Maximum likelihood analysis of mammalian p53 indicates the presence of positively selected sites and higher tumorigenic mutations in purifying sites

Khan, M.M.G.; Ryden, A.M.; Chowdhury, M.S.; Hasan, M.A.; Kazi, Julhash U.

Published in:
Gene

DOI:
10.1016/j.gene.2011.05.011

2011

Maximum likelihood analysis of mammalian p53 indicates the presence of positively selected sites and higher tumorigenic mutations in purifying sites

Maola M. G. Khan[1,7], Anna-Margareta Rydén[2], M. Sanaullah Chowdhury[3], MD. Ashraful Hasan[4] and Julhash U. Kazi[5,6*]

[1]Biochemistry and Molecular Biology, Jahangirnagar University, Bangladesh
[2]Genome Stability Research, National Cancer Center Research Institute, Japan
[3]Department of Computer Science & Engineering, University of Chittagong, Bangladesh
[4]Infectious Diseases Medical Research Center, Hallym University, South Korea
[5]Laboratory of Computational Biochemistry, KN Biomedical Research Institute, Bagura Road, Barisal, Bangladesh
[6]Quality Control Section, Opsonin Pharma Limited, Bagura Road, Barisal, Bangladesh
[7]Chemical Library Validation Team, Chemical Biology Core Facility, Chemical Biology Department, RIKEN Advanced Science Institute, Saitama, Japan.

* Corresponding author:
Julhash U. Kazi
Quality Control Section, Opsonin Pharma Limited
Bagura Road, Barisal-8200, Bangladesh
Tel. +880-431-64074, Fax. +880-431-64075,
E-mail. lcb.kazi@gmail.com

Abbreviations: UMD, Universal Mutation database; PAML, Phylogenetic Analysis by maximum Likelihood; PHYLIP, PHYLogeny Inference Package.

ABSTRACT:

The tumor suppressor gene TP53 (p53) maintains genome stability. Mutation or loss of p53 is found in most cancers. Analysis of evolutionary constrains and p53 mutations reveal important sites for concomitant functional studies. In this study, phylogenetic analyses of the coding sequences of p53 from 26 mammals were carried out by applying a maximum likelihood method. The results display two branches under adaptive evolution in mammals. Moreover, each codon of p53 was analyzed by the PAML method for presence of positively selected sites. PAML identified several statistically significant amino acids that undergo positive selection. The data indicates that amino acids responsible for the core functions of p53 are highly conserved, while positively selected sites are predominantly located in the N-and C-terminus of p53. Further analysis of evolutionary pressure and mutations showed the occurrence of more frequent tumorigenic mutations in purifying sites of p53.

Keywords: p53, Phylogenetic analysis, Maximum likelihood.

INTRODUCTION:

p53 is one of the most notable tumor suppressor genes. It is often regarded as the ‘guardian of the cell’ for conserving genome stability by prevention of mutations (Lane, 1992). It is important in multicellular organisms to activate DNA repair mechanism and arrest the cell cycle at the G1/S phase (Nitta et al., 1997; Keimling and Wiesmuller, 2009). p53 directly participates in the DNA repair process by recognizing DNA lesions and by binding with the proteins that are involved in the excision repair complex (Yoshida and Miki, 2010). However, if the DNA lesion is beyond recovery, p53 initiates apoptosis and thus maintains genomic stability (Li et al., 2008). It also plays critical roles in embryonic development and angiogenesis inhibition (Hofseth et al., 2004).

Mutations in this tumor suppressor gene, p53, are often fatal and cause cancer; mutation or loss of p53 activity has been reported in more than 50% of all human cancer cases (Levine et al., 1991). Divided over tissues, statistical data show that occurrence of p53 mutations is most prevalent in lung (70%), colon (60%), stomach (45%) and breast (20%) cancer (Fenoglio-Preiser et al., 2003; Gasco et al., 2003; Iacopetta, 2003; Toyooka et al., 2003). Inactivation of the p53 gene occurs due to missense and nonsense mutations or insertions/deletions of several nucleotides, which leads to either expression of mutant protein or absence of protein (Benard et al., 2003). Deleterious or deregulated expression of p53 has been attributed to heterogeneously distributed mutations over p53; the mutations are frequent either at highly susceptible DNA regions such as CpG dinucleotides or in codons that encode key residues of the protein (Soussi and Beroud, 2003).

Human p53 consists of 393 amino acids and have at least five domains with specific functions. The transactivation domain (residues 1-42) is localized at the amino terminus and is responsible for the activation of transcription factors. The transactivation domain is followed by a proline rich domain (residues 43-92) that also contains a second transactivation domain. The proline rich domain is important for the apoptotic activity of p53 and is highly conserved. The DNA binding domain (residues 100-300) is the central part of p53 and is regarded as the ‘hot spot’ for mutations as most of the mutations that cause cancer are found in this domain. The oligomerization domain (residues 307-355) is important for tetramerization of p53 and the carboxy-terminus domain (residues 356-393) downregulates the binding of DNA to the central domain and it contains the nuclear localization signals (Harris, 1996; Arrowsmith, 1999; Harms and Chen, 2005).

Variable selection pressure among lineages and the presence of positively selected sites can be tested using different methods. In the present study, we report a phylogenetic analysis of p53 using the maximum likelihood method of PAML (phylogenetic analysis by maximum likelihood) package. PAML utilizes lineage- and site-specific substitution models to detect selection pressure on DNA and protein (Yang and Nielsen, 2002; Yang and Swanson, 2002). The codon substitution models elucidate the mechanism of sequence evolution by comparing synonymous and non-synonymous substitution rates (Goldman and Yang, 1994). The ratio of non-synonymous (d_N) and synonymous (d_S) substitution rate ω (=d_N/d_S) indicates the selection pressure at protein level. A rate of non-synonymous substitution over synonymous substitution higher than 1 (i.e., $\omega > 1$) denotes positive selection. Alongside, $\omega < 1$ and $\omega = 1$ indicate purifying selection and neutral evolution respectively. The lineage-specific model, also called free ratio model, allow the ω ratio to vary among lineages and assume no variation in ω among sites. Thus the lineage-specific model is suitable to detect positive selection along lineages (Yang, 1998; Yang and Nielsen, 1998). On the other hand, the site specific models allow the ω ratio to vary among sites but not among lineages (Yang and Swanson, 2002).

