Supplementary Material for "Non-linear dynamics of non-synonymous (d_N) and synonymous (d_S) substitution rates affects inference of selection"

Contents

1 Supplementary figures

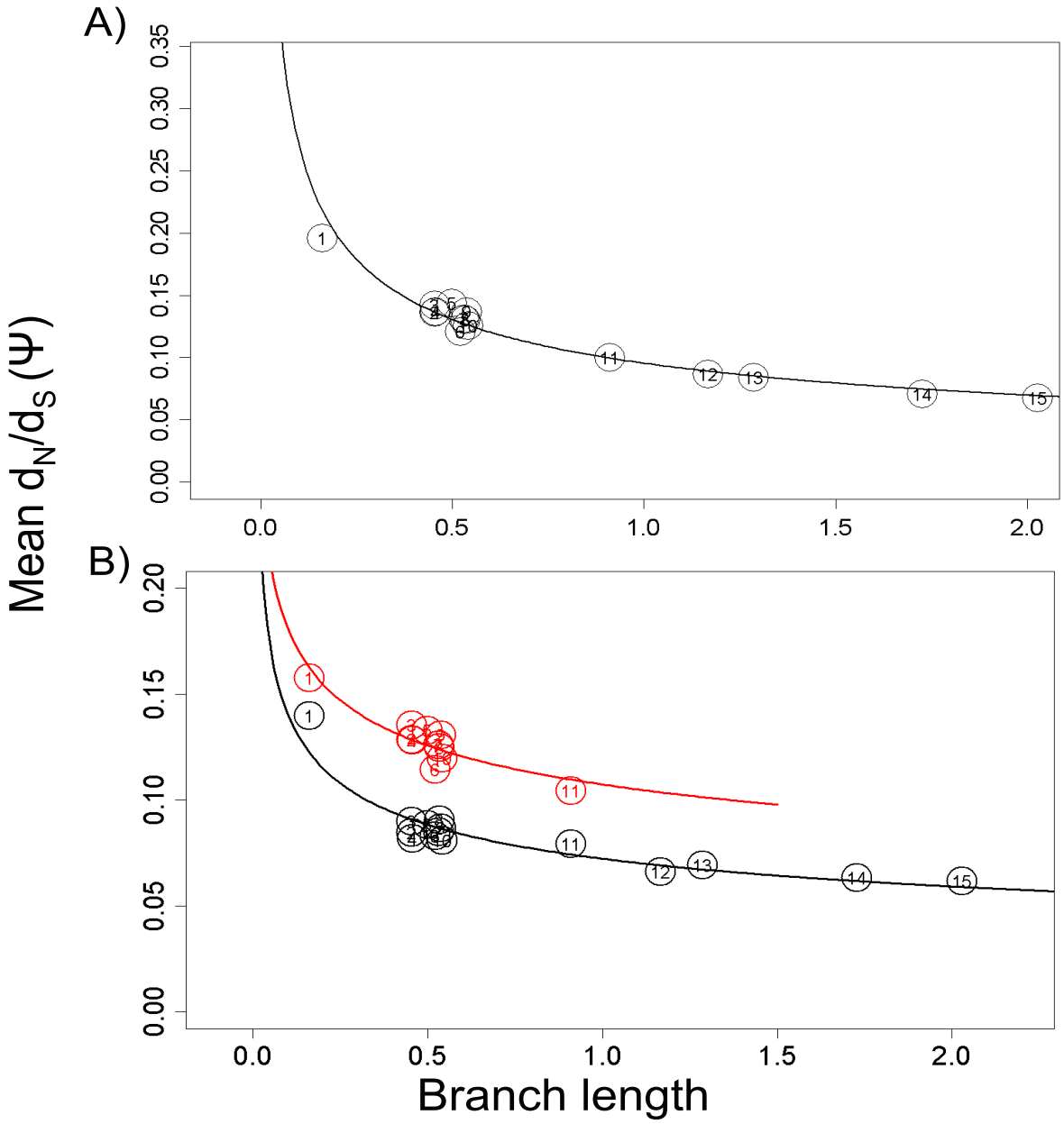

Figure S1: Relationship of ψ and evolutionary distance (branch length) shown for pair wise alignments of mouse and 15 other species. A) All possible orthologues between two species are included. B) Restriction to core sets of genes that are common to all species pairs. Red: 11-way core set of 4572 ortholgoues genes retrieved from all possible pair wise comparisons from mouse-rat to mouse-opossum. Black: 15-way core set of 113 genes common to all possible pair wise comparisons from mouse-rat to mouse-zebra finch. The fitted lines are based on log-log regression models (all 2-way orthologues: $p < 0.001$, $R^2_{adj} = 0.97$, core set 1: $p < 0.001$, $R^2_{adj} = 0.79$, core set 2: $p < 0.001$, $R^2_{adj} = 0.89$). Number code: 1: rat; 2: human; 3: rhesus macaque; 4: chimp; 5: bushbaby; 6: mouse lemur; 7: rabbit; 8: dog; 9: elephant; 10: cow; 11: opossum; 12: platypus; 13: chicken; 14: Xenopus; 15: zebra fish. Branch length estimates from Miller *et al.* (2007).

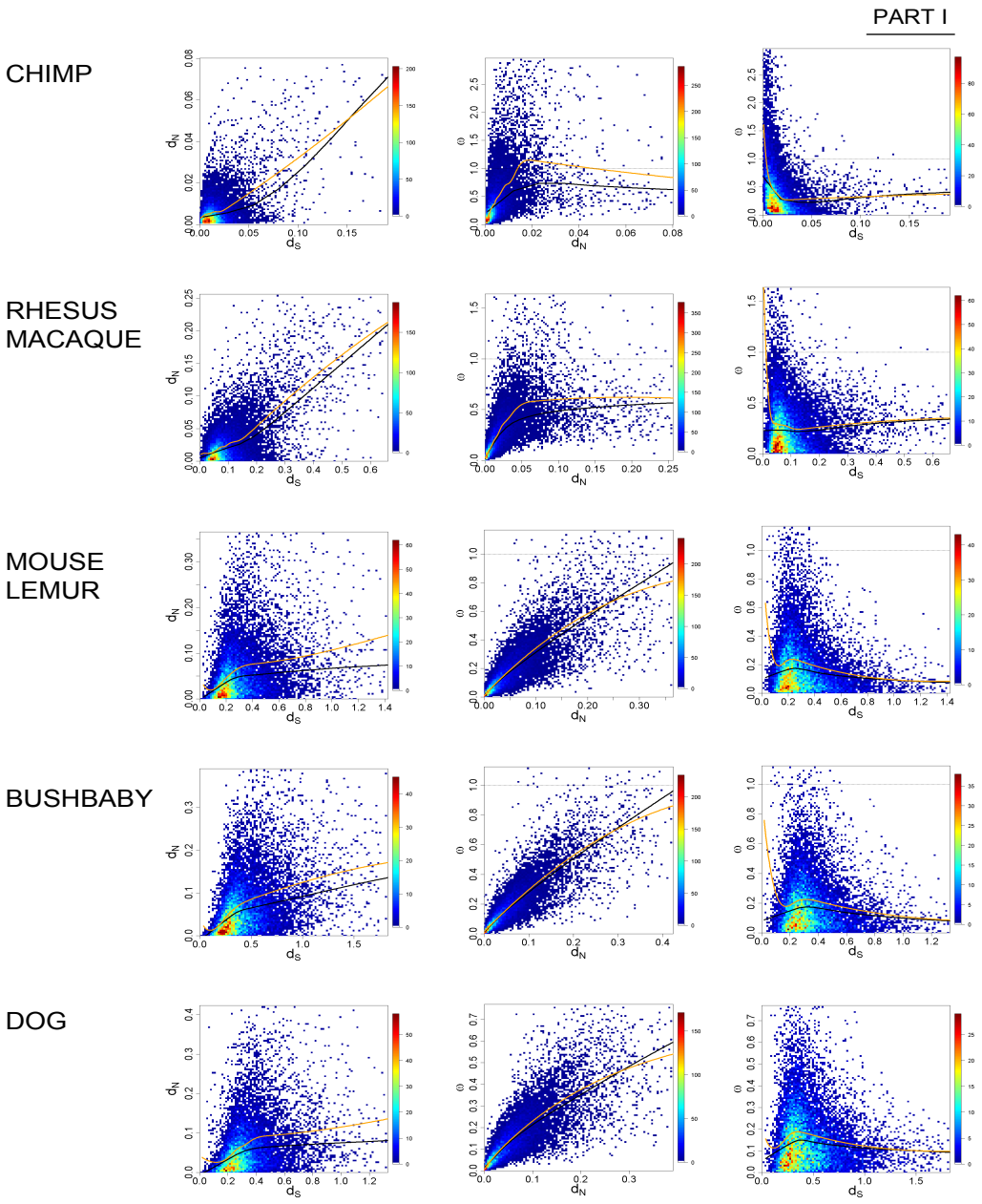

Figure S2: Relationships between measures of protein evolution based on pairwise alignments with human and other species. Left: d_N versus d_S, Middle: ω versus d_N, Right: ω versus d_S. The relationships are depicted as heatmaps and summarized by regression splines selected by BIC model selection (orange line). The number of genes found in each pixel is symbolized by the different colours. Pairwise comparisons are ordered by evolutionary distance starting from the human-chimp comparison to human-dog comparison. The axes are scaled such that they include 99.5% of the data ranges.

