Dispensability of Mammalian DNA

Cory McLean[1] and Gill Bejerano[1,2*]

[1]Department of Computer Science and [2]Department of Developmental Biology,
Stanford University, Stanford, CA 94305, USA

08/11/08

Contents

*To whom correspondence should be addressed. E-mail: bejerano@stanford.edu

S1 Orthology Assignment and Paralogous CNEs

When multiple rodent sequences match a single human sequence, it can be difficult to computationally determine its true rodent ortholog. The UCSC chaining and netting algorithm attempts to address this challenge by relying on longer range continuity (and a better chain score) at the orthologous locus (Kent et al. 2003). In the second step of our computational pipeline (Supplemental Fig. S1) we discard from further consideration any conserved region for which a unique ortholog cannot be determined this way.

Much interest exists concerning the evolution of genes with close paralogs (Kafri et al. 2005). In the context of conserved non-exonic sequences, our previous work has shown that very few (under 4%) of these sequences can be assigned one or more paralogs based on sequence homology (Bejerano et al. 2004). There are 218,824 distinct non-exonic sequences conserved at 90%id or more between human, macaque, and dog. Only 4,646 of them (2.12%) resemble any other sequence in this set. We discard 7,112 non-exonic sequences conserved at 90%id or more between human, macaque, and dog due to ambiguity. 273 (3.84%) of the 7,112 possess a eutherian paralog, suggesting an ascertainment bias, which however affects only a small fraction of all non-exonic sequence we analyze. Because CNEs are very short compared to genes, and poorly understood, the careful study of paralogy among them lies beyond the scope of this manuscript.

S2 Mouse gene knockouts without an observed phenotype

To assess the prevalence of published mouse gene knockout experiments in which no measurable phenotype was identified, we examine results published in the Mouse Genome Informatics database (Eppig et al. 2007). The database contains 4,234 genes annotated with phenotypes resulting from knockout experiments. 510 genes (12.0%) give rise to no measurable phenotype in at least one knockout experiment. Of these, 284 genes (6.7%) have no measurable phenotype for any knockout experiments. Genes for which no measurable knockout phenotype was identified but have not yet been reported make this estimate a conservative one, and some studies estimate 20% of genes may produce no knockout phenotype (Barbaric et al. 2007). Reported ancient genes for which no measurable knockout phenotype was identified include *Gli1*, *Hoxa7*, and *Dach2*, which are conserved to *Tetraodon nigroviridis* (bony vertebrates), *Xenopus tropicalis* (tetrapods), and *Gallus gallus* (amniotes), respectively.

