or human, (iii) they were not anti-sense of any known gene; and (iv) we could design conserved primers between human and mouse. These genes are: AHIF, BIC, BORG, CHH, CMP, DD3, G90, GAS, H19, KCNQ1, NCRM5, NTAB, PCGEM1, S2088, SRA and UHG. Since we could not retrieve any correct sequences from the other 12 species for SRA and KCNQ1, these were excluded from the analysis.

Protein-coding genes: alignment of multiple species sequences of 19 nuclear genes was kindly provided by S. J. O'Brien and W. Murphy. For details and list of genes see (*6*). We partitioned the data in 57 regions of approximately 289 nucleotides. This partition should not influence our results because all analysis is based on averaging across regions. The only metric that could be affected is the clustering, and this will be affected in a conservative way since fusing regions from different genes may generate spurious clustering due to differential levels of substitution rate and amino acid composition.

The sequences obtained in this study are deposited in Genbank under accession numbers CC935641 to CC936712. The sequences for the regulatory regions were obtained from http://globin.cse.psu.edu/ for the beta-globin LCR alignment and from Genbank for the ApoAI, TNFa and GH1 (ApoAI: J04066, M83242, X06659, X64263, Z14124; TNFalpha: L11698, U68414, AF195667, AF011926; GH1: J03071,U02293, U58113, Z46663).

The dog genome sequence:

Sequence data kindly provided by the The Institute for Genomic Research (TIGR), representing approximately 1.5x coverage of the dog genome, was derived from 6.22 million sequencing reads. The end-sequencing of 2 kb and 10 kb clones was

conducted under contract at Celera Genomics as described previously for human (*2*) and reads were assembled with Celera Assembler (*2*). The assembly output consisted of 1.09 million contigs (mean length, 1393 bases) and 0.85 million singletons.

Identification of orthologous dog sequences:

The previously described collections of 2262 CNGs and 1229 CODs from human chromosome 21 (*5*) were searched against the complete collection of assembled dog sequences using BLASTN 2.0MP (http://blast.wustl.edu) with an E-value cutoff of 0.001. The best dog hit for each human sequence was then searched back against the repeat-masked human genome sequence (NCBI build 30, June 2002) using the same cutoff value. Only reciprocal best hits were considered further. For the set of 2262 CNGs, 1638 (73%) had a reciprocal best dog hit (E<0.001), and 1406 (62%) satisfied additional criteria of at least 90% coverage, and at least 70% nucleotide identity. For the set of 1229 CODs, 976 (80%) had a reciprocal best dog hit (E<0.001), and 749 (61%) satisfied additional criteria of at least 90% coverage, and at least 70% nucleotide identity.

**Sequence analysis.**

Note: For the purpose of increasing the sample size, the concatenated COD sequence was split in 57 portions of 289 bps (the average size of CNGs) and the 57 sequences were treated as independent. This does not bias our analysis since we consider the CODs as one group, but it helps account for variability.

Sequence change per million years:

Functionally constrained sequences accumulate substitutions at lower rate than neutrally evolving sequences. Nucleotides of CODs undergo two types of changes, silent (synonymous) that do not alter the amino acid and replacement (non-synonymous) that alter the amino acid. Many ncRNAs form secondary structures and the accumulation of substitutions can be high if that structure is maintained (*7*). Some regulatory regions are highly conserved in distant species (*8-11*).]

We calculated how much of the sequence changes on average per million years. We counted the number of substitutions that have occurred on the tree for each sequence and divided by the size of the sequence and the number of million years each tree spanned in its branches. Note that since different regions had different species coverage, each region had a different denominator of divergence time. The estimates

of branch lengths of the phylogenetic tree used in our analyses are given in Table S4. All the estimations were done with the program PAML3 (*12*). Specifically, ancestral states were reconstructed for the sequences and observed or transient substitutions were placed on each branch of the phylogenetic tree.