In the present study we conducted a comprehensive phylogenetic analysis of p53 from mammals applying the maximum likelihood method. We intended to delineate the evolutionary pressure among lineages and furthermore present a more detailed study on each codon of p53. To avoid false positives by distant lineages, we focused on mammalian p53 rather than distant members of the p53 superfamily. Previous studies are biased by incorporating p53 sequences from non-mammalian species (Walker et al., 1999; Pintus et al., 2006). Furthermore, the relevance of our study is enhanced by independent statistical analysis of identified mutation sites. Our analysis identified two branches that are under adaptive evolution and several codons under positive selection. Furthermore, analysis of the p53 mutations and evolutionary pressure revealed the presence of mutations in evolutionary preserved or purifying sites. Our results confirm the correlation of evolutionary constrains and occurrence of cancerous mutations in mammalian p53.

METHODS

Assembly and alignment of protein sequences: Amino acids and nucleotide sequences of p53 from 26 mammals were retrieved from publically available databases. The species concerned are: *Bos indicus, Bos Taurus, Canis familiaris, Cavia porcellus, Cercopithecus aethiops, Cricetulus griseus, Delphinapterus leucas, Felis silvestris catus, Homo sapiens, Macaca fascicularis, Macaca fuscata fuscata, Macaca mulatta, Marmota monax, Meriones unguiculatus, Mesocricetus auratus, Microcebus murinus, Mus musculas, Ochotona princeps, Oryctolagus cuniculus, Ovis aries, Pan troglodytes, Pongo pygmaeus, Rattus norvegicus, Sorex araneus, Spalax judaei, Tupaia glis belangeri*. The p53 sequences of *Microcebus murinus, Ochotona princeps, Pongo pygmaeus* and *Sorex araneus* were obtained from the Ensemble genome browser (http://uswest.ensembl.org/index.html). Additional p53 sequences from other species were downloaded from the NCBI database (http://www.ncbi.nlm.nih.gov/). The accession numbers of the entries for p53 are listed in supplementary table S1. Retrieved sequences were aligned by using the program ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/).

Phylogentic analysis of amino acid sequences of p53**:** A phylogenetic tree was prepared based on the maximum-likelihood (ML) method of PHYLIP, PHYLogeny Inference Package, (http://evolution.genetics.washington.edu/phylip.html). Protein sequences were aligned using ClustalW2 with standard settings, and were further analyzed by employing the program 'proml' of PHYLIP. The amino acid substitution matrix JTT (Jones-Taylor-Thornton) was used in PHYLIP and the bootstrap value was set to 1000. A consensus tree was prepared using the application 'Consense'. The ω values (ratio of non-synonymous to synonymous substitutions) were calculated from pairwise ML analysis of all species, as discussed below. The phylogenetic tree was supplemented with ω values by using the graphical program 'Canvas 11'.

Testing for selection pressure: Selection pressure on p53 was analyzed by applying the maximum likelihood method of the PAML program package, version 4 (Yang, 2007). The phylogenetic tree that was previously generated by PHYLIP was used as input data in PAML to analyze the selection pressure. The 'codeml' application of PAML implements both lineage-specific model and site-specific substitution models for detecting codons under positive, purifying and neutral selection. If not mentioned elsewhere in this paper, a site refers to an amino acid or codon rather than a nucleotide. The lineage-specific model (model =1, free ratio) allow for variable ω ratio among lineages, while the site specific models (M0, M1, M2, M7 and M8) allow the ω ratio to vary among sites but not among lineages. To detect positive selection along lineages the control file was set as model = 1, NSsites = 0 and to detect positive selection at specific sites, the control file was set as model = 0, NSsites = 0, 1, 2, 3, 7, 8. Models M0 (one ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (beta) and M8 (beta and ω) were applied. Model M1 (neutral) assumes two classes of sites: the conserved sites (indicating purifying selection) at which ω = 0 and the neutral sites at which ω = 1. Model 3 (discrete) uses three site classes, with the proportions (p_0, p_1, p_2), and the ω ratios (ω_0, ω_1 ω_2). Model M7 (beta) uses a beta distribution. Model M8 (beta and ω) allows for sites with ω>1. The following parameters were used in the control file: Codonfreq = 2, clock = 0, aaDist = 0, model = 0, cleandata = 0, fix_kappa = 0, fix_omega = 0. Codonfreq = 2 specifies the codon substitution model that assumes average nucleotide frequencies at the three nucleotide positions, clock = 0 means no clock and rates are entirely free to vary among branches, aaDist = 0 specifies whether equal amino acid distances are assumed, model = 0 means one ω ratio for all branches, cleandata = 0 indicates keeping of sites with ambiguity data, fix_kappa = 0 and fix-omega = 0 specifies the value to be estimated from the data rather than a fixed value.

Maximum likelihood estimation and likelihood ratio test: Three paired models M1-M2, M0-M3 and M7-M8, were used for the likelihood ratio test (LRT). LRT of M1-M2 and M7-M8 are particularly useful to determine positive selection of nucleotides. However, M1-M2 comparison is more robust than the M7-M8 comparison. The M0-M3 comparison tests the variable pressure among sites, rather than a test of positive selection. The M0-M3 LRT is compared to the χ^2 distribution with 4 degrees of freedom (d.f.), whereas both the M1-M2 and M7-M8 LRTs have 2 d.f. The χ^2 distribution for the appropriate d.f. were obtained by running the 'chi2' program in the PAML package. Sites under positive selection were identified by Bayes' Empirical Bayes' (BEB) posterior probability for models M2 and M8 with significant LRT values.

Analysis of selection pressure and occurrence of p53 mutations in cancer: The UMD-p53 mutation database was a kind gift from Prof. Thierry Soussi (University P.M. Curie, France and Karolinska Institutet, Sweden) through personal communication. This curated database compiles somatic and germline mutations as well as polymorphisms of the p53 in different cancers (Beroud and Soussi, 2003). As selection criteria for amino acids, a mean ω value of ≥ 0.02 or mutations more than 100 were set. We found that 115 amino acids fulfill these criteria and these amino acids were further analyzed to determine the correlation between selection pressure and occurrence of p53 mutations.

RESULTS:

Phylogenetic analysis of mammalian p53: To analyze the phylogeny of p53 we retrieved the amino acids and nucleotide sequences of 26 mammals from public databases. The amino acid sequences of p53 were analyzed by the maximum likelihood (ML) method of PHYLIP. We obtained an unrooted phylogenetic tree, as shown in figure 1, based on the ML analysis of PHYLIP. However, the ML analysis of p53 nucleotide sequences produced almost identical tree topology. The p53 sequences were further analyzed by the PAML to assess the selection pressure on both lineages and codons.

