Supplemental Material

Extreme expansion of the olfactory receptor gene repertoire in African elephants and evolutionary dynamics of orthologous gene groups in 13 placental mammals

Yoshihito Niimura[1,2†], Atsushi Matsui[1,2], Kazushige Touhara[1,2]

1. Department of Applied Biological Chemistry, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan
2. ERATO Touhara Chemosensory Signal Project, JST, The University of Tokyo, Tokyo 113-8657, Japan

Dataset S1. Nucleotide sequences of the OR genes in African elephant, horse, cow, rabbit, guinea pig, and mouse.

Dataset S2. Amino acid sequences of the OR genes in African elephant, horse, cow, rabbit, guinea pig, and mouse.

Dataset S3. List of genes contained in each OGG.

Table S1. Number of OR genes from each species in each OGG.

Table S2. OGGs that contained the largest number of intact genes within each species

Species	OGG	# of intact genes
Human	OGG2-14	5
Chimpanzee	OGG2-14	8
Orangutan	OGG2-1	9
Macaque	OGG2-14	4
Marmoset	OGG2-25	6
Mouse	OGG2-4	21
Rat	OGG2-8	24
Guinea pig	OGG2-6	22
Rabbit	OGG2-1	28
Cow	OGG2-1	43
Dog	OGG2-1	25
Horse	OGG2-12	15
Elephant	OGG2-2	84

Table S3. OGGs showing lineage-specific gene expansion

OGG	Species	Expansion rate (*R*)	# of intact genes	Adjusted total # of genes	# of different orders
OGG2-22	Elephant	0.926	46	49.67	5
OGG2-2	Elephant	0.738	84	113.87	7
OGG2-226	Elephant	0.695	7	10.07	5
OGG2-167	Horse	0.682	10	14.67	5
OGG2-169	Guinea pig	0.682	9	13.20	6
OGG2-94	Dog	0.658	10	15.20	5
OGG2-36	Elephant	0.657	23	35.00	7
OGG2-65	Elephant	0.645	12	18.60	6
OGG2-82	Elephant	0.645	11	17.07	6
OGG2-20	Elephant	0.637	28	43.93	7

The expansion rate *R* represents the extent of lineage-specific gene expansion for each OGG and each species. Let us consider the ratio *r*(OGG, species) = *n*(OGG, species) / *N*(OGG), where *n*(OGG, species) is the number of intact genes for a given species in a given OGG and *N*(OGG) is the total number of intact genes for the 13 species examined. For example, if gene expansion has occurred in the mouse lineage in a given OGG, it is likely that the number of genes in rats in the same OGG also becomes large. Therefore, the ratio *r*(OGG, species) tends to be smaller when evolutionarily closed related species are included in the 13 species examined. For this reason, we used the 'adjusted' total number of genes *N_adj*(OGG), in which 7 different orders are considered, rather than the total number of genes among the 13 species, and calculated the expansion rate *R* as *n*(OGG, species) / *N_adj*(OGG). For the species that is neither rodents nor primates, *N_adj*(OGG) is calculated as the sum of *n*(OGG, elephant), *n*(OGG, cow), *n*(OGG, dog), *n*(OGG, horse), *n*(OGG, rabbit), *n*(OGG, rodents), and *n*(OGG, primates), where *n*(OGG, rodents) and *n*(OGG, primates) are the mean of the numbers of intact genes among three rodent species and that among five primate species, respectively. On the other hand, for rodent or primate species, *n*(OGG, species) instead of *n*(OGG, rodents) or *n*(OGG, primates) was used for the calculation of *N_adj*(OGG) to let the value of *R* be one or less. For example, OGG2-22 contains 43 intact genes from elephant, one intact gene each from cow, dog, rabbit, rat, and mouse, and none from other species. The value of *R* for elephant for OGG2-22 is calculated as 46 / (46 + 1 + 1 + 0 + 1 + 2/3 + 0) = 0.926. In this table, the top 10 OGGs that show high *R* values and contain genes from five or more different orders are listed.

Table S4. Number of OR gene clusters

	# of clusters	# of 5+ clusters
Mouse	72	34
Human	98	35
Elephant	503	148
Intact	57	22
Truncated (one-end)	34	24
Truncated (both-ends)	412	102

An OR gene cluster is defined with the criterion that any distances between two neighboring OR genes (including pseudogenes and truncated genes) in a cluster are <500 kb regardless of the presence of other genes. “5+ cluster” indicates an OR gene cluster containing five or more OR genes (including pseudogenes and truncated genes). Elephant OR gene clusters were classified into three categories, intact clusters, one-end truncated clusters, and both-end truncated clusters. A one-end truncated cluster is defined such that the distance between the end of the OR gene cluster (an OR gene located at the end of the cluster) and the end of the scaffold containing the cluster is <500 kb for one end. A both-end truncated cluster is defined such that the distances are <500 kb for both ends. The other clusters were defined as intact clusters.