Clustering:

Clustering was determined by using a modified method from (*13*). With this method we test whether the variable sites are distributed randomly along the sequence or there are hot-spots or cold-spots of substitutions. Significance levels are determined by randomly permuting the same number of variable sites in the same length of the sequence and calculating the statistic for each of a 1000 permutations to build the null distribution. The modification in this manuscript is the following: instead of determining the largest differential of density of variable sites we determined the sum of differentials for the entire length of the sequence and estimated p-values using the same metric for the permutations. This modification makes the test more sensitive to numerous but small clusters of substitutions.

Number of substitutions per variable site.

This metric is meant to estimate how many times each nucleotide has independently changed on the tree. One particularity is that not all the sequences had the same alignment coverage (due to PCR amplification failure in some species) and therefore we could not consider all the changes of each nucleotide on the complete phylogenetic tree. For this reason we obtained the residuals of the linear regression between the average number of substitutions per variable site and the rate per million years for each sequence. These residuals of this regression were the values used to discriminate between the different classes of sequences.

Strand asymmetry in substitutions:

A→T vs. T→A and C→G vs. G→C substitutions were counted for each sequence on the phylogenetic tree. Only these types of substitutions were considered to avoid the effect of biases that result in G+C content change. The A→T and T→A (and C→G and G→C) substitutions were treated as events with equal prior probability and by assuming a binomial distribution we calculated the probability of the observed data. These probability values were used as the metrics for AT and CG asymmetries.

Divergence:

The values of divergence described in Figure 1 and Table S1 were calculated with MEGA2 (www.megasoftware.net) using the Kimura-2-parameter estimate. This

is a standard metric of divergence that takes into account multiple hits and transtion/transversion bias in substitutions. More sophisticated distance measures, such as Tamura-Nei and REV were also implemented and showed identical patterns (data not shown).

Independence of the above characteristics:
A few of the characteristics above are correlated with each other within some of the sequence classes. However, this correlation was very weak and in almost all cases it explains less than 5% of the variance, and therefore our results are not biased.

## Supplementary information

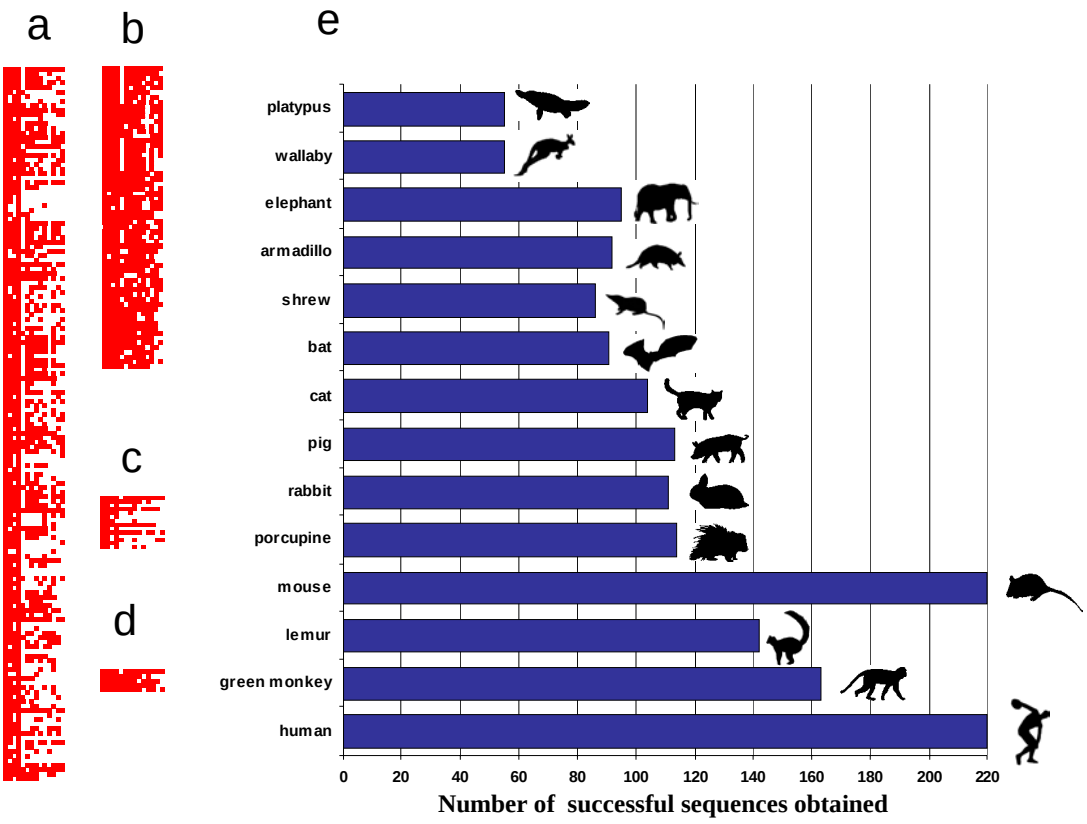
PCR amplification.
The method we used to retrieve the sequences is conservative since it relies on PCR amplification using conserved oligonucleotides. We therefore predict that an even higher proportion of CNGs are conserved in multiple species. Consistently, amplification with a second pair of conserved primers allowed us to retrieve approximately 25% more sequences from a subset of CNGs from species that failed amplification with the first primer pair (data not shown). Moreover, the high level of conservation and the fact that these sequences are single copy in both the human and the mouse genomes suggest that they are orthologous.