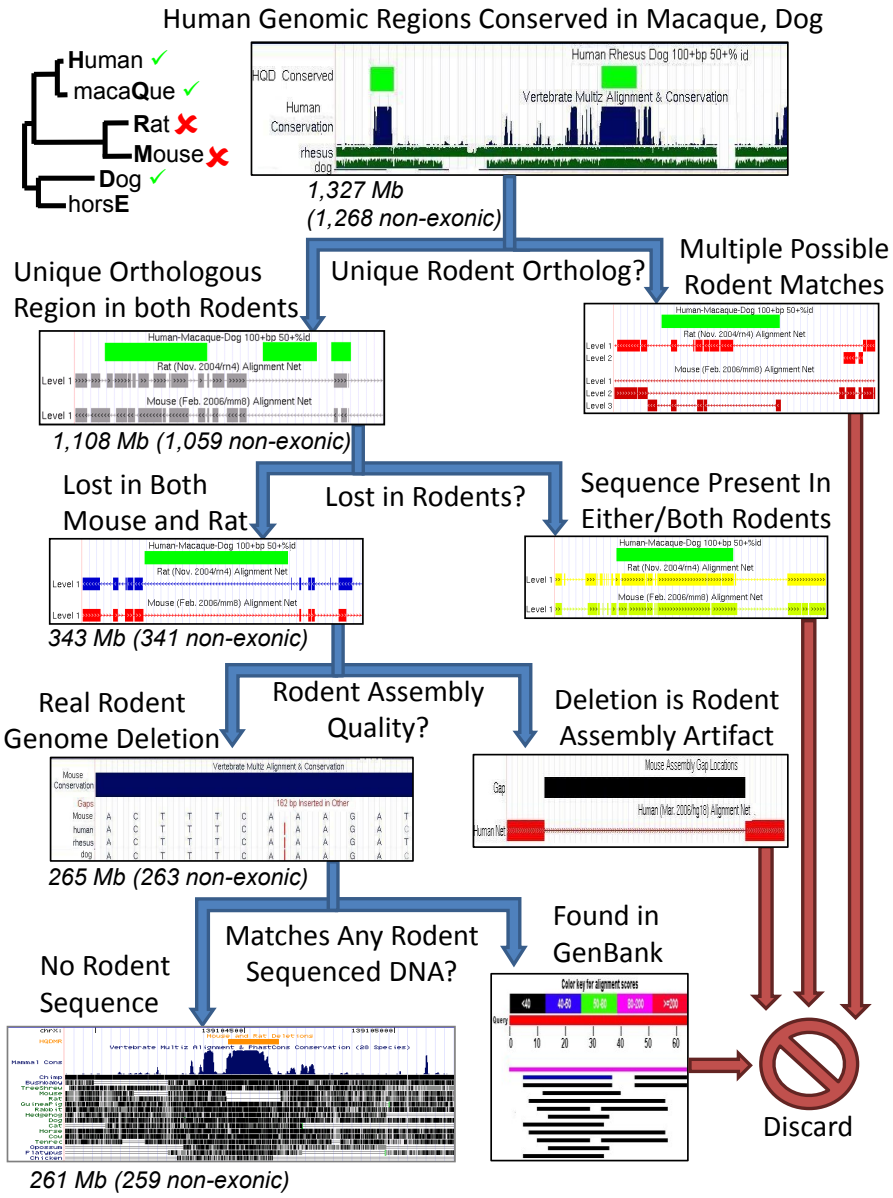

Figure S1: **The computational pipeline used to discover rodent-specific genomic DNA losses.** Total human sequence remaining after each step is displayed below icon. Human-dog-horse and mouse-rat-dog computations were performed the same way.

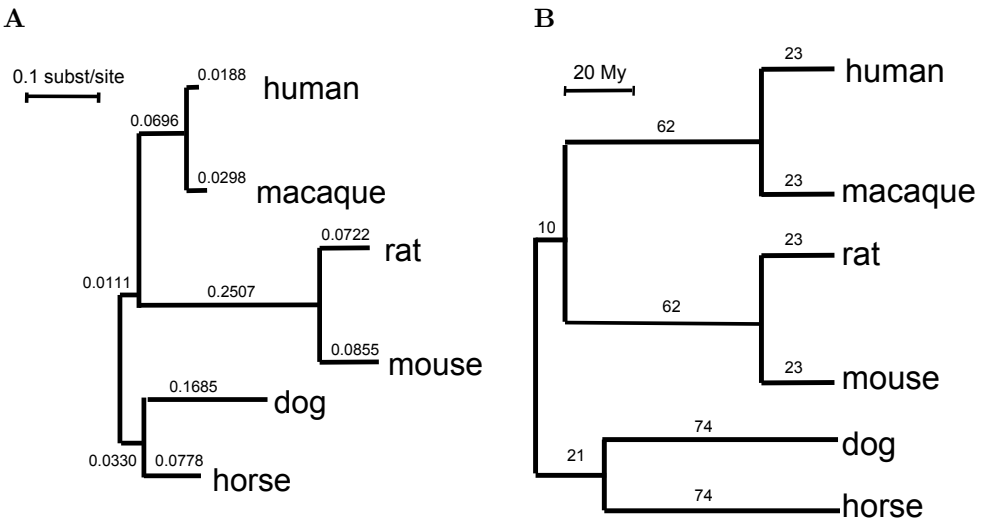

Figure S2: **Phylogenies of species used to calculate pan-mammalian non-exonic loss rates. (A)** Molecular evolution phylogenetic tree adapted from Margulies et al. (2005), showing rapid rodent evolution. **(B)** Evolutionary time phylogenetic tree adapted from Kumar and Hedges (1998), with updated rodent divergence given in International Rat Genome Sequencing Consortium (2004), showing similar branch lengths for the primate and rodent pairs.