Table S5. Comparison of OR gene clusters between mouse and African elephant

Mouse			African elephant						
ID	# of OR genes [a,b]	Length (Mb) [b]	ID	# of OR genes [a,b]	Length (Mb) [b]	# of clusters [c]			
						Total	I	T(o)	T(e)
Mm1.1	8	0.14	s109.1	15	0.33	1	1	0	0
Mm1.2	7	0.31	s33.3	9	0.35	1	1	0	0
Mm1.3	18	0.41	s33.2	35	0.99	1	1	0	0
Mm2.1	35	1.03	s6.1	72	1.38	1	1	0	0
Mm2.2 Mm7.4 Mm10.4 Mm11.1 Mm11.2 Mm11.4 Mm16.3	418 (276, 18, 66, 19, 8, 23, 8)	8.08 (5.07, 0.43, 1.22, 0.45, 0.18, 0.39, 0.34)	s1.5 s1.6 s2.2 s21.1 s46.2 s61.1 s61.2 s96.2 s110.1 s133.1 s115.1 s142.1 s154.1 s155.2 s160.1 s161.1 s162.1 s165.1 s175.1 s179.1 s184.1 s187.1 s188.1 s189.1 s192.1 s194.1 s197.1 s199.1 s200.1 s203.1 s206.1 s210.1 s212.1 s213.1 s215.1 s218.1 s219.1 s222.1 s230.1 s231.1 s234.1 s240.1 s241.1 s242.1 s257.1 s258.1 s263.1 s269.1 s274.1 s275.1 s285.1 s292.1 s300.1 s312.1 s314.1 s315.1 s316.1 s327.1 s334.1 s347.1 s356.1 s366.1 s369.1 s392.1 s399.1 s400.1 s403.1 s409.1 s419.1 s432.1 s443.1 s446.1 s466.1 s481.1 s500.1 s523.1 s573.1 s607.1	1854 (12, 87, 171, 76, 6, 24, 10, 11, 48, 59, 50, 105, 36, 24, 65, 42, 74, 64, 13, 50, 39, 39, 33, 55, 30, 14, 12, 31, 6, 13, 33, 28, 22, 9, 27, 11, 32, 7, 21, 20, 18, 13, 21, 19, 23, 16, 8, 16, 5, 8, 11, 13, 10, 8, 10, 8, 8, 7, 10, 7, 6, 8, 5, 5, 5, 8, 5, 9, 6, 6, 5, 5, 5, 5, 8, 5, 5, 5)	37.14 (0.69, 3.65, 3.58, 1.39, 0.14, 0.49, 0.20, 0.26, 1.54, 2.01, 0.83, 1.62, 1.34, 0.59, 0.95, 0.90, 0.96, 0.92, 0.69, 0.72, 0.71, 0.80, 0.66, 0.57, 0.59, 0.47, 0.44, 0.44, 0.16, 0.44, 0.40, 0.41, 0.37, 0.17, 0.30, 0.31, 0.35, 0.38, 0.28, 0.28, 0.26, 0.24, 0.27, 0.39, 0.32, 0.21, 0.20, 0.23, 0.15, 0.16, 0.21, 0.28, 0.13, 0.11, 0.10, 0.12, 0.21, 0.11, 0.10, 0.08, 0.10, 0.09, 0.08, 0.06, 0.06, 0.07, 0.10, 0.09, 0.06, 0.09, 0.06, 0.04, 0.07, 0.06, 0.05, 0.04, 0.05, 0.03)	78	2	9	67
Mm2.3 Mm14.3	74 (46, 28)	2.10 (0.93, 1.18)	s64.1 s118.2 s174.1 s201.1 s246.1 s256.1 s373.1 s445.1 s476.1	242 (72, 70, 39, 20, 14, 11, 6, 5, 5)	4.98 (1.55, 1.41, 0.73, 0.49, 0.35, 0.19, 0.11, 0.06, 0.07)	9	0	2	7
Mm4.1	8	0.16	s6.4	8	0.22	1	1	0	0
Mm4.2	5	0.15	s6.3	12	0.25	1	1	0	0
Mm4.4 Mm10.2	25 (15, 10)	0.79 (0.27, 0.52)	s26.3 s34.1	37 (29, 8)	1.10 (0.87, 0.24)	2	2	0	0
Mm6.3	25	0.78	s91.1 s172.1	37 (27, 10)	1.35 (0.74, 0.61)	2	0	1	1
Mm7.2	8	0.13	s4.2	9	0.15	1	1	0	0
Mm7.6	158	2.89	s21.4 s79.2 s116.1 s180.1 s211.1 s237.1 s255.1 s277.1 s282.1	353 (5, 15, 218, 43, 26, 12, 10, 14, 10)	5.42 (0.11, 0.33, 3.14, 0.63, 0.44, 0.23, 0.24, 0.14, 0.16)	9	0	2	7
Mm7.7	28	0.47	s21.3 s330.1	44 (36, 8)	0.64 (0.51, 0.13)	2	1	0	1
Mm7.8	49	1.08	s21.2 s387.1	64 (59, 5)	1.10 (1.02, 0.08)	2	1	0	1
Mm7.9	22	0.54	s72.1	7	0.16	1	0	1	0
Mm9.2	46	1.56	s26.6 s181.1 s195.1 s223.1 s229.1 s238.1 s260.1 s262.1 s333.1 s341.1 s346.1 s467.1	176 (67, 25, 14, 16, 9, 12, 5, 8, 5, 5, 5, 5)	3.92 (1.64, 0.74, 0.30, 0.30, 0.18, 0.23, 0.07, 0.16, 0.07, 0.09, 0.07, 0.08)	12	0	1	11
Mm9.3	118	2.48	s50.1 s58.1 s131.1 s144.1 s158.1 s191.1 s236.1	286 (6, 88, 83, 48, 41, 9, 11)	7.93 (0.09, 2.33, 2.12, 1.64, 1.07, 0.39, 0.29)	7	0	2	5
Mm11.6	45	1.01	s47.1	35	0.94	1	1	0	0
Mm13.1 Mm17.3	68 (13, 55)	1.88 (0.63, 1.25)	s108.2 s135.1 s122.1 s301.1	136 (21, 30, 79, 6)	3.75 (0.49, 1.11, 2.06, 0.09)	4	1	2	1
Mm14.4	8	0.46	s118.1	7	0.55	1	1	0	0
Mm15.1	10	0.31	s2.1	40	1.04	1	1	0	0
Mm16.4	31	0.60	s13.1 s66.1	34 (16, 18)	0.78 (0.39, 0.39)	2	0	2	0
Mm19.1	79	2.09	s71.1 s115.2 s168.1	128 (29, 58, 41)	2.49 (0.64, 1.16, 0.69)	3	1	1	1
MmX.1	5	0.74	s100.1	10	0.51	1	1	0	0

a. The total number of intact genes, truncated genes, and pseudogenes is shown.