Next, we calculated the rate of non-synonymous and synonymous substitutions of each lineage to determine if they are under positive selection. The rates of synonymous and non-synonymous substitutions were determined by the free ratio model of PAML. The free ration model assumes an independent ω ratio for each branch that is discussed in detail along with other models of PAML in the following sections. Our results identified two lineages under positive selection. *Mesocricetus auratus* (golden hamster) and *Cricetulus griseus* (Chinese hamster) belong to the first lineage with a non-synonymous substitution rate of 37.2 and a synonymous substitution rate of 33.9. The second adaptive lineage have a non-synonymous substitution rate of 13.6 and a synonymous substitution rate of 9.4 and contains 9 species including human, chimpanzee, orangutan, monkey, shrew and lemur. Among these, human and chimpanzee have identical amino acid sequences.

Testing for variable selection pressures and positive selection in mammalian p53: Each codon of p53 across mammals was tested for selection pressure by the site-specific models of PAML. Model M1 assumes two site classes with $\omega_0 = 0$ and $\omega_1 = 1$ and does not allow sites with $\omega > 1$. Model M2 adds a third site class and allow for the presence of positively selected sites. In our experiments M1 and M2 show the same the likelihood value (-10123.9) and thus the likelihood ratio test was not possible for these models. Model M3 that adopts three site classes suggested ~10% sites which are under positive selection with $\omega_2 = 1.017$. M3 identified thirty five amino acids under positive selection; nine of them are significant at $P > 0.99$, eight are significant at $P > 0.95$ and the remaining eighteen are significant at $P > 0.5$ (table 1). Thus this model indicates variable ω ratios among different sites of p53. The sequence alignment of these 35 amino acids is shown in figure 2. A sequence alignment of these amino acids indicates the presence of indels in p53 evolution. Among them 35L, 81T and 383L were positively selected by M2 and M8 along with M3.

Maximum likelihood estimates of M3 suggest that there are three site classes in proportions $p_0 = 0.480$, $p_1 = 0.415$, and $p_2 = 0.103$ with the ratios $\omega_0 = 0.015$, $\omega_1 = 0.292$ and $\omega_2 = 1.017$. These proportions indicate the prior probabilities of any sites that belong to each of the three classes. Prior probabilities are altered after the codons of different species are analyzed and are called posterior probabilities. As shown in figure 3, the posterior probabilities for site 1 are 0.97704, 0.02296, and 0.00000; this indicates that this site is under purifying selection. On the other hand, the posterior probabilities for site 51 are 0.00000, 0.00463, and 0.99537; this indicates that this site is under positive selection. The posterior probabilities for each site of p53 as calculated by M3 are shown in figure 3. It clearly indicates the presence of purifying sites in the DNA-binding domain of p53 and the positively selected sites on the N-terminus and C-terminus of p53.

Model M7 assumes a beta distribution for ω that is limited to the interval (0, 1). Model M8 adds one more site class to M7, with the ω ratio estimated from the data. Model M8 suggests that ~4% sites are under positive selection with $\omega = 1.279$ (table 1).

Likelihood ratio test to detect positive sites: We applied the site-specific likelihood models to detect the positive sites in p53 (Yang and Nielsen, 2002). Three pairs of models were used for this purpose; M1 (neutral) vs. M2 (selection), M0 (one-ratio) vs. M3 (discrete) and M7 (beta) and M8 (beta and ω). The LRT of M0-M3 indicates the variability of ω ratio among sites rather than the exact test of positive selection. However, the LRT of M7-M8 indicates the direct test for the presence of positively selected sites ($\omega > 1$). When we compared M3 with M0, M3 is significantly better than M0 at $P < 0.001$. Moreover, the LTR between M7 and M8 is also significant at $P < 0.05$ (table 2).

Tumorigenic mutations are more frequent in purifying sites of p53: Next we compared the selection pressure on each codon of p53 with the number of mutations in different cancers. The mean ω values were obtained from M3 (discrete model). The frequencies of codon-specific mutations for p53 were taken from the UMD-p53 database (Beroud and Soussi, 2003). For the comparison of mutations and mean ω, we selected the amino acids based on two criteria; a minimum mean ω value of 0.02 or 100 minimum mutations. Thus we identified 115 amino acids out of 393 amino acids of p53. Comparison between mean ω and mutation indicates that mutations are more frequent in the purifying sites of p53, whereas mutations are rare in the sites with high mean ω values (figure 4). Worth noting is that the vast majority of sequences contain only data from the DBD rather than the full coding sequence. Generally, for mutational analysis exons 5-9 are sequenced that cover codons 126-331 which are mostly in the DBD. Therefore, mutations outside the DBD may be underrepresented.

From the M3 model we found 13 sites with ω values more than 1 and for all of these sites the mutation rate is very low. Among these sites only 2 are localized in the DNA binding domain of p53. On the contrary, we found 13 sites with total mutations more than 300. However, for these sites the mean ω value is much lower than 1 (table 3), and all of these sites are in the DNA binding domain. Thus our results suggest that the mutations are more frequent in the sites that are under purifying selection, and that there is an inverse correlation between number of mutations and ω (supplementary fig. S1). The codons 175, 248 and 273 are considered as the 'hot spot' of mutations (Walker et al., 1999) and according to our analysis the total number of mutations for these codons are 1366, 1928 and 1816, respectively, whereas the mean ω values are 0.025, 0.032 and 0.042. However, mutations are rare in the positively selected sites. For instance, codon 383 has a mean ω value of 1.016 with 0 mutations.

DISCUSSION:

In this manuscript we present a specific correlation of evolutionary constraints and tumorigenic mutations of mammalian p53 compared to previous literature (Walker et al., 1999; Pintus et al., 2006; Pintus et al., 2007). We have analyzed a large number of data sets to get a detailed phylogenetic analysis of mammalian p53. Our observations suggest several statistically significant positively selected sites (amino acid), and at least two branches under adaptive evolution (figure 1). Furthermore, we present higher resolution of p53 structure-function and mutational relationship. The results indicate that certain amino acids are maintained under great evolutionary constraints within the DNA binding domain that is critical for the functions of the p53. Interestingly, the most prevalent tumorigenic mutations have been observed in those well conserved sites or sequences.