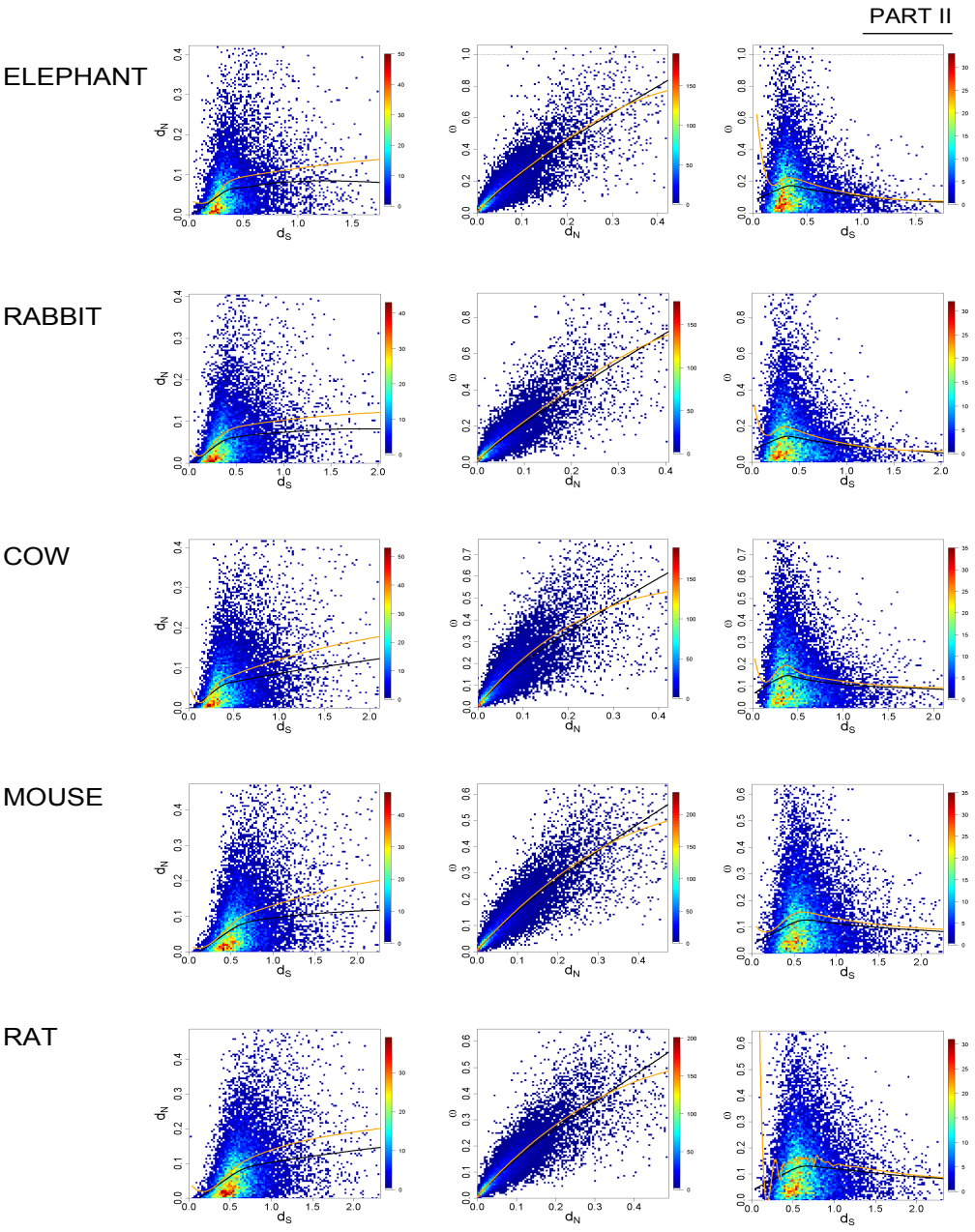

Figure S3: Relationships between measures of protein evolution based on pairwise alignments with human and other species. Left: d_N versus d_S, Middle: ω versus d_N, Right: ω versus d_S. The relationships are depicted as heatmaps and summarized by regression splines selected by BIC model selection (orange line). The number of genes found in each pixel is symbolized by the different colours. Pairwise comparisons are ordered by evolutionary distance starting from the human-elephant comparison to human-rat comparison. The axes are scaled such that they include 99.5% of the data ranges.

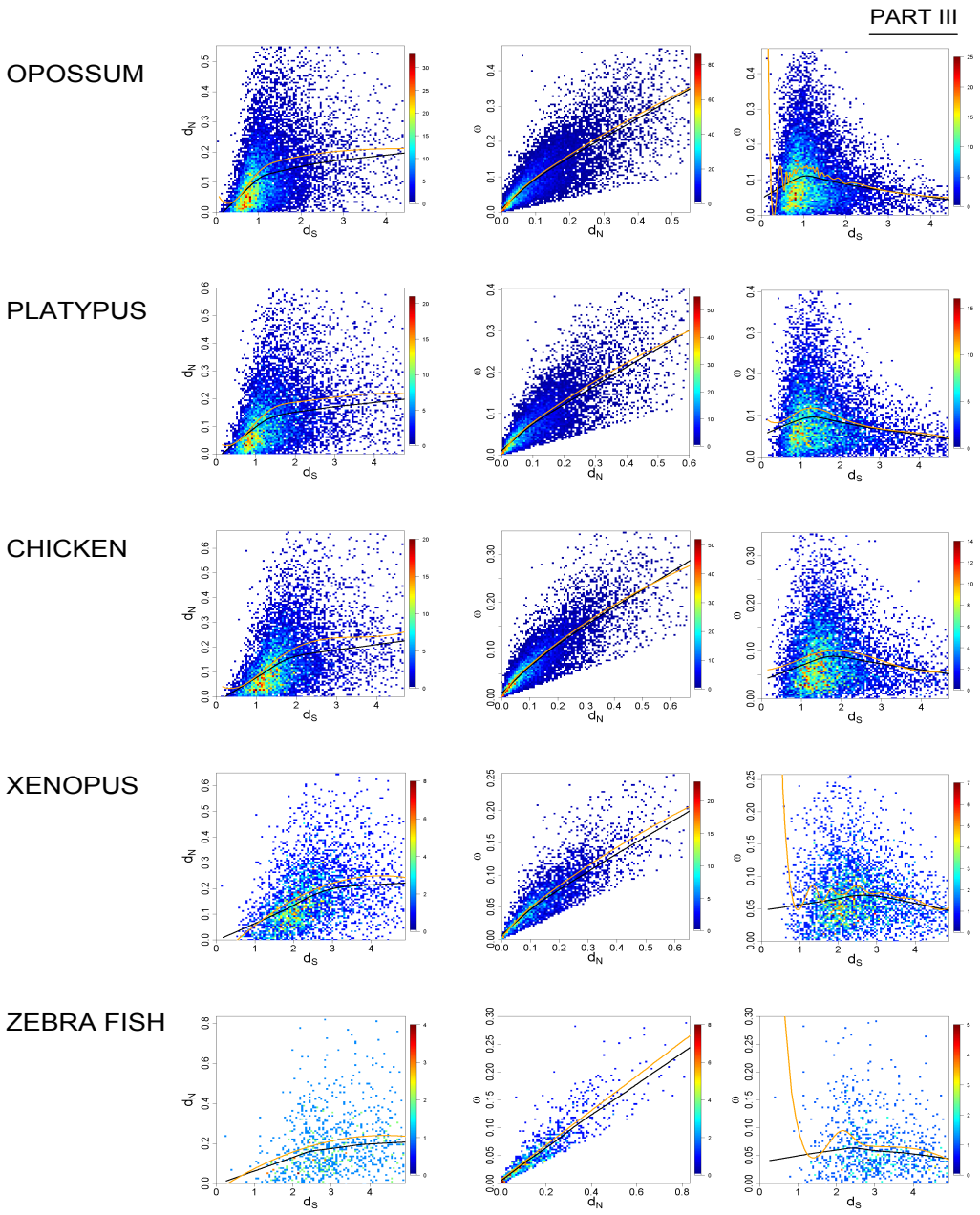

Figure S4: Relationships between measures of protein evolution based on pairwise alignments with human and other species. Left: d_N versus d_S, Middle: ω versus d_N, Right: ω versus d_S. The relationships are depicted as heatmaps and summarized by regression splines selected by BIC model selection (orange line). The number of genes found in each pixel is symbolized by the different colours. Pairwise comparisons are ordered by evolutionary distance starting from the human-opossum comparison to human-zebra fish comparison. The axes are scaled such that they include 99.5% of the data ranges.