b. When two or more clusters are shown in one line, the total number or the total length for all clusters are shown, and the number or the length of each cluster is shown in parentheses.

c. "I", "T(o)", "T(b)" represent intact clusters, one-end truncated clusters, and both-end truncated clusters, respectively. See the legend of Table S4.

Table S5 (cont'd).

OR gene cluster(s) in mouse and those in African elephant are shown in the same line when mouse and elephant clusters contain OR genes belonging to the same OGG. Only 5+ clusters are considered. For example, a mouse cluster Mm1.2 contains OR genes in OGG2-184, OGG2-203, OGG2-360, and OGG2-380, while an elephant cluster s33.3 contains OR genes in OGG2-184, OGG2-203, OGG2-360, OGG2-380, OGG2-454, OGG2-497, and OGG2-609; no other 5+ clusters in mouse or elephant contain OR genes in these OGGs. Each mouse cluster is designated by a chromosome number and an index number of a cluster on the chromosome; for example, Mm1.2 is the second cluster from the centromere on chromosome 1. Each elephant cluster is designated by a scaffold number and a cluster index number; for example, s33.3 is the third cluster on scaffold33.

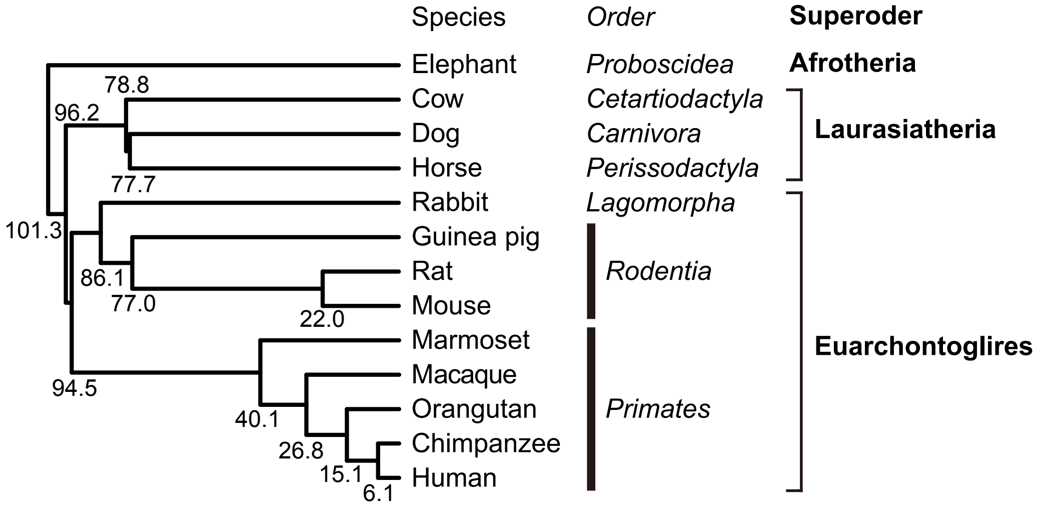


Figure S1. Phylogeny of 13 placental mammal species investigated in this study. The divergence time (million years ago) is shown at each node. For each node, the median of divergence times from TimeTree (http://www.timetree.org/) was used.

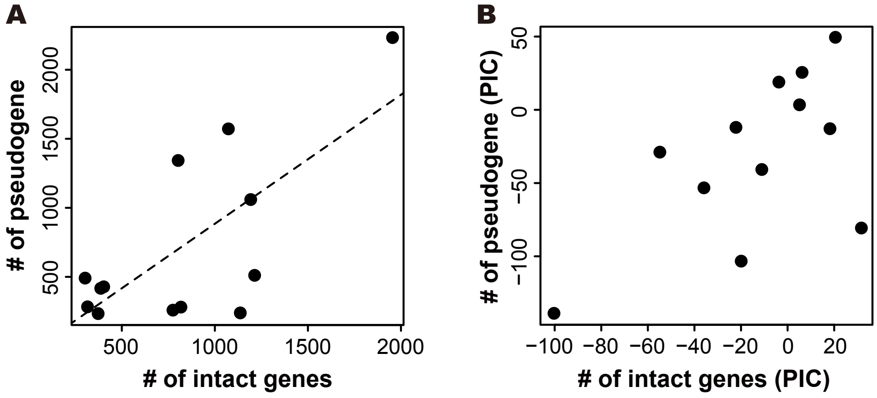


Figure S2. Correlation between the number of intact genes and the number of pseudogenes. (*A*) *r* = 0.713; *p* = 0.0062. The dashed line indicates the regression line, *y* = -49.6 + 0.934*x*. (*B*) PIC indicates a phylogenetically independent constant. Significant correlation was observed even after removing phylogenetic dependence (*r* = 0.593; *p* = 0.041).