The non-synonymous to synonymous substation rate (ω) measures selective pressure at the protein level and indicates molecular evolution. Positive selection can be detected by identifying the amino acids where $\omega > 1$. Previous studies on p53 phylogeny have indicated natural and positive selection at both branch and amino acid level. Polymorphism of human p53 is maintained by natural selection (Beckman et al., 1994). Moreover, natural selection of p53 has also been found within and among salmon species (Ford, 2000). Positive selection has also been reported in p53 evolution (Koonin et al., 2005; Rodin and Rodin, 2005). Studies on p53 and p63/p73 families indicated stepwise positive selection in the evolutionary history (Pintus et al., 2006; Pintus et al., 2007).

In our study we restrained the dataset to only mammals to avoid false positives. Compared to previous studies (Walker et al., 1999; Pintus et al., 2006), our dataset is larger yet more specific. Since the dataset originates solely from mammalian sequences, it is also more relevant for conclusions on phylogeny and human p53 mutational hot-spots. Our phylogenetic analysis of mammalian p53 revealed two novel branches under adaptive evolution. Moreover, the sites that we identify in our study are more relevant for evolutionary study of human p53 and mutational analysis pertaining to cancer. Furthermore, the significance of the positively selected sites was confirmed by independent statistical validation using LRT. Our analysis offer definite refinement in p53 analysis, with emphasis on the DBD covering codons 126-331. For example, based on vertebrate p53 sequences Pintus et al. reported 6S and 47P to be under positive selection, albeit at low significance level (Pintus et al., 2006). However, according to our results these amino acids are not under positive selection. Moreover, they identified 11 positively selected sites at 5% significance level, whereas we found 17 sites at the same significance interval.

According to our studies, the rate of transition is much higher than that of transversion in mammalian p53 (table 1). All the mutational hot spots (codon 175, 245, 248, 273, 282) contain CpG dinucleotide and the cytosine of this dinucleotide is often methylated in mammalian cells (Tornaletti and Pfeifer, 1995b; Soussi and Beroud, 2003). Both exogenous carcinogens (for example UV) and endogenous mutagens (such as altered cell metabolism) can target methylated CpG dinucleotide, leading to a high rate of transitions in p53 (Tornaletti and Pfeifer, 1995a; Denissenko et al., 1997). Distribution of mutations in p53 shows various mutational hot spots (figure 4) and the spectra of hot-spots differ among various tumor types. The rate of transition and transversion also varies among different cancer. For example, in colon cancer the rate of transition at CpG dinucleotides is higher, whereas in lung cancer the rate of transversion is higher (Iacopetta, 2003).

The correlation between evolutionary conservation and somatic mutations in cancer of p53 can give insight into other tumor suppressor genes for mutational analysis by targeting the highly conserved regions. At present, the collective information on somatic and germline mutations of other tumor suppressors and oncogenes are not adequate for significant statistical analysis (Soussi et al., 2006). Once the mutation database of other cancer-related genes becomes sufficient, it will be possible to examine the role of tumorigenic mutations in the evolutionary purifying sites to assess the gain or loss of function of the respective genes.

Acknowledgement: We are grateful to Prof. Thierry Soussi (University P.M. Curie, France, and Karolinska Institutet, Sweden) for providing the curated UMD-p53 mutation database. We are also grateful to Stiftelsen Olle Engkvist Byggmästare for financial support.

References:

Arrowsmith, C.H. Structure and function in the p53 family. *Cell Death Differ* **6** (1999), pp. 1169-73.

Beckman, G., Birgander, R., Sjalander, A., Saha, N., Holmberg, P.A., Kivela, A. and Beckman, L. Is p53 polymorphism maintained by natural selection? *Hum Hered* **44** (1994), pp. 266-70.

Benard, J., Douc-Rasy, S. and Ahomadegbe, J.C. TP53 family members and human cancers. *Hum Mutat* **21** (2003), pp. 182-91.

Beroud, C. and Soussi, T. The UMD-p53 database: new mutations and analysis tools. *Hum Mutat* **21** (2003), pp. 176-81.

Denissenko, M.F., Chen, J.X., Tang, M.S. and Pfeifer, G.P. Cytosine methylation determines hot spots of DNA damage in the human P53 gene. *Proc Natl Acad Sci U S A* **94** (1997), pp. 3893-8.

Fenoglio-Preiser, C.M., Wang, J., Stemmermann, G.N. and Noffsinger, A. TP53 and gastric carcinoma: a review. *Hum Mutat* **21** (2003), pp. 258-70.

Ford, M.J. Effects of natural selection on patterns of DNA sequence variation at the transferrin, somatolactin, and p53 genes within and among chinook salmon (Oncorhynchus tshawytscha) populations. *Mol Ecol* **9** (2000), pp. 843-55.

Gasco, M., Yulug, I.G. and Crook, T. TP53 mutations in familial breast cancer: functional aspects. *Hum Mutat* **21** (2003), pp. 301-6.

Goldman, N. and Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11** (1994), pp. 725-36.

Harms, K.L. and Chen, X. The C terminus of p53 family proteins is a cell fate determinant. *Mol Cell Biol* **25** (2005), pp. 2014-30.

Harris, C.C. Structure and function of the p53 tumor suppressor gene: clues for rational cancer therapeutic strategies. *J Natl Cancer Inst* **88** (1996), pp. 1442-55.

Hofseth, L.J., Hussain, S.P. and Harris, C.C. p53: 25 years after its discovery. *Trends Pharmacol Sci* **25** (2004), pp. 177-81.

Iacopetta, B. TP53 mutation in colorectal cancer. *Hum Mutat* **21** (2003), pp. 271-6.

Keimling, M. and Wiesmuller, L. DNA double-strand break repair activities in mammary epithelial cells--influence of endogenous p53 variants. *Carcinogenesis* **30** (2009), pp. 1260-8.

Koonin, E.V., Rogozin, I.B. and Glazko, G.V. p53 gain-of-function: tumor biology and bioinformatics come together. *Cell Cycle* **4** (2005), pp. 686-8.

Lane, D.P. Cancer. p53, guardian of the genome. *Nature* **358** (1992), pp. 15-6.

Levine, A.J., Momand, J. and Finlay, C.A. The p53 tumour suppressor gene. *Nature* **351** (1991), pp. 453-6.

Li, Y.Z., Lu, D.Y., Tan, W.Q., Wang, J.X. and Li, P.F. p53 initiates apoptosis by transcriptionally targeting the antiapoptotic protein ARC. *Mol Cell Biol* **28** (2008), pp. 564-74.