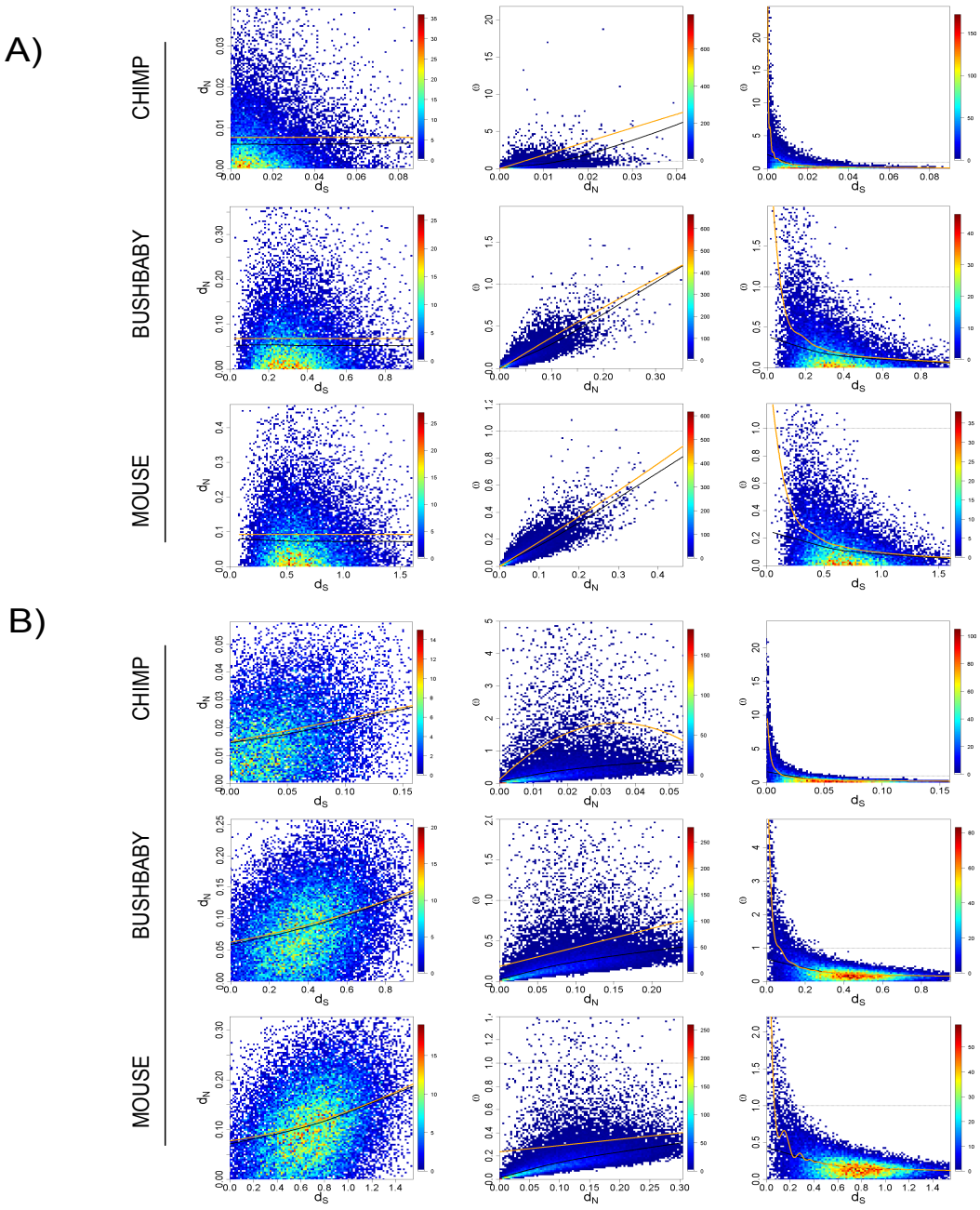

Figure S5: Relationships between measures of protein evolution. Left: d_N versus d_S, Middle: ω versus d_N, Right: ω versus d_S. The relationships are depicted as heatmaps and summarized by regression splines selected by BIC model selection (orange line). The number of genes found in each pixel is symbolized by the different colours. A) Random uncorrelated draws from two gamma distributions whose rate and shape parameters were estimated from pairwise comparisons between human-chimp, human-bushbaby and human-mouse. B) Random correlated draws from two Gaussian distributions whose parameters $N(\mu, \sigma)$ were estimated from the same pairwise comparisons. The axes are scaled such that they include 99.5% of the data ranges.

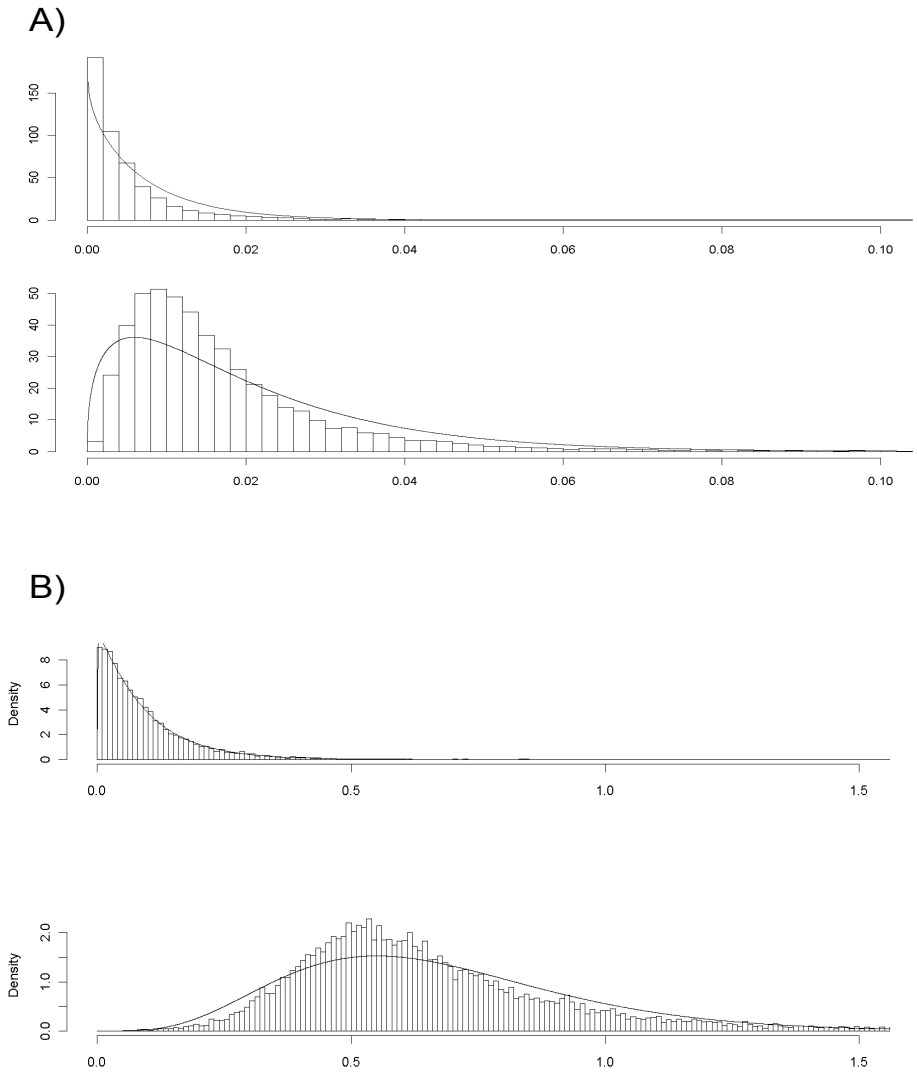

Figure S6: Distributions of d_N and d_S for A) human-chimp, and B) human-mouse pair wise alignments. The straight line depicts a gamma distribution whose parameters were estimated from the empirical data.