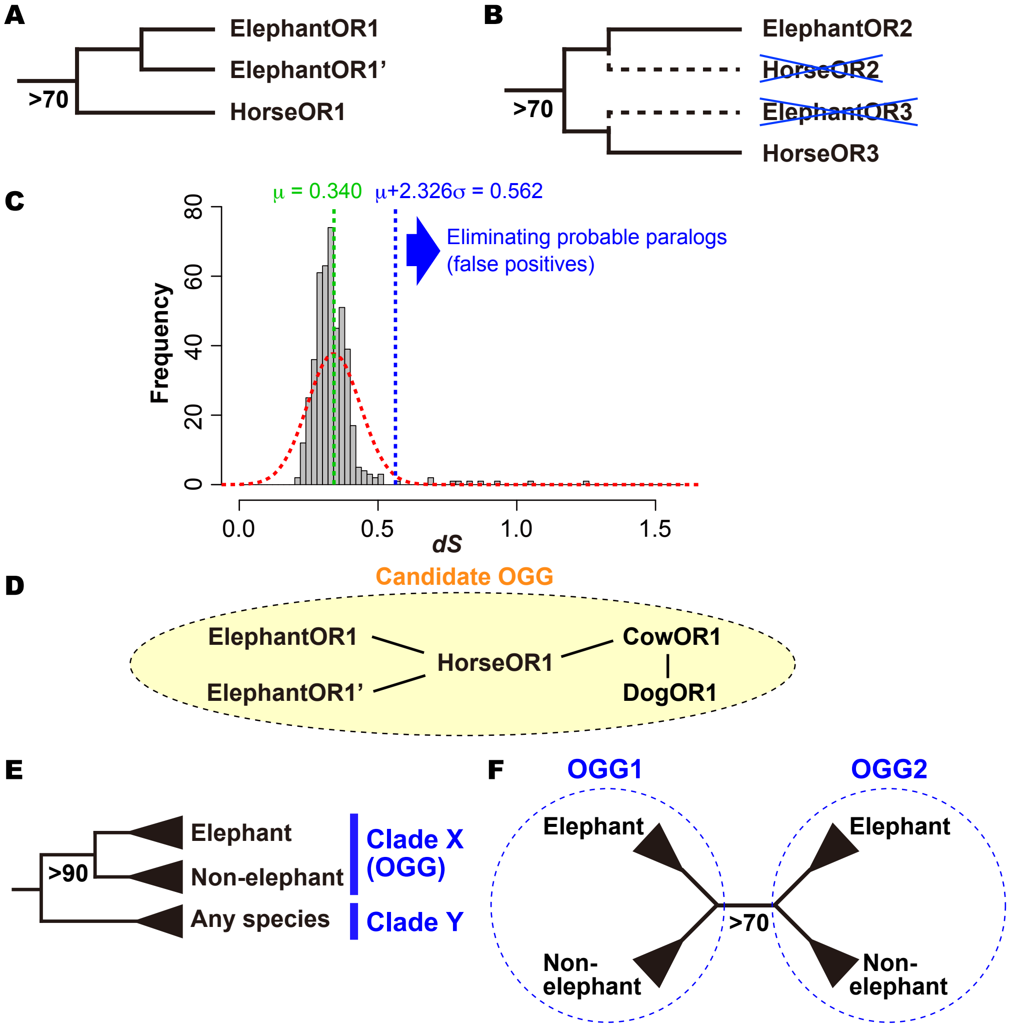


Figure S3. Identification of OGGs. (*A*) Identification of candidate orthologous gene pairs. In this case, (ElephantOR1, ElephantOR1’) and HorseOR1 were regarded as a candidate orthologous gene pair between elephant and horse. Note that both ElephantOR1 and ElephantOR1’ are orthologous to HorseOR1 because gene duplication occurred in the elephant lineage after speciation. (*B*) HorseOR2 and ElephantOR3 were independently lost in the horse and elephant lineages, respectively. In this case, ElephantOR2 and HorseOR3 were wrongly regarded to be orthologous to each other, although they are actually paralogs. The evolutionary distance between ElephantOR2 and HorseOR3 tends to become longer than that for true orthologs between elephant and horse. (*C*) Distribution of *dS* values for candidate orthologous gene pairs between elephant and horse. The red dotted line indicates the normal distribution with the mean (μ) and the standard deviation (σ) that were calculated from the *dS* values. The green and blue dotted lines indicate μ and μ + 2.326σ, respectively. Candidate orthologous gene pairs with *dS* values larger than the value indicated by the blue line were eliminated as probable paralogs (false positives), and the remaining pairs were used for the following analysis. (*D*) ‘A friend of a friend is a friend’ strategy. A line connecting two genes represents a candidate orthologous gene pair. In this case, all five genes are supposed to be members of a single candidate OGG, even though some orthologous relationships (*e.g.*, HorseOR1 vs. DogOR1) were not detected. (*E*) In this rooted tree, clade X contains both elephant and non-elephant genes and is supported with >90% bootstrap value. If clade X and clade Y contain gene(s) from at least one common species, genes in clade X were considered to form an OGG. (*F*) In an unrooted tree, when a clade contains both elephant and non-elephant genes with >70% bootstrap support, the genes in the clade were considered to form an OGG.