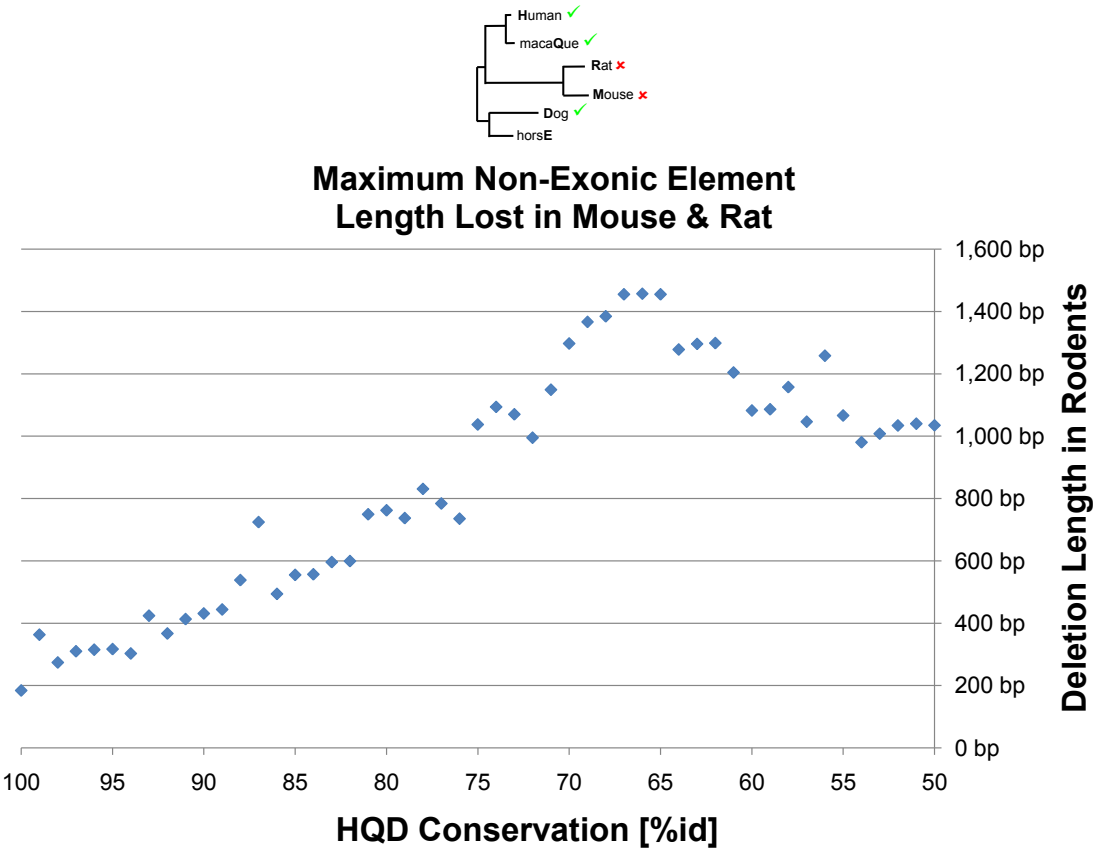

Figure S3: **Maximum length of primate-Dog non-exonic DNA lost in rodents.**

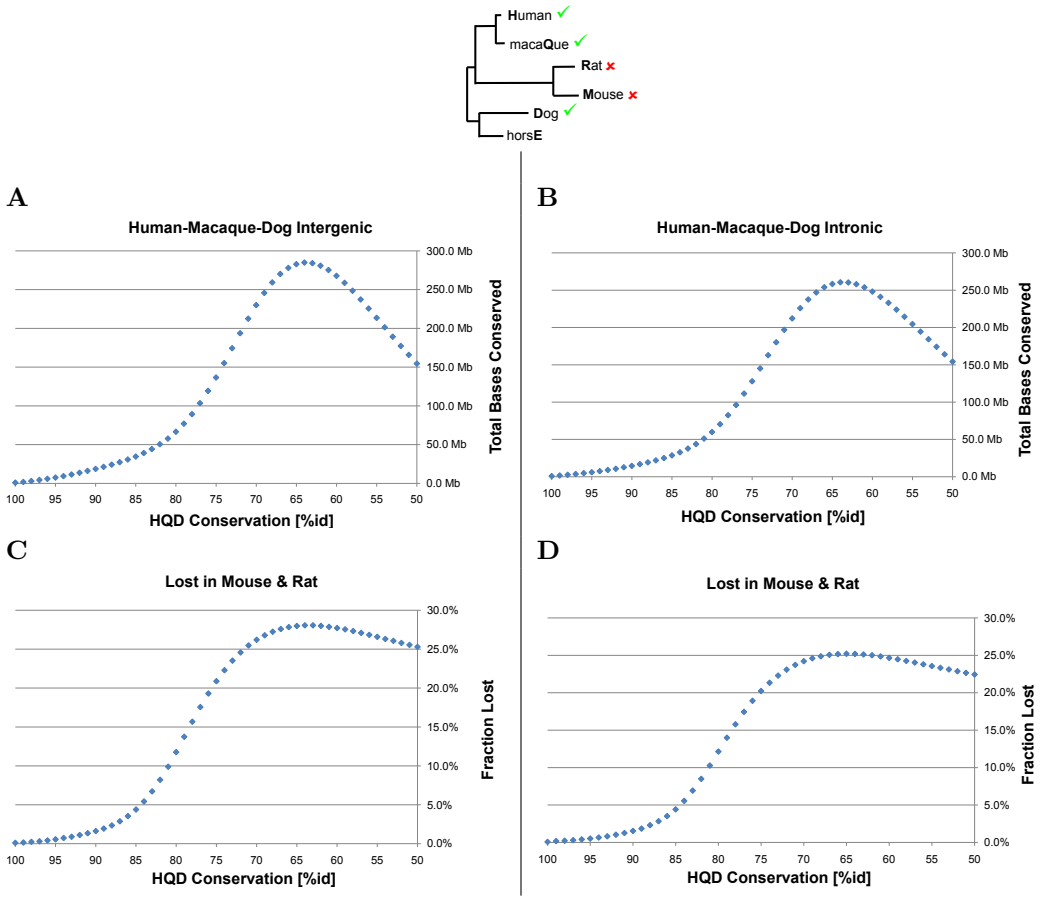

Figure S4: **Abundance and loss rate of intergenic and intronic primate-Dog sequences.** **(A)** Abundance of primate-Dog intergenic DNA. **(B)** Abundance of primate-Dog intronic DNA. **(C)** Fraction of primate-Dog intergenic DNA lost in rodents. **(D)** Fraction of primate-Dog intronic DNA lost in rodents.