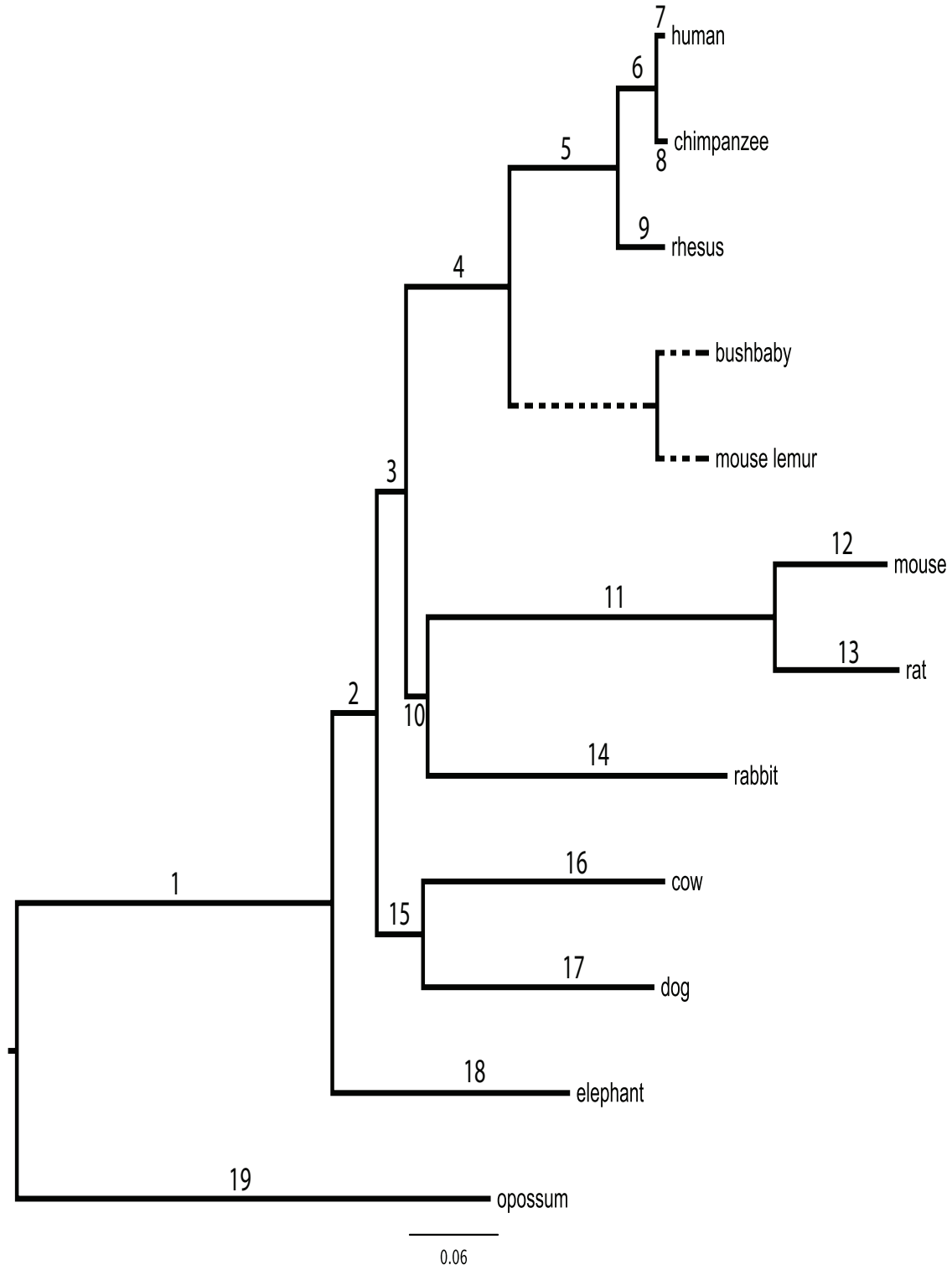

Figure S7: Redrawn subset of phylogenetic tree from Miller *et al.* (2007). Dashed branches are added and have not been used in the analyses. Number coding of branches is used in Figure 1C.

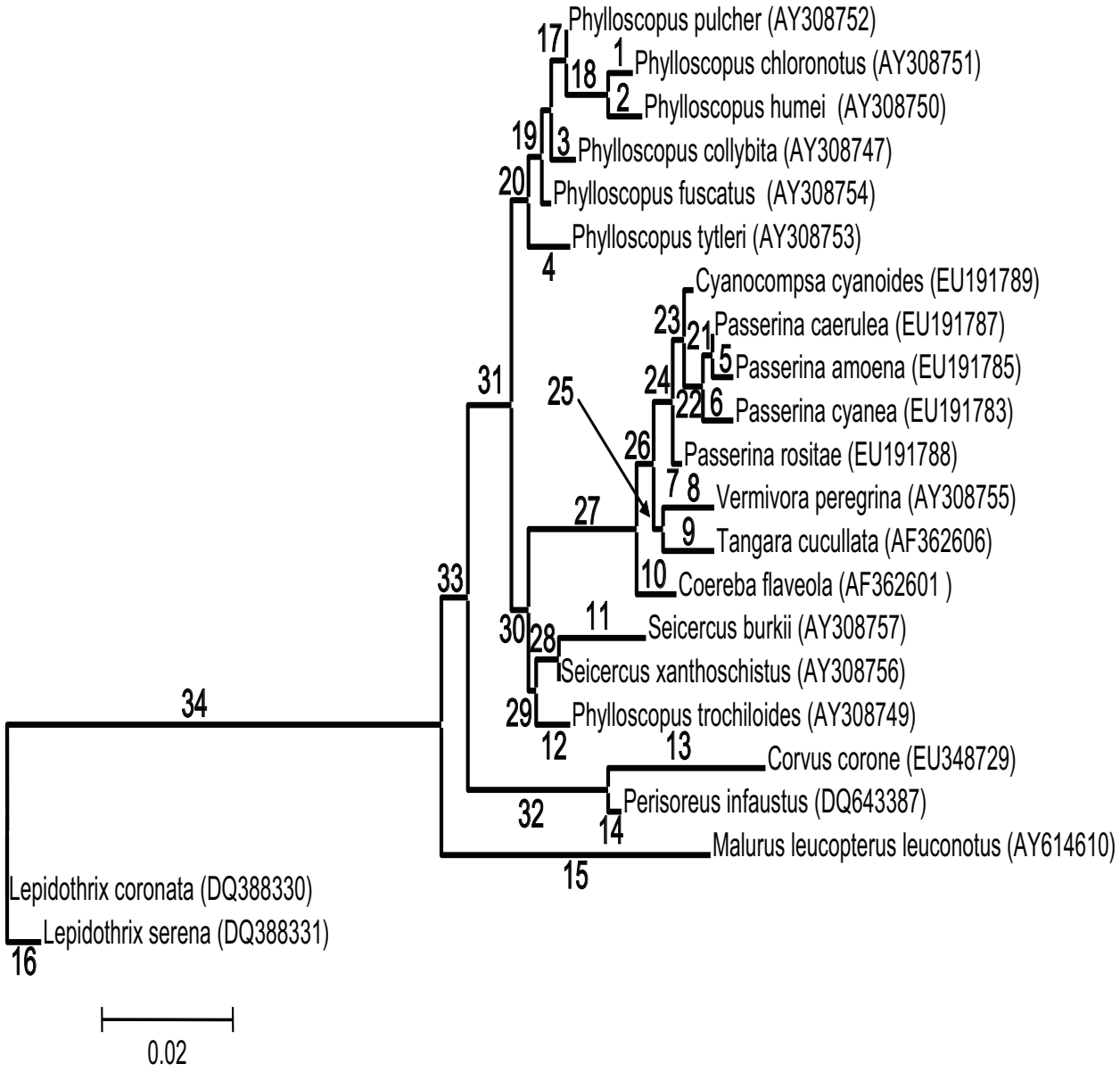

Figure S8: Gene tree of 22 passerine species based on Mc1r sequences. Note that although this tree will most probably not capture the correct phylogenetic relationships between the species, it is useful in understanding the relationship between ω and branch length. Number coding of branches is used in Figure 2B.