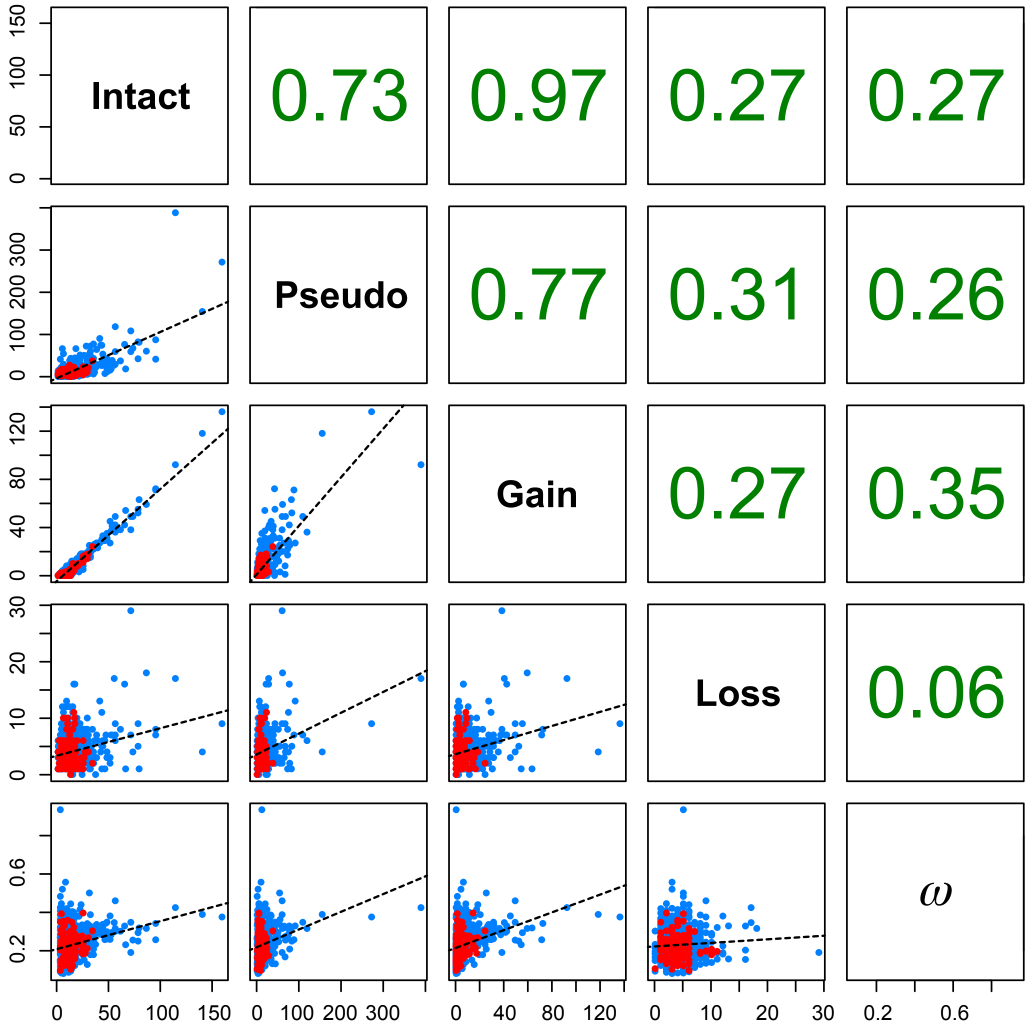


Figure S4. Pairwise comparisons among the number of intact genes, that of pseudogenes, that of gene gains, that of gene losses, and ω values in each OGG. The lower triangular part shows scatter plots. In each scatter plot, red and blue dots represent Class I and Class II genes, respectively, and the dashed line indicates the regression line. Each number in the upper triangular part indicates the correlation coefficients, *r*. The numbers of intact genes and those of pseudogenes are strongly correlated with each other (*r* = 0.73). ω values are significantly correlated with the numbers of intact genes ($r = 0.27$, $p = 3.8\times10^{-13}$) and those of gene gains ($r = 0.35$; $p < 2.2\times10^{-16}$), while no significant correlation was observed between ω values and the numbers of gene losses (*r* = 0.063, *p* = 0.095). The correlation between the numbers of gene losses and those of pseudogenes is not particularly stringent (*r* = 0.31). This apparently counter-intuitive observation can be explained by in the following way. Suppose that a pseudogenization event happened after gene duplication; the number of pseudogenes has increased by one. However, this event does not contribute to the number of gene losses, because only the number of intact genes is considered for estimation by the reconciled-tree method (Niimura and Nei 2007), and the number of intact genes does not change by this event. Therefore, the number of pseudogenes is not equal to the number of gene losses in many cases.