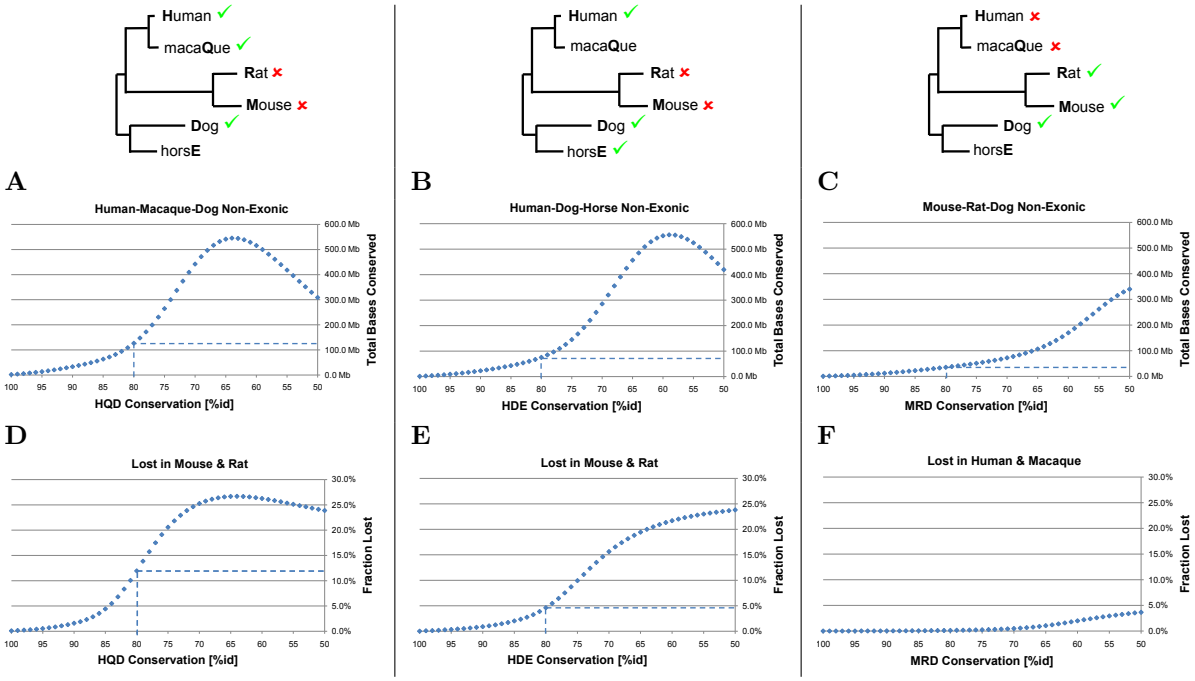

Figure S5: **Abundance and loss rate of non-exonic sequences for all three alignment topologies.** **(A)** Abundance of primate-Dog non-exonic DNA. **(B)** Abundance of Human-Dog-Horse non-exonic DNA. **(C)** Abundance of rodent-Dog non-exonic DNA. **(D)** Fraction of primate-Dog non-exonic DNA lost in rodents. **(E)** Fraction of Human-Dog-Horse non-exonic DNA lost in rodents. **(F)** Fraction of rodent-Dog non-exonic DNA lost in primates. Y-axis values at 80%id are shown as dashed-lines for visualization purposes.

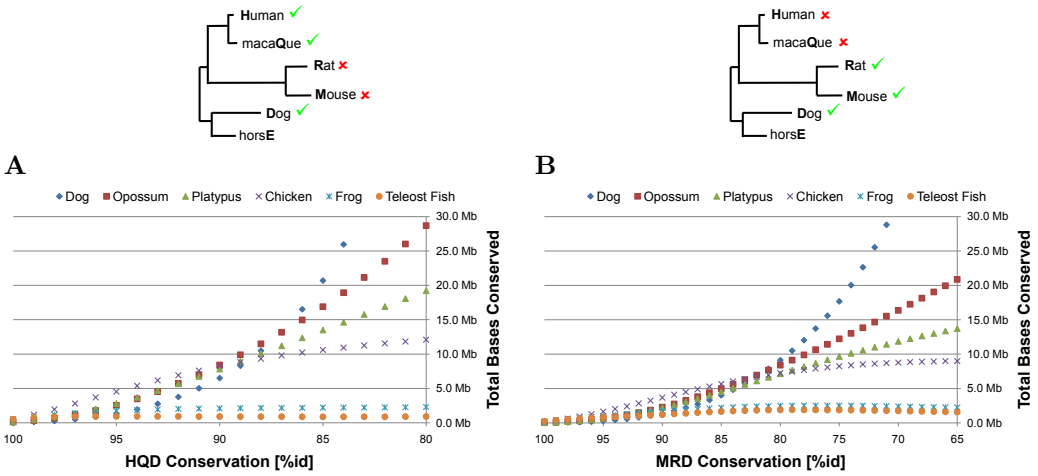

Figure S6: **Deep sequence conservation and abundance. (A)** Abundance of non-exonic primate-Dog bases decomposed by conservation depth. Note x-axis scale and truncated dog conservation depth (up to 66 Mb at 80%id). **(B)** Abundance of non-exonic rodent-Dog bases decomposed by conservation depth. Note x-axis scale and truncated dog conservation depth (up to 61 Mb at 65%id).