2 Simulations

We can simulate data given the process described below in section 2.1. Considering multiple genes, we will first deal with the case when T is the same for all genes, second, the case when T is variable, and third, the case when the substitution rate is variable (and T fixed).

2.1 The model

Let's consider a particular gene for which orthologous genes exist in a pair of species and that these two species diverged T_D units of time ago (see fig. S9), where time is measured in units of $2N$ generations, and N is the population size (of species 1, species 2 and the ancestral species). For this particular gene, the total (scaled) mutation rate for synonymous sites is denoted $\theta_S/2$ and the total (scaled) mutation rate for non-synonymous sites is denoted $\theta_N/2$ so that

$$\frac{\theta_N/2}{\theta_S/2} = \alpha.$$

Note that we will use the term site instead of "possible change", and that we assume that non-synonymous sites ("possible changes") are 3 times as common as synonymous sites. Thus, we view these two sets of sites (and the two mutation processes) as independent of each other.

Let's assume that we have sampled one lineages from each species. Under the infinite sites model (a reasonable model for closely related species at least, and for more divergent species pairs, we assume that every mutation is "seen" despite the possible event of multiple hits for empirical data), mutations are added to a lineage proportional to the length of the branch. In other words, the number of mutations M on a branch of length t is Poisson distributed with parameter $\theta/2t$, $M \sim \text{Po}(\theta/2t)$. Neutral theory dictates that for neutral sites the mutation rate is equal to the substitution rate (Kimura & Ohta, 1971). If we assume that the two species have evolved under identical conditions, we can consider the mutation rate to be equal to the substitution rate (r). Note that whether a particular site is fixed for a specific allele in a species has no effect on the model or the result because we only sample pairs of lineages.

The time till coalescence for two lineages (after they have entered the ancestral population) is denoted T_2. This waiting time is exponentially distributed $T_2 \sim \text{Exp}(1)$, with parameter 1. The total coalescence time for the two lineages is $T_D + T_2 = T$. Assuming no recombination within a gene, all sites in a particular gene (both synonymous and non-synonymous) evolve according to the same genealogy, i.e., all sites within a gene have the exact same coalescent times. We start by assuming that all genes have the same T, we will consider the case of variable T later.

2.2 Expected number of substitutions

The expected number of synonymous substitutions between the two lineages is $E(M_S) = E[\text{Po}(2T \times r_S/2)] = Tr_S$. Similarly, the expected number of non-synonymous substitutions between the two

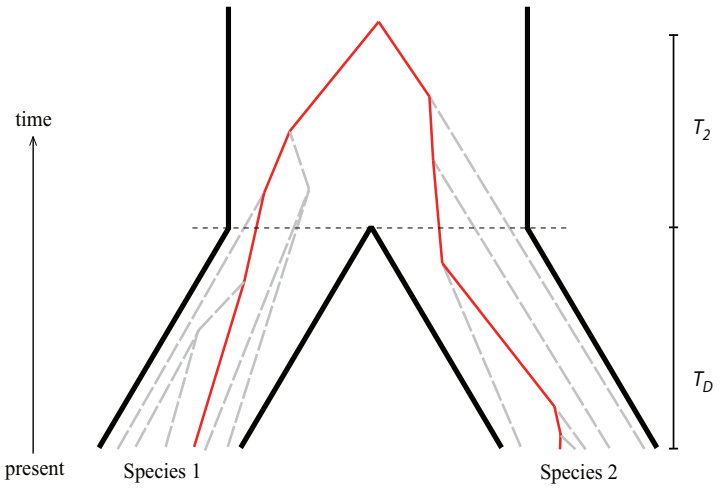

Figure S9: A model of two species. One lineage (red) have been sampled from each species. The divergence time is denoted T_D and the waiting time for two lineages to coalesce is denoted T_2.

lineages is $E(M_N) = E[\text{Po}(2T \times r_N/2)] = Tr_N$. Thus,

$$\frac{E[M_N]}{E[M_S]} = \frac{E[\text{Po}(2T \times r_N/2)]}{E[\text{Po}(2T \times r_S/2)]} = \frac{Tr_N}{Tr_S} = \alpha.$$

If the synonymous mutation rate *per site* is equal to the non-synonymous mutation rate *per site* ($\theta_S = \theta_N$; no selection), we expect 3 times as many non-synonymous substitutions as synonymous substitutions, i.e., $\frac{r_N/2}{r_S/2} = \alpha = 3$. Denote the number of non-synonymous sites by n and the number of synonymous sites by s then $n/s = 3$. On average, the number of non-synonymous substitutions per (non-synonymous) site ($M_N/n = d_N$) would be expected to equal the number of synonymous substitutions per (synonymous) site ($M_S/s = d_S$), so that $E(d_N)/E(d_S) = 1$.

Note however, that the expected value for the ratio of M_N/M_S will not recover the ratio of the non-synonymous rate and the synonymous rate (see for example Heijmans, 1999),

$$E\left[\frac{M_N}{M_S}\right] \neq \frac{E[M_N]}{E[M_S]},$$

nor will the expected value for the ratio d_N/d_S recover the ratio of the expected number non-synonymous substitutions per site and the expected number of synonymous substitutions per site,

$$E\left[\frac{d_N}{d_S}\right] \neq \frac{E[d_N]}{E[d_S]}.$$

Thus, the mean of M_N/M_S across genes, will be a biased estimator of the ratio of the number of non-synonymous substitutions and the number of synonymous substitutions and the mean of d_N/d_S will be a biased estimator of the number of non-synonymous substitutions per site relative

to the number of synonymous substitutions per site. Therefore, a "species d_N/d_S" computed from averaging d_N/d_S across genes may potentially be misleading. In the following, we denote:

$$\omega = \frac{d_N}{d_S},$$

$$\bar{\omega} = \sum_{i \in C} [\omega_i]/n = \sum_{i \in C} \left[\frac{d_{N,i}}{d_{S,i}}\right]/n,$$

where the set C contains all genes with $d_S > 0$, and n is the number of genes in C, and

$$\psi = \frac{\sum d_N}{\sum d_S}.$$

(Note that when we compare $\bar{\omega}$ and ψ, we use the same set of genes (C) to compute $\bar{\omega}$ and ψ.) We will evaluate the bias of $\bar{\omega}$ by:

$$\text{bias} = \frac{\bar{\omega} - E[d_N]/E[d_S]}{E[d_N]/E[d_S]},$$

for all genes with $d_S > 0$.

2.3 Fixed T

We can think of a fixed value of T as the case of multiple genes (in the genome), where we ignore the stochastic process of picking a common ancestor for two lineages in the ancestral species.

2.3.1 The neutral case

We start by considering the neutral case, that is when $r_N = 3r_S$ and $\alpha = 3$.

Simulation 1: Neutral case – human-chimp
Let's assume that the effective population sizes of humans and chimps is 10,000 individuals each and the generation time is 25 years. A divergence time of 6 million years corresponds to 12 units of time (scaled in 2N generations). Assume that $r_S = 5/3$, then $r_N = 5$ since $\alpha = 3$ and $M_S = \text{Po}(r_S T) = \text{Po}(20)$ and $M_N = \text{Po}(r_N T) = \text{Po}(60)$. If we assume that there are 1000 synonymous sites and 3000 non-synonymous sites, we have $E(d_N) = E[\text{Po}(60)]/3000 = 0.02$ and $E(d_S) = E[\text{Po}(20)]/1000 = 0.02$, and $E(d_N)/E(d_S) = 1$.

If we simulate 1,000,000 replicate genes, we find a $\overline{\omega}$ of 1.2471 (excluding the cases when $M_S = 0$), $\psi = 0.9998$ (excluding cases when $M_S = 0$) and that $\overline{\omega}$ is upward biased about 24.7% – the correlation between ω and d_S equals -0.8297, and correlation between ω and d_N equals 0.4517.

The distribution of ω-values can be seen in fig. S10.