Nitta, M., Okamura, H., Aizawa, S. and Yamaizumi, M. Heat shock induces transient p53-dependent cell cycle arrest at G1/S. *Oncogene* **15** (1997), pp. 561-8.

Pintus, S.S., Fomin, E.S., Ivanisenko, V.A. and Kolchanov, N.A. [Phylogenetic analysis of the family of p53]. *Biofizika* **51** (2006), pp. 640-9.

Pintus, S.S., Fomin, E.S., Oshurkov, I.S. and Ivanisenko, V.A. Phylogenetic analysis of the p53 and p63/p73 gene families. *In Silico Biol* **7** (2007), pp. 319-32.

Rodin, S.N. and Rodin, A.S. Origins and selection of p53 mutations in lung carcinogenesis. *Semin Cancer Biol* **15** (2005), pp. 103-12.

Soussi, T. and Beroud, C. Significance of TP53 mutations in human cancer: a critical analysis of mutations at CpG dinucleotides. *Hum Mutat* **21** (2003), pp. 192-200.

Soussi, T., Ishioka, C., Claustres, M. and Beroud, C. Locus-specific mutation databases: pitfalls and good practice based on the p53 experience. *Nat Rev Cancer* **6** (2006), pp. 83-90.

Tornaletti, S. and Pfeifer, G.P. Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancers. *Oncogene* **10** (1995a), pp. 1493-9.

Tornaletti, S. and Pfeifer, G.P. UV light as a footprinting agent: modulation of UV-induced DNA damage by transcription factors bound at the promoters of three human genes. *J Mol Biol* **249** (1995b), pp. 714-28.

Toyooka, S., Tsuda, T. and Gazdar, A.F. The TP53 gene, tobacco exposure, and lung cancer. *Hum Mutat* **21** (2003), pp. 229-39.

Walker, D.R., Bond, J.P., Tarone, R.E., Harris, C.C., Makalowski, W., Boguski, M.S. and Greenblatt, M.S. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene* **18** (1999), pp. 211-8.

Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15** (1998), pp. 568-73.

Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24** (2007), pp. 1586-91.

Yang, Z. and Nielsen, R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* **46** (1998), pp. 409-18.

Yang, Z. and Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19** (2002), pp. 908-17.

Yang, Z. and Swanson, W.J. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* **19** (2002), pp. 49-57.

Yoshida, K. and Miki, Y. The cell death machinery governed by the p53 tumor suppressor in response to DNA damage. *Cancer Sci* (2010).

Figure Legends:

Figure 1. Phylogeny of p53. The two numbers displayed along each branch correspond to maximum-likelihood estimates of the numbers of non-synonymous and synonymous substitutions along that branch. For each branch, the free ratio model that assumes a different d_N/d_S ratio was used. Two branches that are under positive selection are shown in red color.

Figure 2. Alignment of positively selected sites by M3. Model M3 identified 35 sites under positive selection. Amino acids (9) in red color are significant at $P > 0.99$, amino acids (8) in blue color are significant at $P > 0.95$ and amino acids (18) in black color are significant at $P > 0.5$. Amino acids 35L, 81T and 383L were positively selected by M2, M3 and M8. Numbering of amino acids is based on the sequences of human p53.

Figure 3. Stacked histogram representing the posterior probabilities for the three site classes with different selective pressures identified by the 'codeml' model M3 (discrete). M3 suggests three site classes in proportions $p_0 = 0.480$, $p_1 = 0.415$, and $p_2 = 0.103$ with the ratios $\omega_0 = 0.015$, $\omega_1 = 0.292$ and $\omega_2 = 1.017$.

Figure 4. Analysis of mean ω and number of mutations. The mean ω (calculated using M3 of PAML) is represented below the X axis and the number of mutations (obtained from UMD-p53 mutation database) is represented above the X axis. The amino acid number of human p53 is shown on the top of each bar. (a) Comparison of amino acids 4 to 136, (b) Comparison of amino acids 151 to 237 and (c) Comparison of amino acids 244 to 383.

Table 1: PAML selection analysis

Model	Likelihood	ts/tv	Average dN/dS	Parameter Estimates		Positively selected sites
				Frequency	dN/dS	
M0, one-ratio	-10334.7	3.2	0.202	$p = 1.000$	$\omega = 0.202$	None
M1, neutral	-10123.9	3.47	0.295	$p_0 = 0.783$	$\omega_0 = 0.099$	Not allowed
				$p_1 = 0.216$	$\omega_1 = 1.000$	
M2, selection	-10123.9	3.47	0.295	$p_0 = 0.783$	$\omega_0 = 0.099$	35L, 81T, 383L
				$p_1 = 0.195$	$\omega_1 = 1.000$	
				$p_2 = 0.021$	**$\omega_2 = 1.000$**	
M3, discrete	-10055.7	3.34		$p_0 = 0.480$	$\omega_0 = 0.015$	** 35L, 38Q, 51E, 65R, 81T, 129A, 148D, 311N, 383L * 52Q, 55T, 59G, 67P,
				$p_1 = 0.415$	$\omega_1 = 0.292$	
				$p_2 = 0.103$	**$\omega_2 = 1.017$**	
M7, beta	-10061.9	3.318		$p = 0.326$		Not allowed
				$q = 1.07$		
M8, beta and ω	-10058.9	3.348		$p_0 = 0.962$	**$\omega = 1.279$**	35L, 38Q, 51E, 65R, 71P, 81T, 129A, 148D, 311N, 383L
				$p = 0.382$		
				$p_1 = 0.037$		
				$q = 1.554$		