Discriminant analysis and regulatory regions.
The three characteristics described in the paper can be analyzed and plotted in a multi-dimensional space. An example of pronounced 3-Dimensional graphical separation is presented in Figure S3 between CNG-high and CODs. This figure shows that CNG-high can be discriminated from CODs in a 3-dimensional space. Interestingly, discriminant analysis shows that the four characteristics can successfully discriminate more than 70% of the three classes of functional sequences (Table S2). In particular, when discriminant analysis was done solely between CNG-high and CODs the correct assignments were 85% and only 9 of the 63 CNG-high were classified as CODs. These results show that CNGs can be discriminated from both classes of transcribed

sequences (CODs and ncRNAs), and this is particularly successful between CNG-high and CODs.

We have hypothesized that some of the CNGs are regulatory regions, and therefore characteristics of known highly conserved regulatory regions should cluster with those of CNGs. To test the validity of this hypothesis we used data from previously identified and well-characterized regulatory regions in the hypersensitive site 2 (HS2) of the beta-globin locus control region (LCR) (see (*14*)for literature review) and the promoters of the TNFα (see (*9*)for literature review), GH1 and ApoAI (http://transfac.gbf.de/TRANSFAC/genes; G000203 and G000282). These regulatory regions show levels of divergence and gap density similar to CNGs when compared between human and mouse, and all of them had been previously sequenced in multiple mammalian species. We calculated the three metrics based on these multiple species alignments and placed their values on the 3-D distributions (see example in Figure S3). In all cases, the well-characterized regulatory regions are grouped with the CNGs. This provides excellent support for the hypothesis that a fraction of the CNGs may function as regulatory elements. In addition, this analysis demonstrates that the metrics we chose to distinguish them from other functional sequences of the genome are appropriate for regulatory regions.

**Figure S1**: Success rate for the retrieval of the orthologous Conserved Non-Genic (CNG) sequences and ncRNAs from 12 mammalian species: a) Alignment coverage of all 191 CNGs (CNGs-all; order of species (from left to right): human, green monkey, lemur, mouse, porcupine, rabbit, pig, cat, bat, shrew, armadillo, elephant, wallaby, platypus) along Hsa21 from centromere (top) to telomere (bottom). Each red square corresponds to a successfully amplified and sequenced CNG (or ncRNA for c and d); b) Alignment coverage of CNG-high (CNGs-high; same order of species as above); c) Alignment coverage of all ncRNAs (ncRNAs-all, same order of species as above) d) ncRNA-high (same order of species as above); e) The number of sequences amplified and sequenced for each of the 12 mammalian species, that corresponded to the orthologous human and mouse CNGs. Mouse and human are at 220 because the sequences were obtained from their available genome.