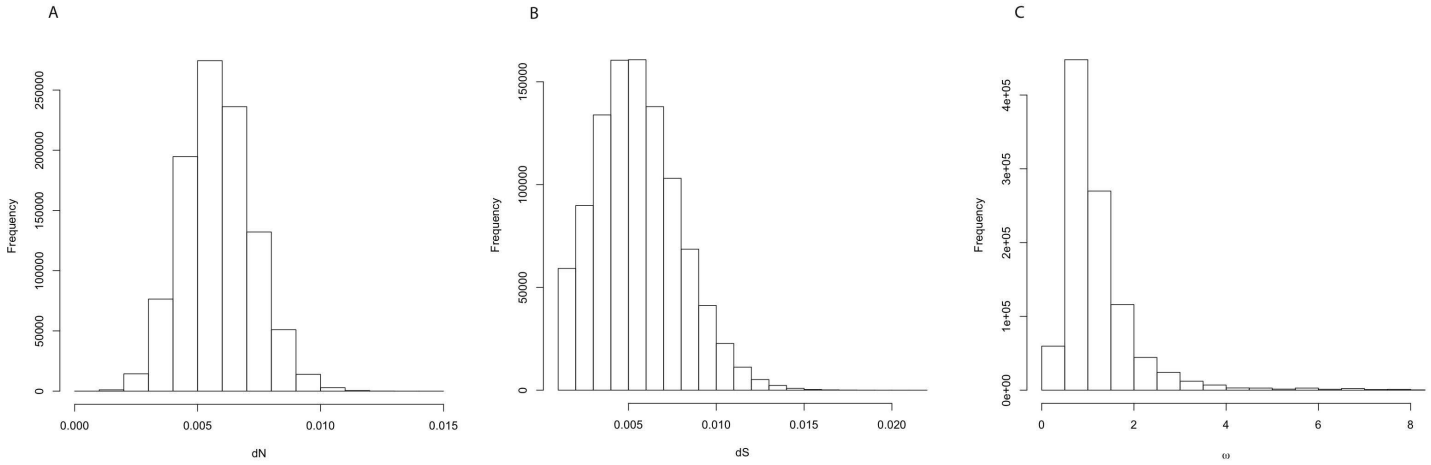

Figure S10: Simulated ω-values (1,000,000 replicates) according to the description under simulation 1. $\overline{\omega}$ is 1.2471 (the expected value is 1; $\psi = 0.9998$). The correlation between ω and d_S is -0.8297 ($p \approx 0$) and the correlation between ω and d_N is 0.4517 ($p \approx 0$).

When comparing these levels of d_N and d_S to empirical data, it is obvious that we observe far to many non-synonymous substitutions relative to synonymous substitutions for the simulated data, and that the ω-values are consequently, on average, much higher than for empirical data (see main text and section 2.6).

2.3.2 Purifying selection on non-synonymous sites

It is likely that purifying selection is the cause of the low level of non-synonymous substitutions observed in empirical data (Ohta, 1992). To adjust the simulations to better mimic empirical data, we can set the substitution rate for non-synonymous sites lower than for synonymous sites. We assume that there is purifying selection acting on non-synonymous sites so that that only a fifth of non-synonymous mutations reach fixation and become substitutions. We can adjust the non-synonymous substitution rate by 3/10, and we have $r_S = 5/3$, $r_N = 3/2$ and $\alpha = 9/10$. If we again assume that there are 1000 synonymous sites and 3000 non-synonymous sites, we have $E(d_N) = E[\text{Po}(18)]/3000 = 0.0060$ and $E(d_S) = E[\text{Po}(20)]/1000 = 0.020$, and $E(d_N)/E(d_S) = 0.30$.

Simulation 2a: Purifying selection – human-chimp
We adjust our model parameters in a manner so that we get the same mean values for d_N and d_S as in our investigated empirical data. In the empirical data, we observed d_N with 0.0062 and d_S with 0.020. We have $E(d_N) = E[\text{Po}(18)]/3000 = 0.0060$ and $E(d_S) = E[\text{Po}(20)]/1000 = 0.020$, and $E(d_N)/E(d_S) = 3/10$. If we again simulate 1,000,000 replicates and exclude all cases when $M_S = 0$, $\overline{\omega}$ is 0.3170 and $\psi = 0.3002$. The upward bias is 5.7%. The distribution of ω can be seen in figure S11. In figure S12 we show density plots of d_N versus d_S, ω versus d_N, and ω versus d_S.

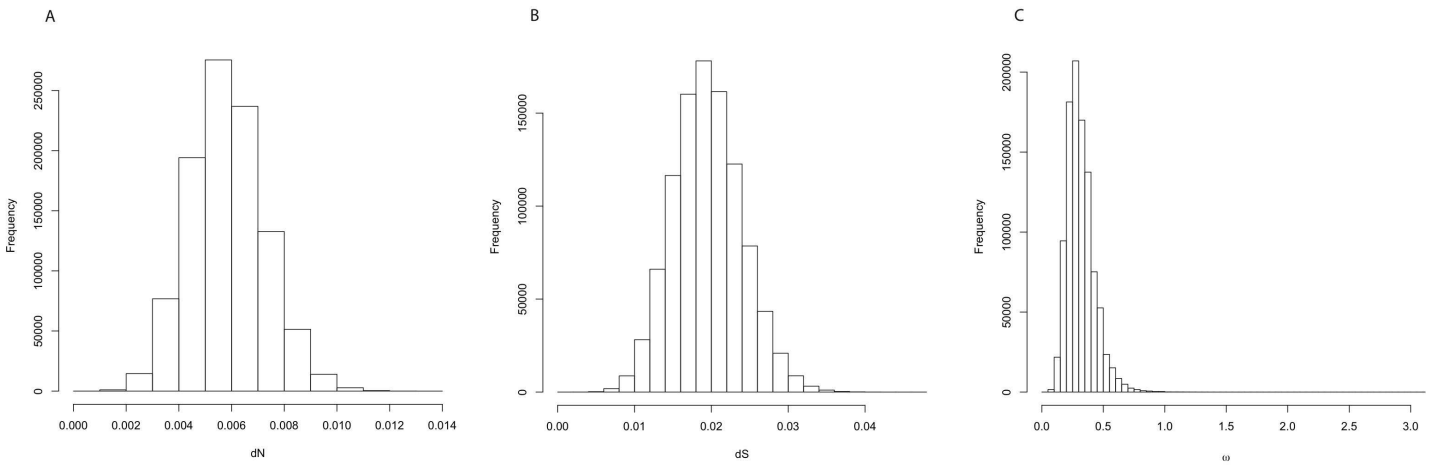

Figure S11: Simulated ω-values (1,000,000 replicates) according to simulation 2a. $\bar{\omega}$ equals 0.3170 and ψ equals 0.3002 (the empirical value for ψ is 0.3061).

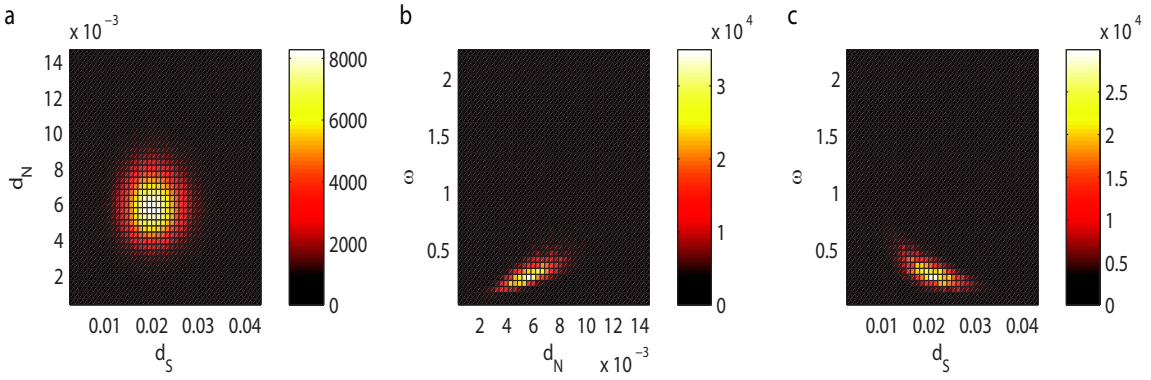

Figure S12: Heatmap of ω (1,000,000 replicates) according to simulation 2a. d_N plotted versus d_S and ω. The correlation between ω and d_S equals -0.674 ($p \approx 0$) and the correlation between ω and d_N equals 0.672 ($p \approx 0$).