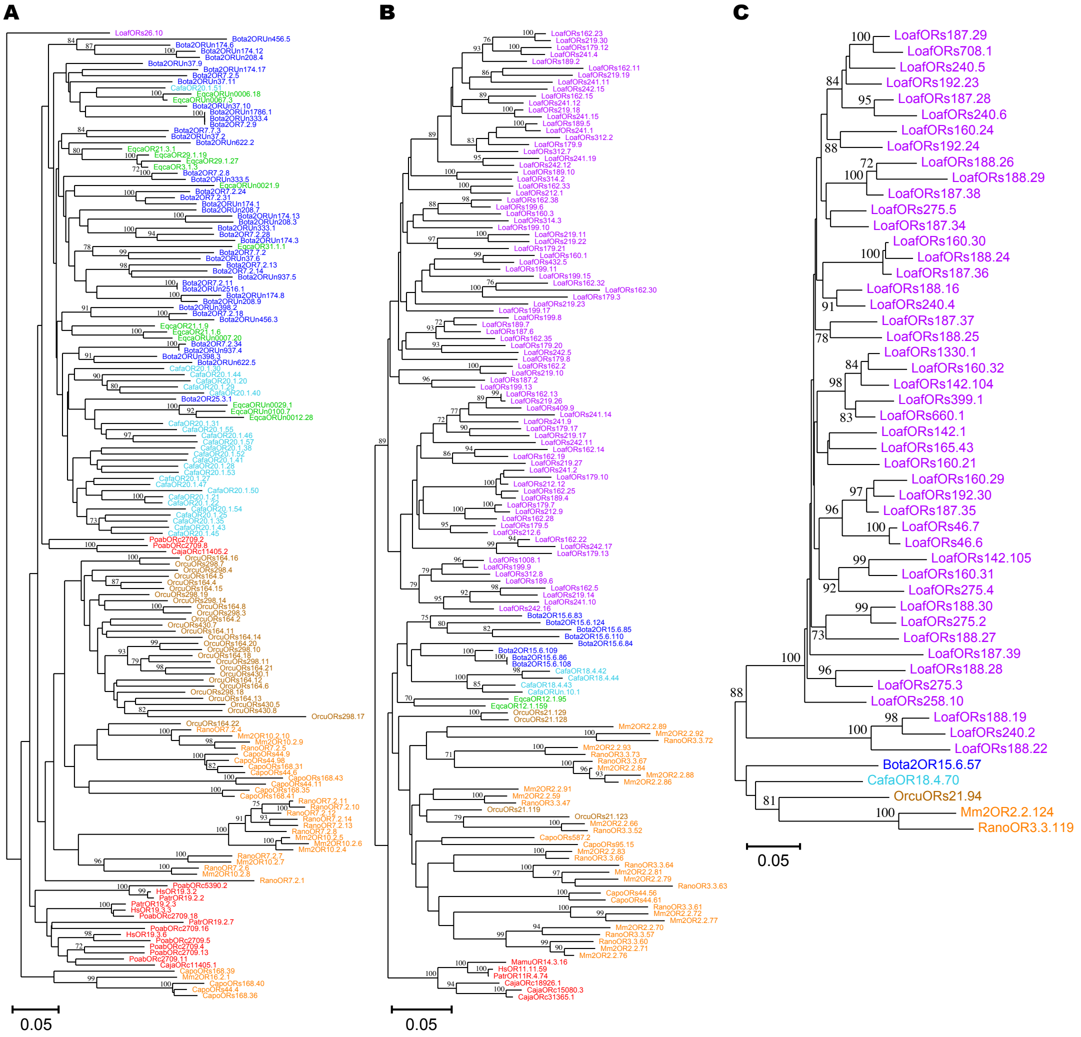


Figure S5. Neighbor-joining phylogenetic trees constructed from all intact OR genes in OGG2-1 (*A*), OGG2-2 (*B*), and OGG2-22 (*C*). Bootstrap values obtained from 500 resamplings are shown only for the nodes with bootstrap values greater than 70%. The color code for each gene is the same as Fig. 3A,C.