**<u>Figure S2</u>:** Examples of alignments of conserved sequences: **a)** Alignments of a highly Conserved Non-Genic sequence; **b)** a Conserved Non-Genic sequence with alternating regions of high and low divergence (clustering); **c)** a protein-coding sequence with the characteristic periodicity of variable sites (*6*); **d)** and a non-coding RNA gene with an apparently random pattern of substitutions.

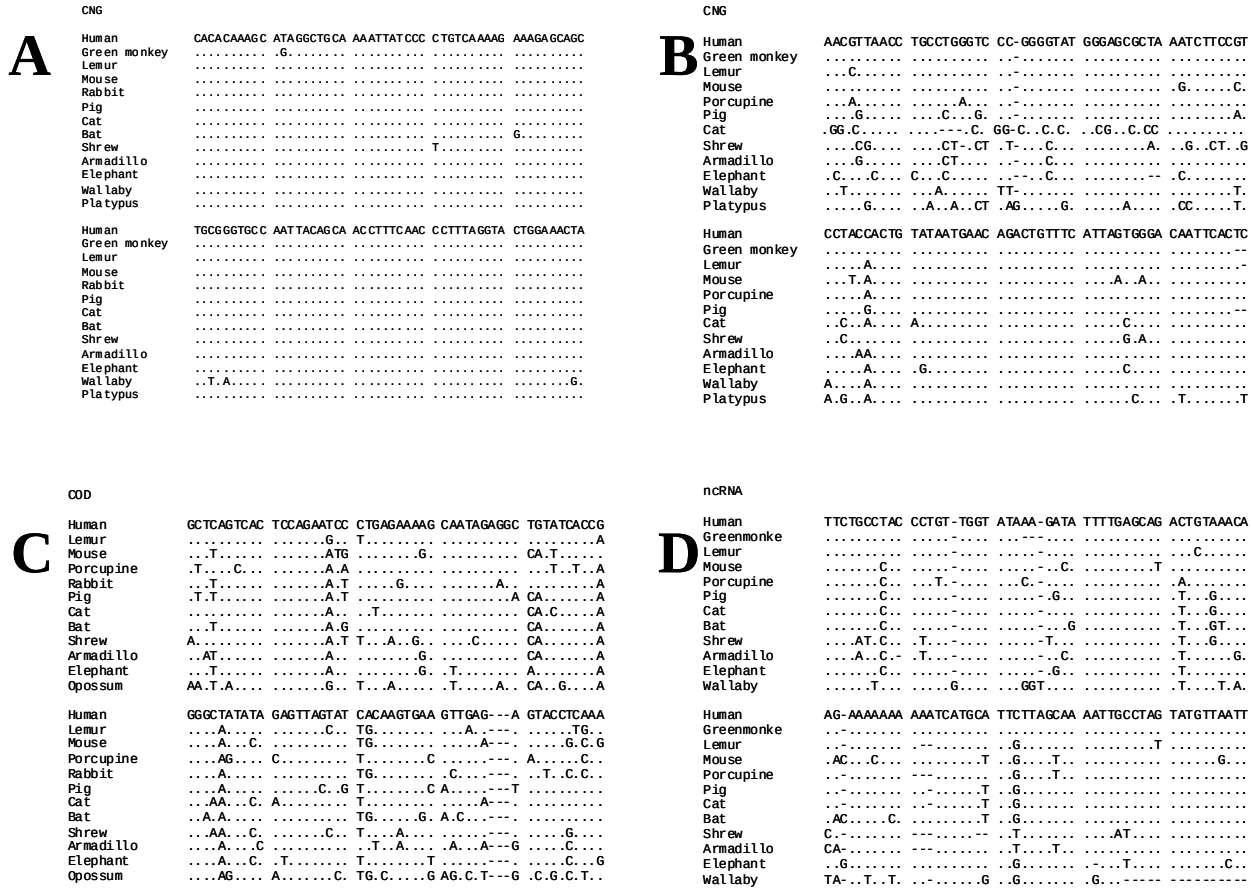


**Figure S3:** An example of three-dimensional representation of residuals of substitutions per variable site (subs/var-site (resid)), AT strand asymmetry and clustering p-values for CNG-high (cyan circles), CNG-low (blue circles), CODs (red circles) and known regulatory regions (REG; green filled squares).