Simulation 2b: Purifying selection – humans vs multiple species

Let's assume that the generation time and population size is modified for each species so that one unit of scaled time corresponds to the same amount of scaled time in all species (that is, we assume that rates of substitution are the same rate in all species). We simulate M_S and M_N for a pair of species where the first could be humans, and the second species diverged from the first 6 million years (12 time units), the third species diverged from the first species 12 million years ago, and so on for the fourth and fifth species at increasing intervals of 6 million years ago, until 90 million years ago. The substitution rates are the same as in simulation 2a and we simulate 1,000,000 replicates.

For these pairwise species-comparisons, we see that $\bar{\omega}$ decreases with T, mean d_S, and mean d_N (fig. S13 shows $\bar{\omega}$ and ψ as functions of d_S). The estimator ψ is nearly constant across the range of T, mean d_S and mean d_N. Table S1 shows a summary of the results at different time points.

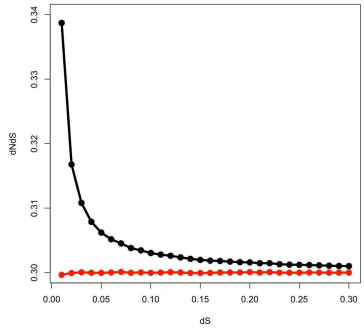

Figure S13: $\overline{\omega}$ and ψ as a functions of mean d_S, according to simulation 2b (1,000,000 replicates for each dot). The black line denotes values for $\overline{\omega}$ and the red line values for ψ.

Table S1: Simulation of different time points using simulation 2b (every time point contains 1,000,000 replicates).

N	S	d_N	d_S	corr d_N, ω	corr d_S, ω	$\overline{\omega}$	ψ	bias (%)	$\omega > 1$ (%)
9	10	0.0030	0.0100	0.57	-0.63	0.3391	0.3000	13.0	0.86
18	20	0.0060	0.0200	0.67	-0.67	0.3170	0.3002	5.7	0.04
27	30	0.0090	0.0300	0.69	-0.68	0.3110	0.3000	3.7	0.00
36	40	0.0120	0.0400	0.70	-0.68	0.3080	0.3000	2.7	0.00
90	100	0.0300	0.1000	0.71	-0.69	0.3031	0.3000	1.0	0.00
180	200	0.0600	0.2000	0.72	-0.69	0.3015	0.3000	0.5	0.00
360	400	0.1200	0.4000	0.72	-0.69	0.3008	0.3000	0.3	0.00
540	600	0.1800	0.6000	0.72	-0.69	0.3005	0.3000	0.2	0.00

2.4 Variable T

The time T_2 is a random variable and it is exponentially distributed ($T_2 \sim \text{Exp}(1)$). For two lineages, the total coalescent time T is the sum of T_D and T_2 (see fig. S9). We augmented our model so that it can deal with different coalescent times for different genes.

Simulation 3: Purifying selection and variable T – human-chimp
We simulate data following simulation 2a, with the modification to allow T_2 to vary. The value for $\overline{\omega}$ is 0.3168, the ψ is 0.3000. Again, we observe a slightly overestimation if we use $\overline{\omega}$ to estimate mean d_N/d_S, the upward bias is 5.6%. The correlation between ω and d_S equals -0.657 and the correlation between ω and d_N equals 0.655. Both correlations are extremely similar to the values that we obtained in simulation 2a. The distributions of ω-values are plotted in figures S14 and S15. We note that the result of this model is very similar to the previous model, the model without variable T_2. This means that incorporating a realistic coalescent time for the pair of lineages in the ancestral species has a negligible effect on the results, at least for the parameters we are interested in here. We will therefore proceed without using this modification of the model.

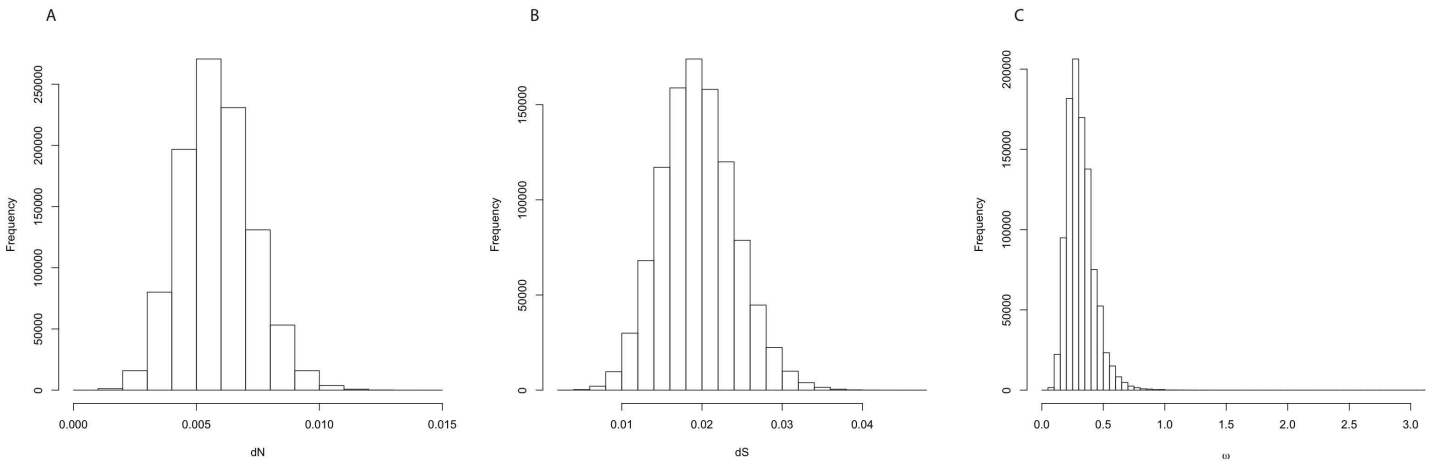

Figure S14: Simulated ω-values (1,000,000 replicates) according to simulation 3. $\overline{\omega}$ equals 0.3168 and $\psi = 0.3000$. The empirical value of ψ is 0.3061 for the human-chimp comparison.

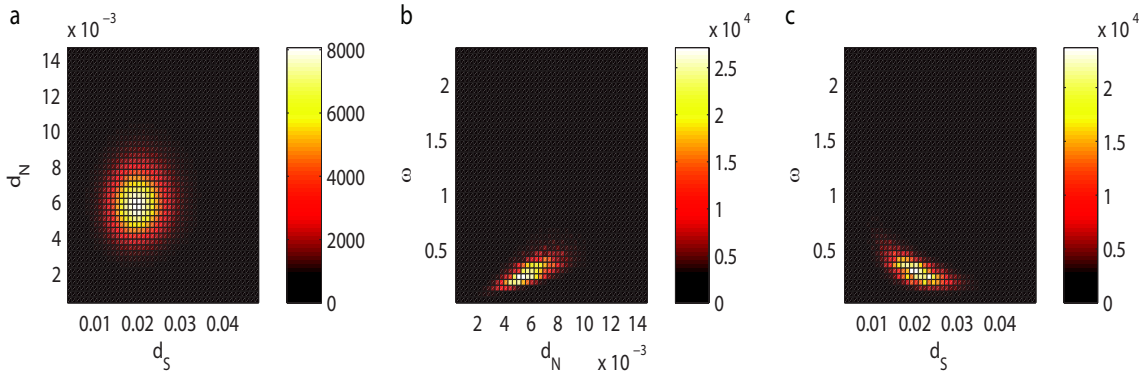

Figure S15: Heatmap of ω. d_N plotted versus d_S and ω versus d_N and d_S according to simulation 3. The correlation between ω and d_S equals -0.657 ($p \approx 0$) and the correlation between ω and d_N equals 0.655 ($p \approx 0$).

2.5 Varying substitution rate

We can let substitution rate vary across genes, for example we can draw a value of the substitution rate for a particular gene from some distribution (e.g. a normal distribution) with mean equal to the mean substitution rate (and with some variance). One can imagine a substitution rate that is variable across genes (but which keeps α constant) or two independent substitution rates for r_N and r_S, or something in between the two previous extreme cases.