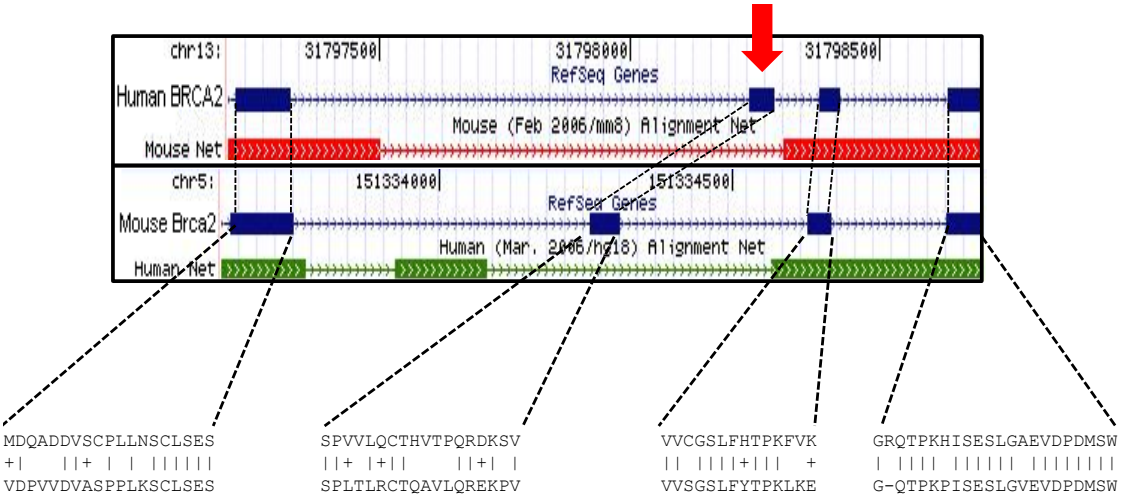

Figure S7: **Mislabeled coding exon loss arising from accelerated evolution.** Red arrow indicates a coding exon of BRCA2 not aligned at the nucleotide level between human and mouse, evident by lack of an overlapping net block (Kent et al. 2003) in both human and mouse. The corresponding amino acid alignment below clearly shows this exon to be orthologous between the two species.

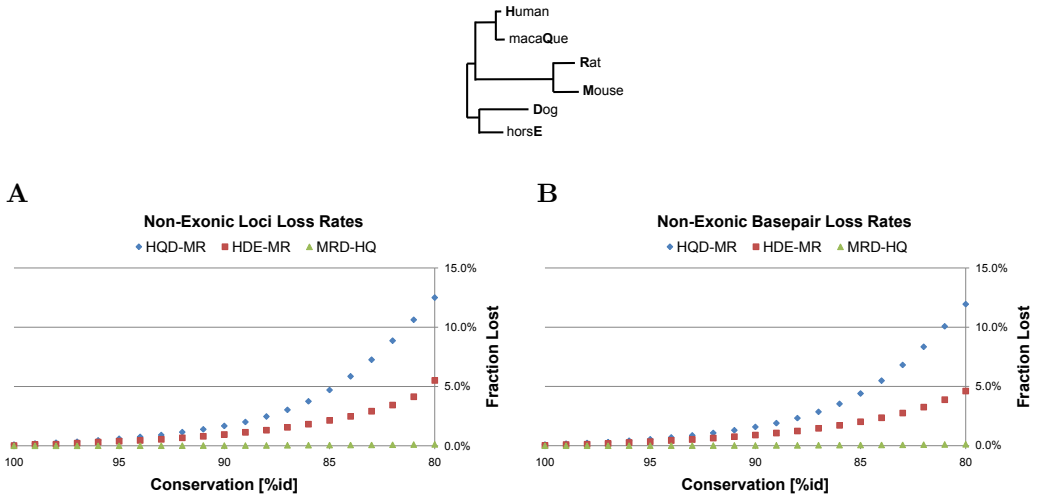

Figure S8: **Cumulative non-exonic loss rates for all three alignment topologies.** **(A)** Fraction of non-exonic loci lost. **(B)** Fraction of non-exonic basepairs lost. Based on the per conservation bin loss rates reported in Figure S5.