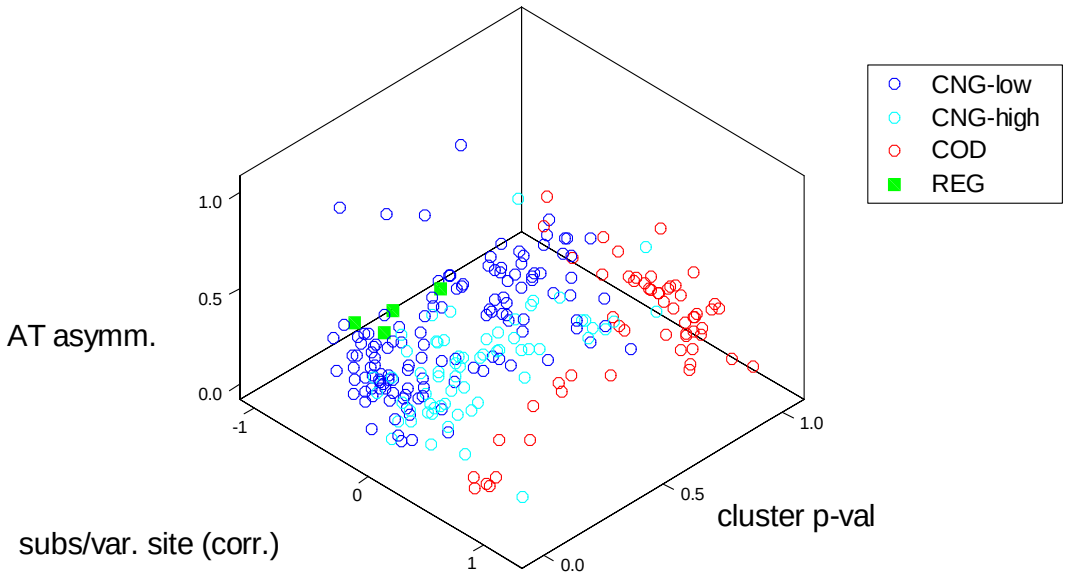
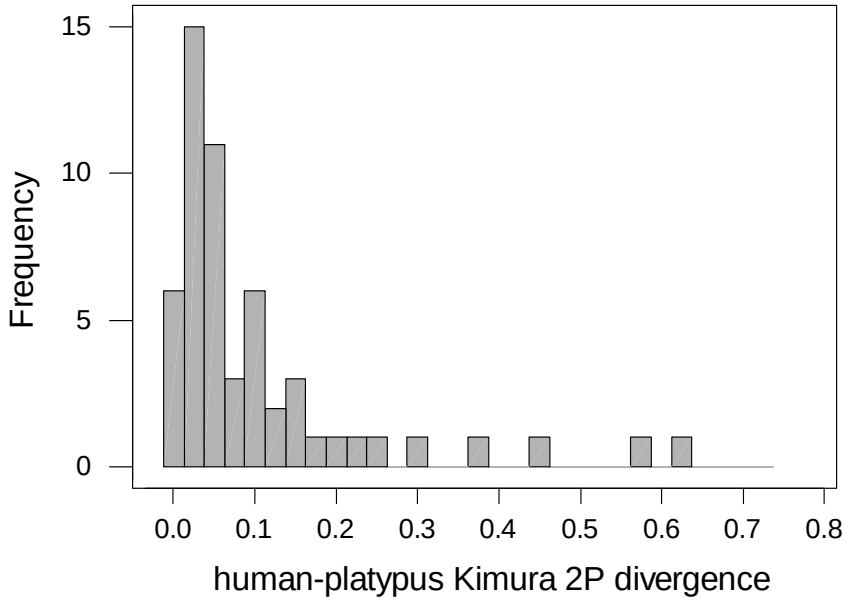