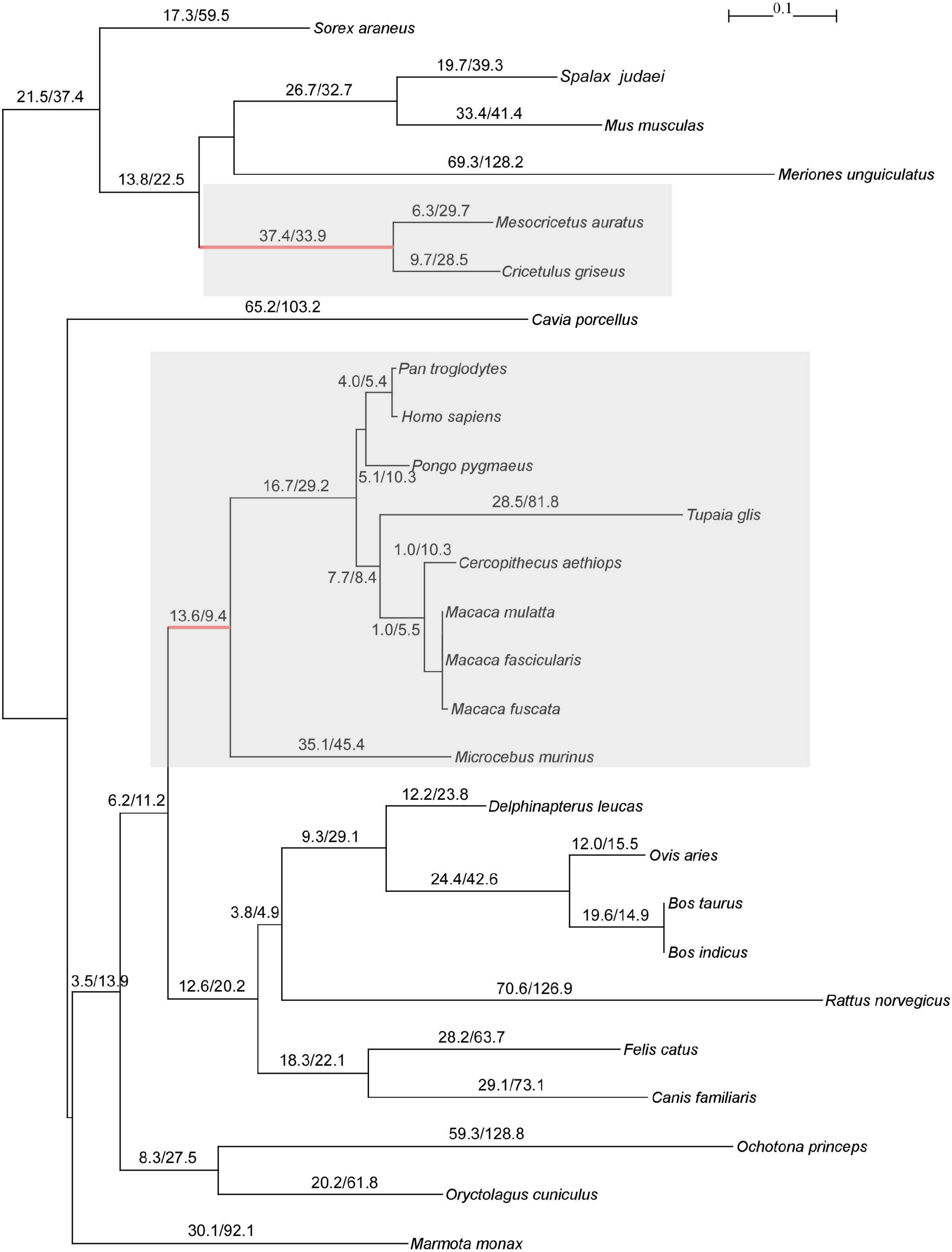
** P > 0.99, * P > 0.95

Table 2: Likelihood ratio statistics (2Δl) for test of positive selection

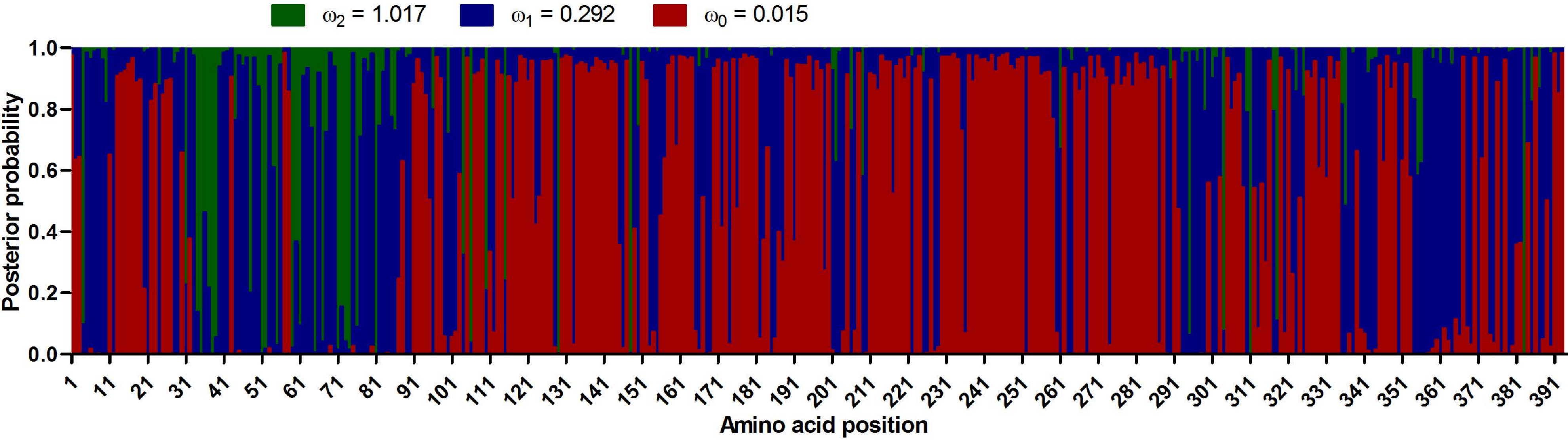
Model	2Δl	Df	P
M2 - M1	0	2	1.0
M0 - M3	558.0	4	< 0.001
M7 - M8	6.0	2	0.049