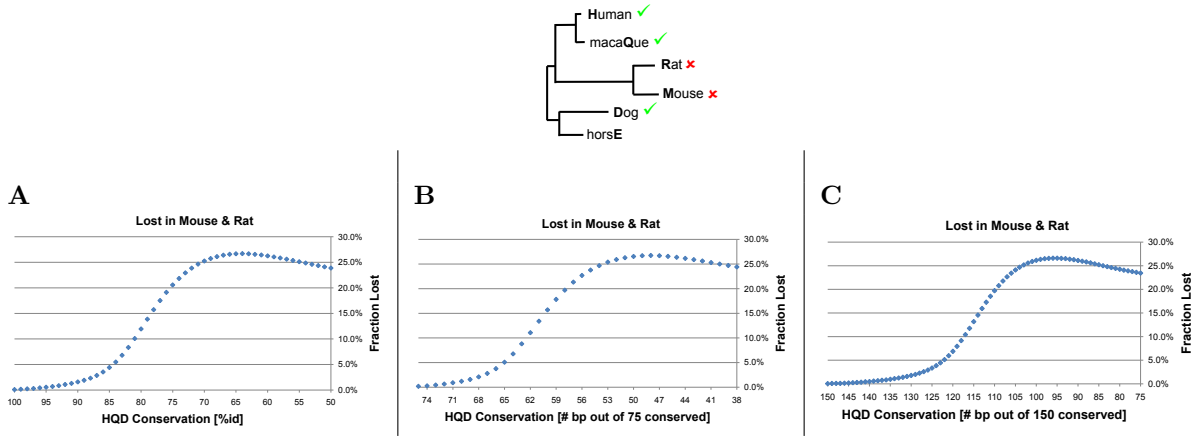

Figure S9: **Loss rate of primate-Dog non-exonic sequences using different window sizes.** **(A)** Fraction of primate-Dog non-exonic DNA lost in rodents using 100 bp windows. **(B)** Fraction of primate-Dog non-exonic DNA lost in rodents using 75 bp windows. **(C)** Fraction of primate-Dog non-exonic DNA lost in rodents using 150 bp windows.

Table S1: **Hypergeometric Gene Ontology Enrichment Test of Rodent-specific Coding Exon Losses of Primate-Dog Alignable Regions**

Function	**Uncorrected P-Value**
polypeptide N-acetylgalactosaminyltransferase activity	0.001199
steroid binding	0.001821
cysteine-type peptidase activity	0.001852
cobalt ion binding	0.002678
regulation of cell motility	0.003485
regulation of locomotion	0.005625
locomotion	0.005837
regulation of cell migration	0.006722
cell motility	0.009093
localization of cell	0.009093
ligand-dependent nuclear receptor activity	0.009842

Showing the most significant of 2,108 GO annotations appearing at least ten times in the background gene set.

Table S2: **Hypergeometric Gene Ontology Enrichment Test of Rodent-specific Non-exonic Losses of $\geq$ 95%id Primate-Dog Regions**

Function	**Uncorrected P-Value**
sequence-specific DNA binding	0.00001538
regulation of transcription	0.0001856
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	0.0001869
transcription factor activity	0.0002324
basal lamina	0.0002545
transcription regulator activity	0.0004294
transcription	0.0004594
regulation of transcription, DNA-dependent	0.0005945
RNA biosynthesis	0.0007776
transcription, DNA-dependent	0.0007776
regulation of cellular physiological process	0.0008805
embryonic development	0.0009508

Showing the most significant of 1,588 GO annotations appearing at least ten times in the background gene set.

Table S3: **Hypergeometric Gene Ontology Enrichment Test of Rodent-specific Non-exonic Losses of $\geq$ 90%id Primate-Dog Regions**

Function	**Uncorrected P-Value**
transcription factor activity	0.00001396
development	0.00001839
system development	0.00002080
sequence-specific DNA binding	0.00002570
anatomical structure development	0.00003076
nervous system development	0.00007730
transcription regulator activity	0.0001035
morphogenesis	0.0001269
organ development	0.0001542
organ morphogenesis	0.0001728
branching morphogenesis of a tube	0.0003129
heparan sulfate proteoglycan metabolism	0.0004138
lung development	0.0004571
cell differentiation	0.0004659
morphogenesis of a branching structure	0.0006124
regulation of transcription	0.0007053
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism	0.0008381

Showing the most significant of 1,859 GO annotations appearing at least ten times in the background gene set.

Table S4: **Alignment based misannotation of orthologous human and mouse coding exons.**