**Figure S4**: Distribution of Kimura 2-parameter divergence values for all human-platypus comparisons.



**TableS1:**Kimura 2-parameter estimate of pairwise divergence for CNGs, CODs and RNAs

| | | human | green monkey | lemur | mouse | porcupine | rabbit | pig | cat | bat | shrew | armadillo | elephant | wallaby / opposum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| green monkey | CNG-all | 0.0279 | | | | | | | | | | | | |
| | COD | *N/A* | | | | | | | | | | | | |
| | ncRNA-all | 0.0502 | | | | | | | | | | | | |
| lemur | CNG-all | 0.0608 | 0.0666 | | | | | | | | | | | |
| | COD | 0.1107 | N/A | | | | | | | | | | | |
| | ncRNA-all | 0.0874 | 0.0988 | | | | | | | | | | | |
| mouse | CNG-all | 0.1499 | 0.1369 | 0.1123 | | | | | | | | | | |
| | COD | 0.1835 | *N/A* | 0.187 | | | | | | | | | | |
| | ncRNA-all | 0.2656 | 0.2828 | 0.1998 | | | | | | | | | | |
| porcupine | CNG-all | 0.0866 | 0.0903 | 0.0816 | 0.1213 | | | | | | | | | |
| | COD | 0.1668 | *N/A* | 0.173 | 0.2046 | | | | | | | | | |
| | ncRNA-all | 0.0996 | 0.1085 | 0.1245 | 0.1895 | | | | | | | | | |
| rabbit | CNG-all | 0.0817 | 0.0864 | 0.0724 | 0.1208 | 0.0999 | | | | | | | | |
| | COD | 0.135 | *N/A* | 0.1396 | 0.1917 | 0.1795 | | | | | | | | |
| | ncRNA-all | 0.1311 | 0.136 | 0.1232 | 0.1942 | 0.1518 | | | | | | | | |
| pig | CNG-all | 0.0861 | 0.0837 | 0.0691 | 0.1357 | 0.0938 | 0.0878 | | | | | | | |
| | COD | 0.1411 | *N/A* | 0.1465 | 0.2043 | 0.1859 | 0.1528 | | | | | | | |
| | ncRNA-all | 0.1391 | 0.1454 | 0.1276 | 0.2127 | 0.145 | 0.1605 | | | | | | | |
| cat | CNG-all | 0.0617 | 0.0666 | 0.0635 | 0.1131 | 0.0795 | 0.0753 | 0.0573 | | | | | | |
| | COD | 0.1245 | *N/A* | 0.1299 | 0.1868 | 0.1741 | 0.1424 | 0.1263 | | | | | | |
| | ncRNA-all | 0.0701 | 0.0912 | 0.1065 | 0.2428 | 0.1305 | 0.1358 | 0.1068 | | | | | | |
| bat | CNG-all | 0.0737 | 0.0888 | 0.0816 | 0.1226 | 0.1287 | 0.0762 | 0.0811 | 0.0707 | | | | | |