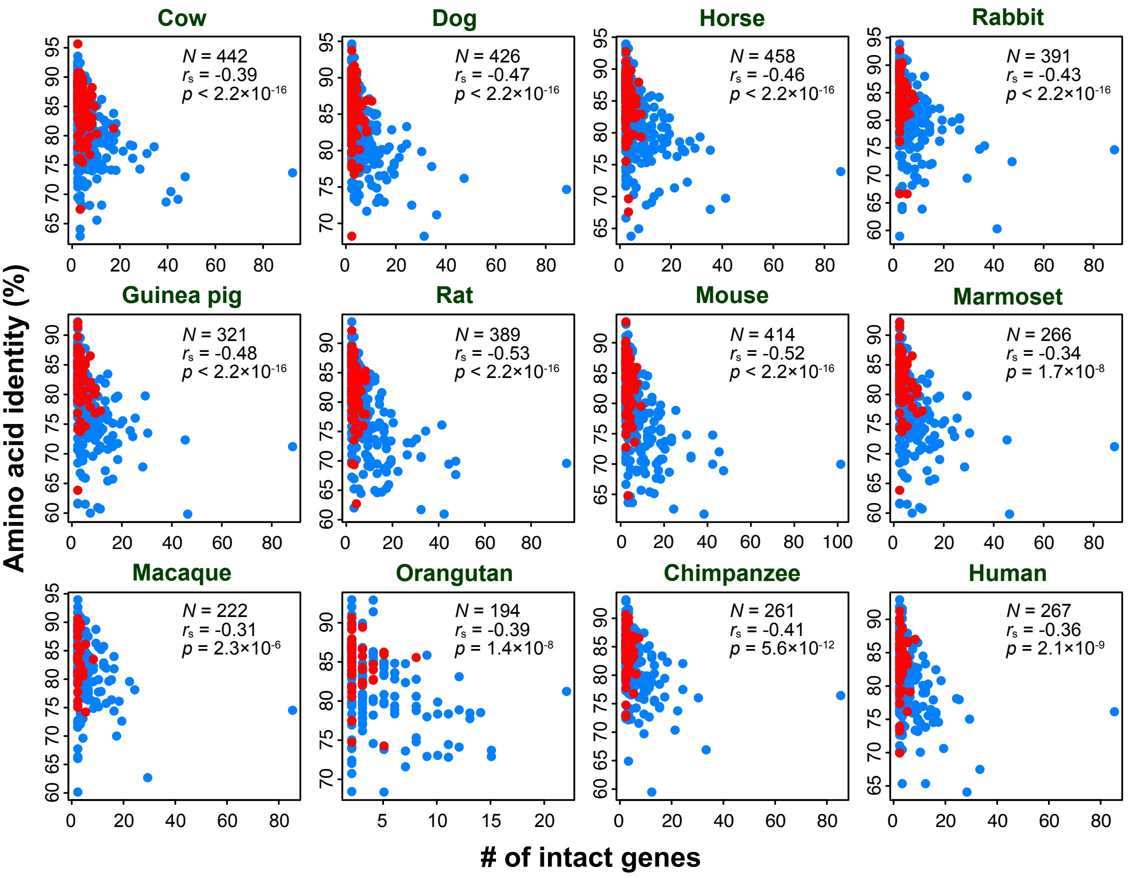
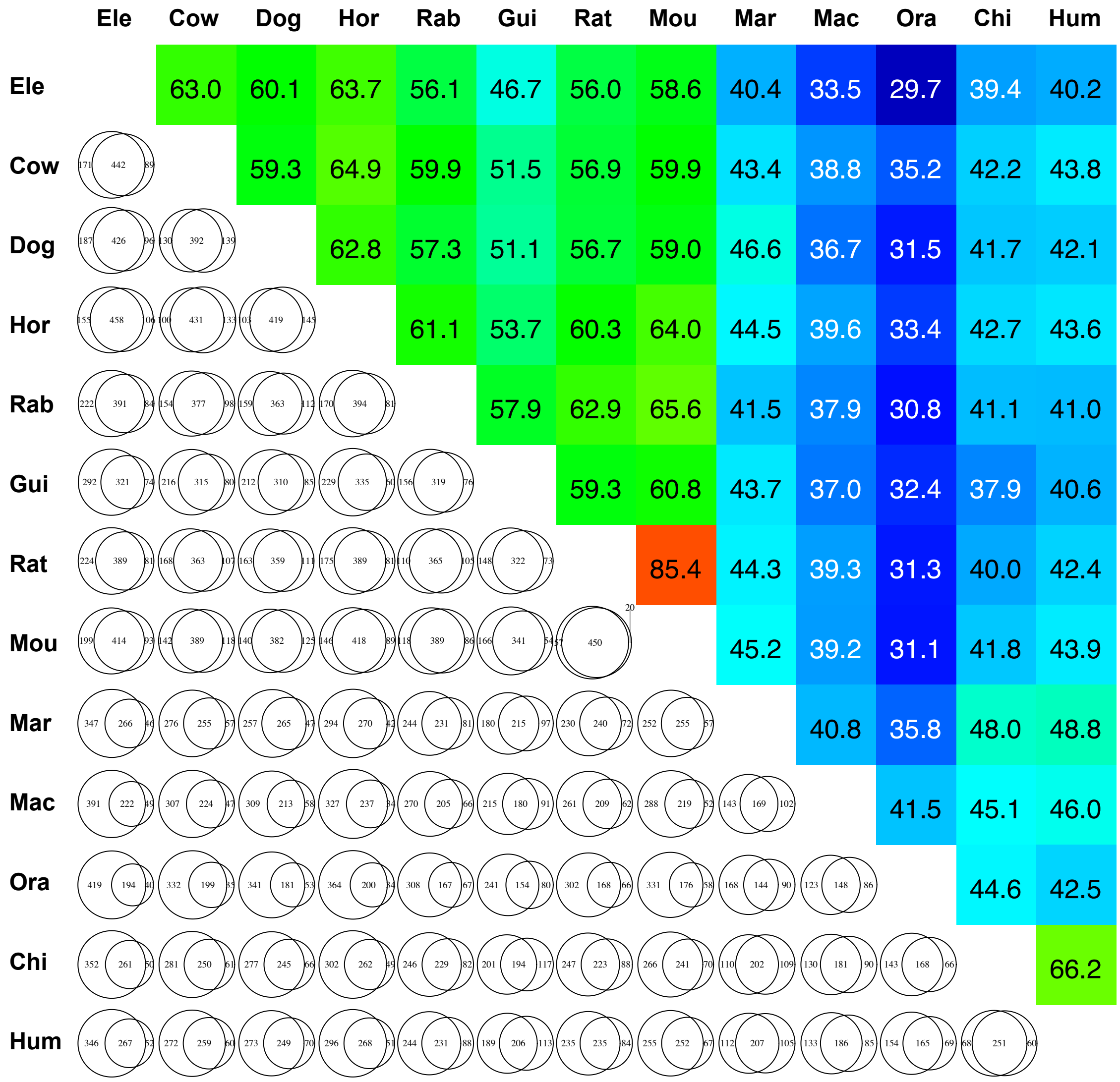


Figure S6. For each OGG, the total number of intact genes in elephant and each of the other species (indicated above each scatterplot) within respective OGGs was negatively correlated with amino acid sequence identity between elephant and the species among intact OR genes within the respective OGGs. When an OGG included two or more genes from either or both species, the mean of the amino acid sequence identities for all possible interspecies combinations of genes were used. OGGs that contained at least one intact gene from both elephant and a given species were considered. The number of OGGs examined (*N*), the Spearman's rank correlation coefficient r_s, and the *p* value were also shown at each scatterplot. Note that this kind of analyses makes sense only for a comparison between elephant and another species. Because the divergence between elephant and the other species has occurred first in the evolution of the 13 species examined, any pairs of an elephant gene and a gene from another species in the same OGG are orthologous to each other. However, the divergence of genes between two non-elephant species in the same OGG may have occurred before the divergence between the two species.