Simulation 4: Variable substitution rate – humans vs multiple species
We choose to evaluate three different levels of correlation between r_S and r_N: uncorrelated, correlation of 0.4 and correlation of 0.8. For the first case, we draw r_N and r_S from two different uncorrelated gamma distributions, r_N-values are drawn from $\Gamma(9/10, 1)$ and the r_S-values are drawn from $\Gamma(1, 1)$ (Simulation 4a). Note that if the shape parameter is set to 1, a gamma distribution is an exponential distribution. The randomly drawn values are then scaled by a factor of 20 to get

the same mean as in the simulations 2a, 2b and 3. In the second and third case, we draw r_N-values and r_S-values from two correlated gamma distributions with $\Gamma(1,1)$ using the R package *splus2R*. In simulation 4b, we assume correlation ρ of 0.4 between the two distributions (ρ was estimated to 0.4 for our empirical data, see 2.6) and in simulation 4c, we assume a correlation of 0.8.

We simulate 1,000,000 replicates for each of these three cases. The estimator ψ is close to the expected value in all three simulations, 4a: 0.2852, 4b: 0.3061 and 4c: 0.3118. The estimator $\overline{\omega}$ is upward biased in all three cases (4a: $\overline{\omega} = 0.9139$, 4b: $\overline{\omega} = 0.7059$, 4c: $\overline{\omega} = 0.4665$). The correlation between ω and d_S is very similar, for all approaches (4a: -0.292, 4b: -0.292, 4c: -0.247). The correlation between ω and d_N equals 0.450 for simulation 4a, $\rho = 0.264$ for simulation 4b, and $\rho = 0.094$ for simulation 4c. Interestingly, if we simulate different time points, $\overline{\omega}$ does not decrease with d_S as in simulation 2b. Rather it increases over time (see fig. S16). In tables S2 and S3, we show the results from simulation 4b and 4c for a divergence time from 6 millions years to 90 million years.

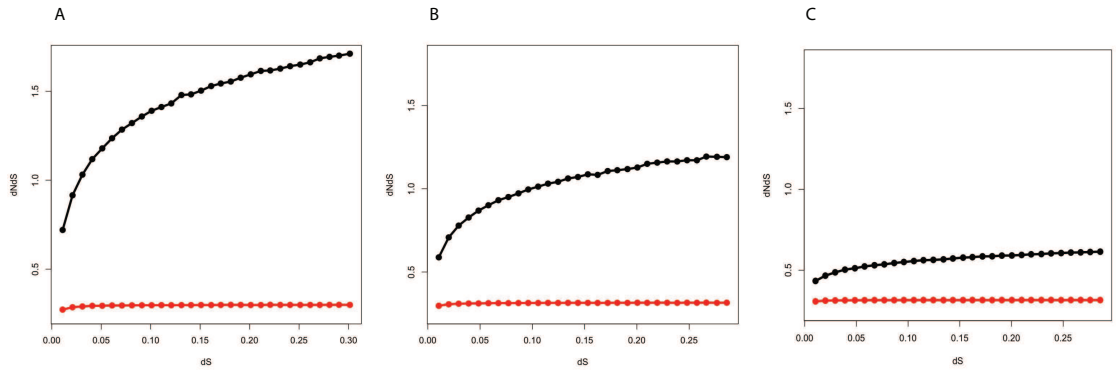

Figure S16: The estimators $\overline{\omega}$ and ψ as a functions of mean d_S according to simulation 4.1, 4.2 and 4.3 (1,000,000 replicates for each point in the plots). The black line denotes values for $\overline{\omega}$ and the red line values for ψ. A. Simulation 4a with two uncorrelated gamma distributions, B. Simulation 4b with two correlated gamma distributions ($\rho = 0.4$), and C. Simulation 4c with two correlated gamma distributions ($\rho = 0.8$).

Table S2: Summary of the results from simulation 4b at different time points (two correlated gamma distributions $\rho = 0.4$). The results for each time point is based on 1,000,000 replicates). N* denotes the expected value for non-synonymous substitutions, S* denotes the expected value for synonymous substitutions.

N*	S*	d_N	d_S	corr d_N, ω	corr d_S, ω	$\overline{\omega}$	ψ	bias (%)	$\omega > 1$ (%)
9	9.5	0.0031	0.0105	0.37	-0.31	0.5906	0.2971	96.9	14.32
18	19	0.0061	0.0200	0.27	-0.29	0.7056	0.3061	135.2	16.37
27	28.5	0.0091	0.0295	0.22	-0.27	0.7764	0.2949	155.8	17.19
36	38	0.0121	0.0390	0.19	-0.26	0.8299	0.3109	176.6	17.69
90	95	0.0301	0.0958	0.13	-0.21	0.9951	0.3142	231.7	18.52
180	190	0.0601	0.1908	0.09	-0.18	1.1300	0.3148	276.7	18.77
360	380	0.1200	0.3811	0.07	-0.14	1.2496	0.3148	231.7	18.91
540	570	0.1800	0.5709	0.06	-0.12	1.3291	0.3153	343.0	18.99

Table S3: Summary of the results from simulation 4c at different time points (two correlated gamma distributions $\rho = 0.8$). The results for each time point is based on 1,000,000 replicates). N^* denotes the expected value for non-synonymous substitutions, S^* denotes the expected value for synonymous substitutions.

N^*	S^*	d_N	d_S	corr d_N, ω	corr d_S, ω	$\overline{\omega}$	ψ	bias (%)	$\omega > 1$ (%)
9	9.5	0.0032	0.0105	0.20	-0.25	0.4341	0.3075	44.7	7.53
18	19	0.0062	0.0200	0.09	-0.25	0.4670	0.3119	55.7	8.26
27	28.5	0.0092	0.0295	0.05	-0.24	0.4878	0.3138	62.6	8.66
36	38	0.0122	0.0390	0.03	-0.23	0.5010	0.3141	67.0	8.81
90	95	0.0303	0.0961	-0.01	-0.20	0.5513	0.3152	83.8	9.30
180	190	0.0603	0.1915	-0.03	-0.17	0.5902	0.3151	96.7	9.47
360	380	0.1199	0.3802	-0.03	-0.14	0.6326	0.3154	110.9	9.63
540	570	0.1803	0.5715	-0.03	-0.12	0.6572	0.3154	119.1	9.60

Our model only assumes purifying selection on non-synonymous sites, despite this, we observe genes with $\omega > 1$ for the three approaches with 19.9%, 16.4% and 8.3% for a divergence time of 12 million years. These are non-trivial fractions of genes that show signals of positive selection, in other words, false positives.

2.6 Empirical Data

Observation from empirical human-chimp data (17.226 genes):

$\overline{\omega} = 0.374509$
$\psi = 0.3060981$
$d_N = 0.006191826$
$d_S = 0.02022824$
$\rho_{\omega - d_N} = 0.846$ (Spearman)
$\rho_{\omega - d_S} = -0.178$ (Spearman)
$\rho_{d_N - d_S} = 0.4$ (Spearman)
Γ-distribution fitted to d_N: $\Gamma(0.9236, 123.8)$
Γ-distribution fitted to d_S: $\Gamma(1.416, 70.00)$

3 Supplementary tables

Table S4: Contingency table summarizing the number of genes having naught, one, two or three synonymous mutations for a candidate set of 50 positively selected genes and genes from the genomic background (Nielsen *et al.*, 2005). The residuals of χ^2 as calculated from a contingency table are normally distributed and can be used as an indication for the significance of individual deviations from expected values. Significance levels are added to the category where genes are either over- or under-represented in the common asterisk convention.

Number of synonymous substitutions	Observed/expected number of genes with a certain number of synonymous substitutions	
	Candidate set	Full set
0	39/15.3***	2773/2796.7
1	7/14.7*	2695/2687.3
2	2/11.5**	2122/2112.5
3	2/8.5*	1570/1563.5

References

Heijmans, R. 1999. When does the expectation of a ratio equal the ratio of expectations?, *Statistical Papers* **40**, 107–115.

Kimura, M. and Ohta, T. 1971. "Theoretical Aspects of Population Genetics", Princeton University Press, Princeton.

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S. L., Nekrutenko, A., Giardine, B., Harris, R. S., Tyekucheva, S., Diekhans, M., Pringle, T. H., Murphy, W. J., Lesk, A., Weinstock, G. M., Lindblad-Toh, K., Gibbs, R. A., Lander, E. S., Siepel, A., Haussler, D., and Kent, W. J. 2007. 28-way vertebrate alignment and conservation track in the ucsc genome browser, *Genome Research* **17**, 1797–1808.

Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees, *PLoS Biology* **3**, 0976–0985.

Ohta, T. 1992. The nearly neutral theory of molecular evolution, *Annual Review of Ecology and Systematics* **23**, 263–286.