| | COD | 0.1202 | *N/A* | 0.1248 | 0.1864 | 0.1727 | 0.1402 | 0.1202 | 0.1089 | | | | | |
| | ncRNA-all | 0.1095 | 0.1059 | 0.1077 | 0.1719 | 0.1306 | 0.1395 | 0.0983 | 0.0853 | | | | | |
| shrew | CNG-all | 0.0811 | 0.0826 | 0.0776 | 0.1281 | 0.1079 | 0.097 | 0.0795 | 0.0741 | 0.0859 | | | | |
| | COD | 0.1685 | *N/A* | 0.1703 | 0.2238 | 0.2146 | 0.1821 | 0.1745 | 0.1581 | 0.1535 | | | | |
| | ncRNA-all | 0.0975 | 0.1503 | 0.1813 | 0.3112 | 0.2356 | 0.2126 | 0.2122 | 0.1399 | 0.2021 | | | | |
| armadillo | CNG-all | 0.0698 | 0.0838 | 0.0799 | 0.114 | 0.0851 | 0.0894 | 0.0709 | 0.0679 | 0.0942 | 0.0892 | | | |
| | COD | 0.1276 | *N/A* | 0.1379 | 0.1997 | 0.1811 | 0.1463 | 0.1422 | 0.1334 | 0.1322 | 0.1706 | | | |
| | ncRNA-all | 0.1365 | 0.1422 | 0.1345 | 0.193 | 0.1471 | 0.1454 | 0.1402 | 0.126 | 0.1168 | 0.2388 | | | |
| elephant | CNG-all | 0.0879 | 0.0902 | 0.0808 | 0.1396 | 0.1116 | 0.0939 | 0.0871 | 0.0801 | 0.0964 | 0.0927 | 0.089 | | |
| | COD | 0.1405 | *N/A* | 0.148 | 0.2094 | 0.1868 | 0.1592 | 0.1556 | 0.1428 | 0.1372 | 0.1888 | 0.1346 | | |
| | ncRNA-all | 0.1423 | 0.1499 | 0.1298 | 0.2114 | 0.1643 | 0.1688 | 0.1446 | 0.1395 | 0.1218 | 0.2341 | 0.1637 | | |
| wallaby/opposum | CNG-all | 0.0633 | 0.0506 | 0.0701 | 0.1236 | 0.0844 | 0.0967 | 0.0857 | 0.0798 | 0.0956 | 0.0661 | 0.0933 | 0.1023 | |
| | COD | 0.2833 | *N/A* | 0.2734 | 0.3063 | 0.301 | 0.287 | 0.2908 | 0.2729 | 0.2741 | 0.3082 | 0.274 | 0.2832 | |
| | ncRNA-all | 0.1682 | 0.174 | 0.2318 | 0.2541 | 0.1762 | 0.2429 | 0.2133 | 0.1858 | 0.1388 | 0.3237 | 0.2659 | 0.207 | |
| platypus | CNG-all | 0.0613 | 0.0536 | 0.0776 | 0.1265 | 0.094 | 0.0926 | 0.0859 | 0.0767 | 0.1031 | 0.0732 | 0.0884 | 0.098 | 0.042 |
| | COD | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* | *N/A* |
| | ncRNA-all | 0.0389 | 0.0708 | 0.1043 | 0.2339 | 0.0798 | 0.1115 | 0.1126 | 0.0563 | 0.1365 | 0.1092 | 0.1245 | 0.1302 | 0.0794 |