Table 3: Comparison of Mean ω with frequency of mutations

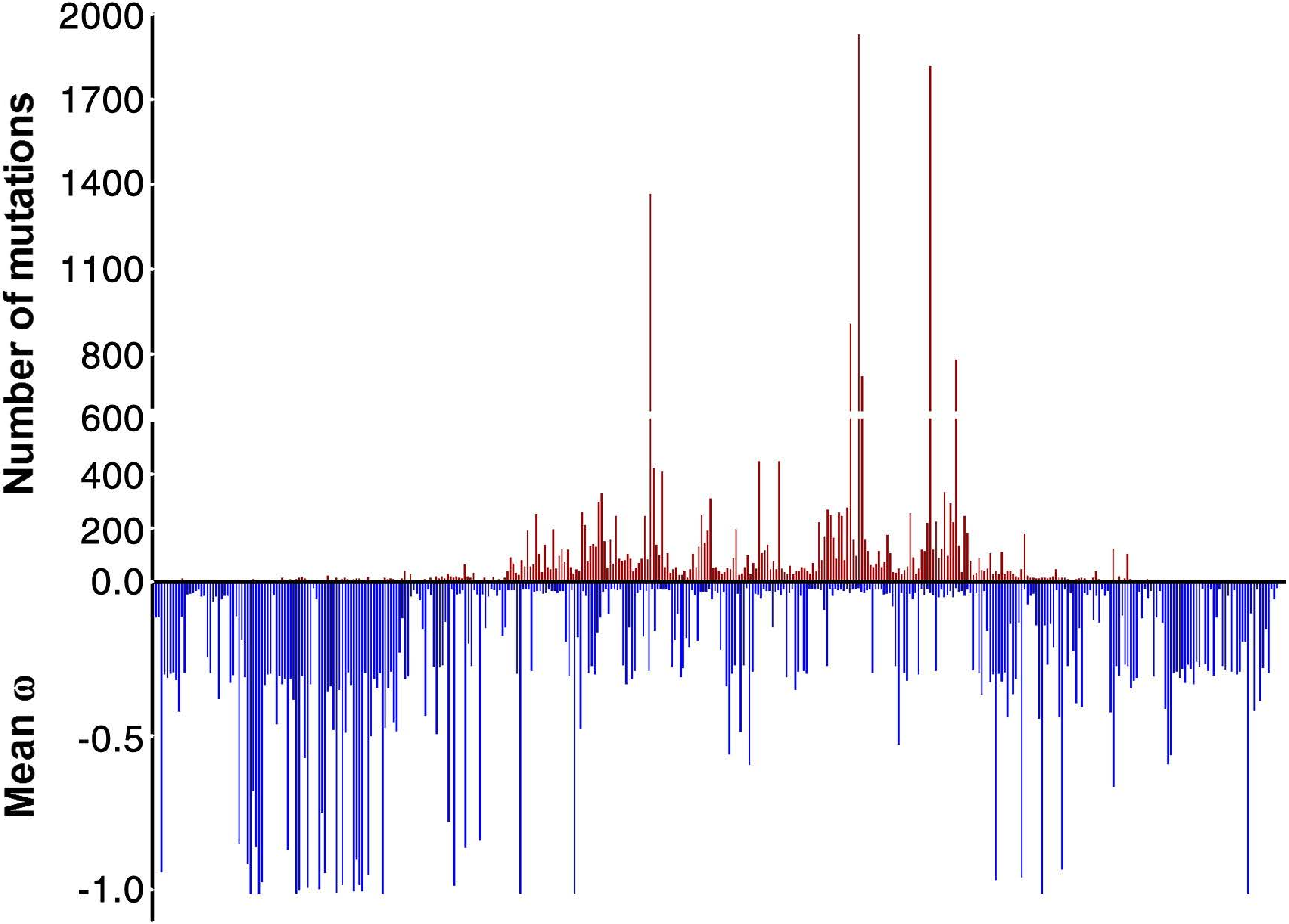
Codon Number	Mean ω	Frequency of mutations		Codon Number	Mean ω	Frequency of mutations
59	1.001	7		196	0.055	311
52	1.004	13		158	0.031	327
71	1.006	9		278	0.042	333
74	1.006	5		179	0.025	408
65	1.012	14		176	0.161	421
148	1.013	31		213	0.054	448
51	1.014	9		220	0.043	449
129	1.015	24		249	0.035	720
311	1.015	12		282	0.031	780
38	1.016	6		245	0.024	908
383	1.016	0		175	0.025	1366
35	1.017	5		273	0.042	1816
81	1.017	5		248	0.032	1928


0.1
17.3/59.5
Sorex araneus
21.5/37.4
19.7/39.3
Spalax judaei
26.7/32.7
33.4/41.4
Mus musculas
69.3/128.2
Meriones unguiculatus
13.8/22.5
6.3/29.7
Mesocricetus auratus
37.4/33.9
9.7/28.5
Cricetulus griseus
65.2/103.2
Cavia porcellus
Pan troglodytes
4.0/5.4
Homo sapiens
Pongo pygmaeus
5.1/10.3
16.7/29.2
28.5/81.8
Tupaia glis
1.0/10.3
Cercopithecus aethiops
7.7/8.4
Macaca mulatta
13.6/9.4
1.0/5.5
Macaca fascicularis
Macaca fuscata
35.1/45.4
Microcebus murinus
6.2/11.2
12.2/23.8
Delphinapterus leucas
9.3/29.1
12.0/15.5
Ovis aries
24.4/42.6
Bos taurus
3.8/4.9
19.6/14.9
Bos indicus
3.5/13.9
12.6/20.2
70.6/126.9
Rattus norvegicus
28.2/63.7
Felis catus
18.3/22.1
29.1/73.1
Canis familiaris
59.3/128.8
Ochotona princeps
8.3/27.5
20.2/61.8
Oryctolagus cuniculus
30.1/92.1
Marmota monax