Gene Name	Hg18 Coordinates			Positive Selection Evidence
ACYP1	chr14	74,591,998	74,599,180	
AHRR	chr5	426,185	446,466	Hughes et al. (2005)
BRCA2	chr13	31,797,502	31,798,307	Crespi and Summers (2006)
C13orf16	chr13	110,790,405	110,793,769	
C19orf24	chr19	1,228,300	1,230,106	
C21orf13	chr21	39,700,148	39,700,266	
C6orf57	chr6	71,330,352	71,333,684	
CASP1	chr11	104,408,414	104,409,933	Wang and Gu (2001)
CCDC27	chr1	3,659,702	3,661,434	
CCDC47	chr17	59,178,315	59,183,023	
CCL25	chr19	8,027,414	8,033,639	Immune system response*
DKFZp762E1312	chr2	234,409,533	234,411,309	
DKFZp762E1312	chr2	234,419,906	234,422,794	
FKSG24	chr19	18,167,042	18,168,552	
HTN3	chr4	70,932,875	70,933,531	Sabatini and Azen (1994)
ICOSLG	chr21	44,474,747	44,479,569	Murphy et al. (2006)
IL4	chr5	132,043,372	132,045,491	Vallender and Lahn (2004)
IL4I1	chr19	55,084,643	55,084,800	Chavan et al. (2002)
MKI67IP	chr2	122,202,466	122,204,891	Bronikowski et al. (2004)
MUC13	chr3	126,129,075	126,129,273	
MUM1	chr19	1,311,947	1,315,329	Immune system response*
PLB1	chr2	28,719,370	28,720,788	
PTPN13	chr4	87,920,257	87,921,128	
TCL1B	chr14	95,221,984	95,223,032	
TNFRSF14	chr1	2,476,243	2,480,078	Immune system response*
TPRX1	chr19	52,994,285	53,003,834	Booth and Holland (2007)

*Immune system response genes are frequently under rapid mutation rate (Vallender and Lahn 2004).
Less than 5% of flagged rodent-specific coding exon losses fell into this category.

References

Barbaric, I., Miller, G., and Dear, T. 2007. Appearances can be deceiving: phenotypes of knockout mice. *Brief Funct Genomic Proteomic*, **6**:91–103.

Bejerano, G., Haussler, D., and Blanchette, M. 2004. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics*, **20 Suppl 1**:i40–48.

Booth, H. A. F. and Holland, P. W. H. 2007. Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. *Gene*, **387**(1-2):7–14.

Bronikowski, A. M., Rhodes, J. S., Garland, T. J., Prolla, T. A., Awad, T. A., and Gammie, S. C. 2004. The evolution of gene expression in mouse hippocampus in response to selective breeding for increased locomotor activity. *Evolution Int J Org Evolution*, **58**(9):2079–2086.

Chavan, S. S., Tian, W., Hsueh, K., Jawaheer, D., Gregersen, P. K., and Chu, C. C. 2002. Characterization of the human homolog of the IL-4 induced gene-1 (Fig1). *Biochim Biophys Acta*, **1576**(1-2):70–80.

Crespi, B. J. and Summers, K. 2006. Positive selection in the evolution of cancer. *Biol Rev Camb Philos Soc*, **81**(3):407–424.

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., and Richardson, J. E. 2007. The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res*, **35**(Database issue):630–637.

Hughes, A. L., Packer, B., Welch, R., Bergen, A. W., Chanock, S. J., and Yeager, M. 2005. Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding Loci. *Genetics*, **170**(3):1181–1187.

International Rat Genome Sequencing Consortium 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**(6982):493–521.

Kafri, R., Bar-Even, A., and Pilpel, Y. 2005. Transcription control reprogramming in genetic backup circuits. *Nat. Genet.*, **37**:295–299.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, **100**(20):11484–11489.

Kumar, S. and Hedges, S. B. 1998. A molecular timescale for vertebrate evolution. *Nature*, **392**(6679):917–920.

Margulies, E. H., Maduro, V. V. B., Thomas, P. J., Tomkins, J. P., Amemiya, C. T., Luo, M., and Green, E. D. 2005. Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc Natl Acad Sci U S A*, **102**(9):3354–3359.

Murphy, K. M., Nelson, C. A., and Sedy, J. R. 2006. Balancing co-stimulation and inhibition with BTLA and HVEM. *Nat Rev Immunol*, **6**(9):671–681.

Sabatini, L. M. and Azen, E. A. 1994. Two coding change mutations in the HIS2(2) allele characterize the salivary histatin 3-2 protein variant. *Hum Mutat*, **4**(1):12–19.

Vallender, E. J. and Lahn, B. T. 2004. Positive selection on the human genome. *Hum Mol Genet*, **13 Spec No 2**:245–254.

Wang, Y. and Gu, X. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics*, **158**(3):1311–1320.