**Table S2**: Discriminant analysis of the three classes of functional conserved sequences (CNG-high, CODs, ncRNA-all) and success to discriminate between them in pairwise and three-way comparisons

| Comparison | Input Group | Assigned Group | | | |
|---|---|---|---|---|---|
| | | CNG | COD | ncRNA | Prop. Correct* |
| CNG-high vs. COD | CNG | 54 | 9 | N/A | 85.7% |
| | COD | 9 | 48 | N/A | 84.2% |
| CNG-high vs. ncRNA-all | CNG | 46 | N/A | 17 | 73% |
| | ncRNA | 5 | N/A | 9 | 64.3% |
| COD-high vs. ncRNA-all | COD | N/A | 53 | 4 | 93% |
| | ncRNA | N/A | 1 | 13 | 93% |
| CNG-high vs. COD vs. ncRNA-all | CNG | 40 | 8 | 15 | 63.5% |
| | COD | 8 | 47 | 2 | 82.5% |
| | ncRNA | 4 | 1 | 9 | 64.3% |

* Proportion of correct assignments

**Table S3: Species and source of genomic DNA for the study**

| Taxonomic group | Species | *Species* | Given by/ Origin | Institution | Institution | Sample nb/ strain/ cell line | Sex |
|---|---|---|---|---|---|---|---|
| Eutheria clade IV | Domestic pig | *Sus scrofa* | Seegene Inc. | Seoul | Korea | GDPI 2016-1 | |
| Eutheria clade IV | Domestic cat | *Felis cattus* | M. Casellini | Caroll Veterinary Cabinet | Petit Lancy, Switzerland | | F |
| Eutheria clade IV | Greater mouse-eared bat | *Myotis myotis* | M. Ruedi | Museum of Natural History Geneva | University of Geneva | MHNG 1828.024 | F |
| Eutheria clade IV | White-toothed shrew | *Crocidura russula* | M. Ruedi/ P. Vogel | Institute of Ecology | University of Lausanne | Cru2.6 | F |
| Eutheria clade III | Brush-tailed porcupine | *Atherurus africanus* | M. Ruedi | Museum of Natural History Geneva | University of Geneva | MHNG 1828.022 | |
| Eutheria clade III | House mouse | *Mus musculus* | Celera Genomics | Rockville | Maryland | | |
| Eutheria clade III | Domestic rabbit | *Oryctolagus cuniculus* | Seegene Inc. | Seoul | Korea | GDRB 2017-1, New Zealand White | |
| Eutheria clade III | Human | *Homo sapiens* | Human Genome Sequencing Consortium | Division of Medical Genetics | University of Geneva | | |
| Eutheria clade III | African green monkey | *Cercopithecus aethiops* | A. Reymond | Division of Medical Genetics | University of Geneva | Cos-7 | |
| Eutheria clade III | Ring-tailed lemur | *Lemur catta* | F. M. Catzeflis | Institut des Sciences de l'Evolution | Universite Montpellier 2 | T1251 | |
| Eutheria clade II | Nine-banded armadillo | *Dasypus novemcinctus* | D. L. Dittmann/ F. Sheldon | Louisiana Museum of Natural History | Lousiana State University | LSUMZ M2397 | |
| Eutheria clade I | African elephant | *Loxodonta africana* | C. Wenker | Zoo Basel | Basel | "Yoga", born 1995 in Botswana | M |
| Metatheria | Tammar wallaby | *Macropus eugenii* | J. A. Marshall Graves | Comparative Genomics Research Group | The Australian National University | | F |
| Monotremata | Duck-billed platypus | *Ornithorhynchus anatinus* | J. A. Marshall Graves | Comparative Genomics Research Group | The Australian National University | | F |
| Monotremata | Duck-billed platypus | *Ornithorhynchus anatinus* | M. Westerman | Department of Genetics | La Trobe University | | |

## References


1. E. S. Lander *et al.*, *Nature* **409**, 860-921 (Feb 15, 2001).
2. J. C. Venter *et al.*, *Science* **291**, 1304-51 (Feb 16, 2001).
3. R. H. Waterston *et al.*, *Nature* **420**, 520-62 (Dec 5, 2002).
4. R. J. Mural *et al.*, *Science* **296**, 1661-71 (May 31, 2002).
5. E. T. Dermitzakis *et al.*, *Nature* **420**, 578-82 (Dec 5, 2002).
6. W. J. Murphy *et al.*, *Science* **294**, 2348-51 (Dec 14, 2001).
7. E. Rivas, S. R. Eddy, *BMC Bioinformatics* **2**, 8 (2001).
8. W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, C. E. Lawrence, *Nat Genet* **26**, 225-8 (Oct, 2000).
9. J. Y. Leung *et al.*, *Proc Natl Acad Sci U S A* **97**, 6614-8 (Jun 6, 2000).
10. E. T. Dermitzakis, A. G. Clark, *Mol Biol Evol* **19**, 1114-21 (Jul, 2002).
11. L. Elnitski *et al.*, *Genome Res* **13**, 64-72 (Jan, 2003).
12. Z. Yang, *Comput Appl Biosci* **13**, 555-6 (Oct, 1997).
13. H. Tang, R. C. Lewontin, *Genetics* **153**, 485-95 (Sep, 1999).
14. L. Elnitski, W. Miller, R. Hardison, *J Biol Chem* **272**, 369-78 (Jan 3, 1997).