Species	4	31	34	35	36	37	38	39	48	51	52	55	59	60	61	65	67	71	72	73	74	76	81	104	106	110	115	129	148	295	304	311	318	336	383
Homo sapiens	P	V	P	L	P	S	Q	A	D	E	Q	T	G	P	D	R	P	P	P	V	A	A	T	Q	S	R	H	A	D	P	T	N	P	E	L
Bos indicus	S	L	S	E	L	S	A	P	T	A	T	D	C	P	N	Q	-	-	P	E	P	A	P	P	N	R	Q	S	D	S	T	N	P	K	P
Bos taurus	S	L	S	E	L	S	A	P	T	A	T	D	C	P	N	Q	-	-	P	E	P	A	P	P	N	R	Q	S	D	S	T	N	P	K	P
Cercopithecus aethiops	P	V	P	L	P	S	Q	A	D	A	Q	T	G	P	D	R	S	P	H	M	A	T	T	H	S	R	H	D	D	P	T	N	P	E	F
Canis familiaris	S	V	S	E	L	C	P	A	E	V	N	D	D	S	D	R	-	-	P	A	T	A	G	P	T	R	H	L	S	P	T	S	Q	E	L
Cricetulus griseus	P	V	T	-	S	S	D	S	E	T	G	E	-	S	G	Q	-	T	A	E	D	V	A	Q	D	R	H	S	N	P	A	N	P	E	L
Cavia porcellus	P	V	D	S	L	S	P	P	E	A	S	G	N	P	D	H	S	V	S	E	A	T	A	R	S	E	K	G	E	L	I	S	P	K	L
Delphinapterus leucas	S	L	S	E	L	S	P	A	E	A	N	D	R	P	D	Q	-	-	P	E	P	A	T	P	S	H	H	A	S	S	A	G	Q	E	L
Felis catus	P	V	S	E	L	S	S	A	E	A	N	D	A	P	D	G	-	-	S	A	V	A	A	P	A	H	Q	P	R	P	T	S	Q	E	P
Mesocricetus auratus	P	V	T	-	S	S	D	S	E	A	G	E	-	P	G	Q	S	A	A	E	D	V	A	Q	D	R	H	S	S	P	A	N	P	E	L
Macaca fascicularis	P	V	P	L	P	S	Q	A	D	A	Q	T	G	P	D	R	S	P	P	M	A	T	T	H	S	R	H	D	D	P	T	N	P	E	F
Macaca fuscata	P	V	P	L	P	S	Q	A	D	A	Q	T	G	P	D	R	S	P	P	M	A	T	T	H	S	R	H	D	D	P	T	N	P	E	F
Microcebus murinus	P	V	S	S	L	S	P	P	E	A	G	T	G	P	D	R	S	E	P	A	A	V	T	P	N	R	H	A	D	P	A	N	P	E	L
Marmota monax	A	V	P	V	L	S	P	P	E	E	N	D	G	P	D	Q	S	A	P	K	A	T	-	P	V	R	H	S	D	P	T	G	P	A	I
Macaca mulatta	P	V	P	L	P	S	Q	A	D	A	Q	T	G	P	D	R	S	P	P	M	A	T	T	H	S	R	H	D	D	P	T	N	P	E	F
Mus musculas	S	I	-	-	S	P	H	C	Q	E	E	E	-	P	S	R	S	A	A	Q	D	V	G	Q	N	H	Q	P	S	L	A	C	Q	K	T
Meriones unguiculatus	P	L	-	-	A	L	E	P	Q	T	S	G	G	A	D	P	C	A	E	G	-	A	A	Q	N	R	Q	S	S	R	A	N	S	K	P
Ovis aries	S	L	S	E	L	S	A	P	E	V	T	D	C	P	N	Q	-	-	P	E	P	-	A	P	N	R	H	S	D	S	T	S	Q	K	P
Oryctolagus cuniculus	S	L	T	S	L	N	P	P	E	A	N	N	D	P	E	R	P	A	P	E	A	A	-	H	N	R	H	C	D	P	S	T	T	E	P
Ochotona princeps	Q	L	T	N	M	T	-	-	-	-	N	-	D	P	G	H	T	P	P	E	S	A	P	P	T	R	Q	S	D	P	S	T	-	E	M
Pongo pygmaeus	P	V	P	L	P	S	Q	A	D	A	Q	I	G	P	D	R	S	S	P	V	G	A	I	Q	S	R	H	A	D	P	T	N	P	E	L
Pan troglodytes	P	V	P	L	P	S	Q	A	D	E	Q	T	G	P	D	R	P	P	P	V	A	A	T	Q	S	R	H	A	D	P	T	N	P	E	L
Rattus norvegicus	S	I	T	T	A	T	G	S	L	Q	D	E	-	P	D	E	L	A	A	Q	E	G	A	Q	N	H	Q	S	T	H	A	S	Q	E	P
Sorex araneus	S	N	P	L	Q	Y	P	E	P	E	N	D	S	S	E	R	-	-	P	G	A	A	L	P	S	S	Q	T	T	S	A	G	P	D	P
Spalax judaei	Q	V	T	-	S	P	N	S	E	A	N	D	-	P	D	Q	P	I	T	G	D	V	A	Q	S	R	H	P	D	L	T	G	P	E	L
Tupaia glis	P	V	P	L	P	S	Q	A	D	E	Q	T	G	P	D	R	P	P	P	V	A	A	T	Q	S	R	H	D	D	S	I	G	P	E	L



ω_2 = 1.017
ω_1 = 0.292
ω_0 = 0.015
Posterior probability
1.0
0.8
0.6
0.4
0.2
0.0
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
251
261
271
281
291
301
311
321
331
341
351
361
371
381
391
Amino acid position


Number of mutations
2000
1700
1400
1100
800
600
400
200
0.0
Mean ω
-0.5
-1.0

Supplementary table S1:

Scientific name	Common name	Nucleotide accession number	Protein accession number
Bos indicus	Zebu	U74486	AAB51214.1
Bos taurus	Bovine	NM_174201.2	NP_776626
Canis familiaris	Dog	AF060514	AAC16909
Cavia porcellus	Guinea pig	AJ009673	CAB43196
Cercopithecus aethiops	Afr. green monkey	X16384	CAA34420.1
Cricetulus griseus	Chinese hamster	U50395.1	AAC53040.1
Delphinapterus leucas	Beluga whale	AF475081	AAL83290.1
Felis silvestris catus	Cat	D26608	BAA05653.1
Homo sapiens	Human	NM_000546	NP_000537.3
Macaca fascicularis	Crab eating macaque	AF456343	AAN64027
Macaca fuscata fuscata	Japanese macaque	AF456344	AAN64028.1
Macaca mulatta	Rhesus macaque	L20442	AAA17994
Marmota monax	Woodchuck	AJ001022	CAA04478.1
Meriones unguiculatus	Mongolian jird	AB033632	BAB69969.1
Mesocricetus auratus	Golden hamster	U07182.1	AAB41344.1
Microcebus murinus	Mouse lemur	ENSMICG00000014853	ENSMICP00000013539
Mus musculas	Mouse	NM_011640.2	NP_035770
Ochotona princeps	Pika	ENSOPRG00000006441	ENSOPRP00000005906
Oryctolagus cuniculus	Rabbit	X90592	CAA62216.1
Ovis aries	Sheep	X81705	CAA57349.1
Pan troglodytes	Chimpanzee	XM_001172099.1	XP_001172099.1
Pongo pygmaeus	Orangutan	ENSPPYG00000007934	ENSPPYP00000008923
Rattus norvegicus	Rat	X13058	CAA31457.1
Sorex araneus	Shrew	ENSSARG00000006206	ENSSARP00000005616
Spalax judaei	Blind subterranean mole rat	AJ783406	CAH03844
Tupaia glis belangeri	Common tree shrew	AF175893	AAF22640.1