	Ele	Cow	Dog	Hor	Rab	Gui	Rat	Mou	Mar	Mac	Ora	Chi	Hum
Ele		63.0	60.1	63.7	56.1	46.7	56.0	58.6	40.4	33.5	29.7	39.4	40.2
Cow	171 442 89		59.3	64.9	59.9	51.5	56.9	59.9	43.4	38.8	35.2	42.2	43.8
Dog	187 426 96	130 392 139		62.8	57.3	51.1	56.7	59.0	46.6	36.7	31.5	41.7	42.1
Hor	155 458 106	100 431 133	103 419 145		61.1	53.7	60.3	64.0	44.5	39.6	33.4	42.7	43.6
Rab	222 391 84	154 377 98	159 363 112	170 394 81		57.9	62.9	65.6	41.5	37.9	30.8	41.1	41.0
Gui	292 321 74	216 315 80	212 310 85	229 335 60	156 319 76		59.3	60.8	43.7	37.0	32.4	37.9	40.6
Rat	224 389 81	168 363 107	163 359 111	175 389 81	110 365 105	148 322 73		85.4	44.3	39.3	31.3	40.0	42.4
Mou	199 414 93	142 389 118	140 382 125	146 418 89	118 389 86	166 341 54	57 450 20		45.2	39.2	31.1	41.8	43.9
Mar	347 266 46	276 255 57	257 265 47	294 270 42	244 231 81	180 215 97	230 240 72	252 255 57		40.8	35.8	48.0	48.8
Mac	391 222 49	307 224 47	309 213 58	327 237 34	270 205 66	215 180 91	261 209 62	288 219 52	143 169 102		41.5	45.1	46.0
Ora	419 194 40	332 199 35	341 181 53	364 200 34	308 167 67	241 154 80	302 168 66	331 176 58	168 144 90	123 148 86		44.6	42.5
Chi	352 261 50	281 250 61	277 245 66	302 262 49	246 229 82	201 194 117	247 223 88	266 241 70	110 202 109	130 181 90	143 168 66		66.2
Hum	346 267 52	272 259 60	273 249 70	296 268 51	244 231 88	189 206 113	235 235 84	255 252 67	112 207 105	133 186 85	154 165 69	68 251 60	

Figure S7. Pairwise comparisons of OR gene repertories among the 13 placental mammals. Each Venn diagram at the lower left indicates the number of OGGs that are present in both or either of the two species compared. For example, for elephant-horse comparison, 458 OGGs are present both in elephant and in horse, while 155 OGGs are present in elephant but absent in horse, and 106 OGGs are present only in horse (elephants and horses retain 613 and 564 OGGs; see Fig. 6). Note that each number indicates the number of OGGs and not that of OR genes. Each number at the upper right represents the fraction (in percentage) of the number of OGGs that are commonly present in both of the two species to that of OGGs present in at least one species. For example, for elephant-horse comparison, the value was calculated as 458 / (155 + 458 + 106) = 0.637. Ele, elephant; Cow, cow; Dog, dog; Hor, horse; Rab, rabbit; Gui, guinea pig; Rat, rat; Mou, mouse, Mar, marmoset; Mac, macaque; Ora, orangutan; Chi, chimpanzee; Hum, human.

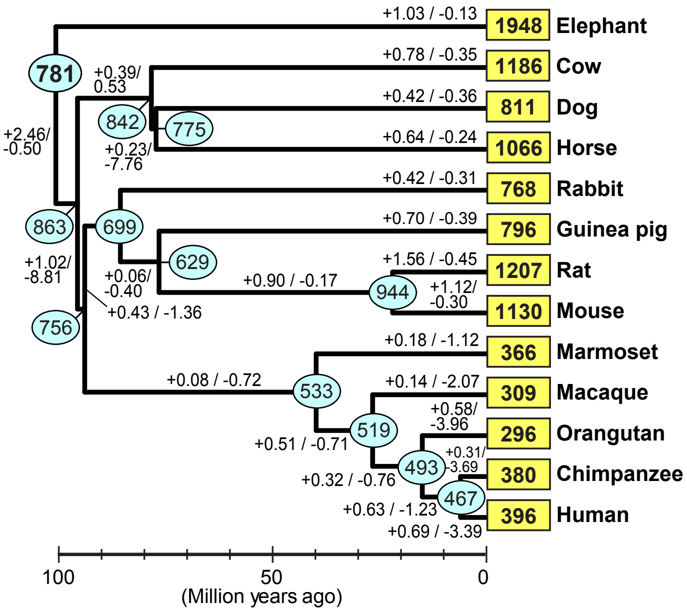


Figure S8. The birth rate β and the death rate δ of OR genes (per gene per 100 million years) in each branch of the phylogeny of 13 placental mammals. β and δ are indicated by plus and minus signs, respectively. These values were calculated from the numbers of gene gains and gene losses shown in Fig. 6. See Methods.

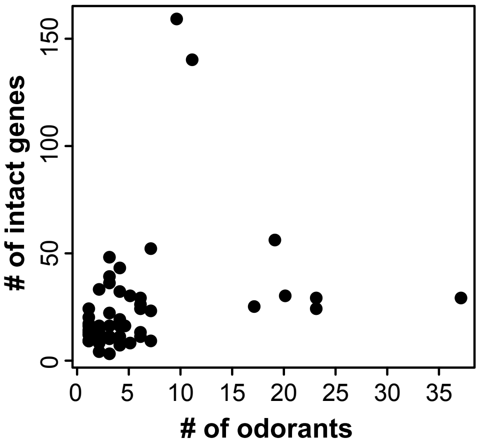


Figure S9. The number of intact OR genes in an OGG is positively correlated with the number of ligands binding to a member OR with statistical significance (r_s = 0.470; p = 0.00020). The data regarding OR-odorant pairs were obtained from Fig. 1 in Saito *et al.* (2009). When two or more ORs belong to the same OGG, the mean of the numbers of ligands among these ORs was used. There were four such pairs of ORs. We investigated the statistical significance of the Spearman's rank correlation coefficient under the assumption that respective OGG was independent and randomly sampled (n = 58).